수의학박사학위논문

Genome-Wide Methylome Analysis in Canine Mammary Tumors and Immune Cells Elucidates Epigenetic Tumor Regulation and Its Application to a Malignancy Prediction

개 유선종양 조직 및 면역세포의 메틸롬 분석을 통한 후성유전학적 암 조절 기전 규명 및 악성종양 예측모델 개발

2023 년 8 월

서울대학교 대학원

수의학과 수의생명과학 전공

(수의생화학)

남 아 름

Genome-Wide Methylome Analysis in Canine Mammary Tumors and Immune Cells Elucidates Epigenetic Tumor Regulation and Its Application to a Malignancy Prediction

Under the supervision of Professor Je-Yoel Cho

DISSERTATION

Presented in Partial Fulfillment of the Requirement for

the Degree of DOCTOR OF PHILOSOPHY

By

A-Reum Nam

Major in Veterinary Biomedical Sciences

(Veterinary Biochemistry)

Department of Veterinary Medicine

The Graduate School

Seoul National University

August 2023

개 유선종양 조직 및 면역세포의 메틸롬 분석을 통한 후성유전학적 암 조절 기전 규명 및 악성종양 예측모델 개발

Genome-Wide Methylome Analysis in Canine Mammary Tumors and Immune Cells Elucidates Epigenetic Tumor Regulation and Its Application to a Malignancy Prediction

지도교수 조 제 열

이 논문을 수의학박사 학위논문으로 제출함
2023 년 7 월

서울대학교 대학원
수의학과 수의생명과학 전공 (수의생화학)
남 아 름

남아름의 수의학박사 학위논문을 인준함
2023 년 8 월

위 원 장 _____ (인)
부 위 원 장 _____ (인)
위 원 _____ (인)
위 원 _____ (인)
위 원 _____ (인)

Abstract

Genome-Wide Methylome Analysis in Canine Mammary Tumors and Immune Cells Elucidates Epigenetic Tumor Regulation and Its Application to a Malignancy Prediction

A-Reum Nam
Major in Veterinary Biomedical Sciences
Department of Veterinary Medicine
Seoul National University

Canine mammary tumor (CMT) has long been considered as a good animal model for human breast cancer (HBC) due to their pathological and biological similarities. However, only a few aspects of the epigenome have been explored in both HBC and CMT. Moreover, DNA methylation studies have mainly been limited to the promoter regions of genes. Genome-wide dysregulation of CpG methylation accompanies tumor progression and characteristic states of cancer cells, prompting a rationale for biomarker development. Understanding how the archetypic epigenetic modification

determines systemic contributions of immune cell types is the key to further clinical benefits.

In Chapter 1, the study focuses on genome-wide methylome profiles in canine mammary tumors (CMT) and adjacent normal tissues, particularly highlighting the intron regions as potential targets for epigenetic regulation. The analysis revealed the identification of numerous tumor suppressors and oncogenes. Notably, differentially methylated genes (DMGs), including intron-DMRs (differentially methylated regions), were enriched in cancer-associated biological processes. Interestingly, two PAX motifs, PAX5 (tumor suppressive) and PAX6 (oncogenic), were frequently observed in hyper- and hypo-methylated intron-DMRs, respectively. The study found an inverse correlation between hyper-methylation at PAX5 motifs in the intron regions of CDH5 and LRIG1 genes and their gene expression, while CDH2 and ADAM19 genes with hypomethylated PAX6 motifs in their intron regions showed up-regulation. These findings were validated both in the originally MBD-sequenced specimens and additional clinical samples. Additionally, the study investigated intron methylation and downstream gene expression of these genes in human breast invasive carcinoma datasets from the TCGA database. The regional methylation alterations were conserved in the corresponding intron regions, resulting in altered gene expression in breast cancer. This study provides evidence supporting the conservation of epigenetic regulation in both CMT and human breast cancer

(HBC), highlighting the importance of intronic methylation in understanding gene regulation in these diseases.

On the other hand, the response of immune cells to cancer plays a crucial role in determining the prognosis of cancer and the efficacy of anticancer treatments. Emerging evidence suggests that immune checkpoints, which are key targets of immunotherapy, are also subject to epigenetic regulation. Consequently, Chapter 2 of this dissertation focuses on investigating the differential DNA methylome profiles in peripheral blood mononuclear cells (PBMCs) obtained from patients with mammary tumors. I conducted methylated CpG-binding domain sequencing (MBD-seq) and investigated the differential methylome landscapes of peripheral blood mononuclear cells (PBMCs) from 76 canines with or without mammary tumors. Through gene set enrichment analysis, it was found that genes involved in the growth and differentiation of T- and B-cells are highly methylated in tumor PBMCs. Furthermore, the study identifies increased methylation and reversed expression in representative marker genes (BACH2, SH2D1A, TXK, UHRF1) that regulate immune cell proliferation. Although there was no dramatic difference in the PBMC methylome between malignant and benign tumors, we devised a machine-learning approach to predict malignancy utilizing our methylome dataset. This study provides valuable insights into the comprehensive epigenetic regulation of circulating

immune cells in response to tumors, offering a new framework for identifying benign and malignant cancers through genome-wide methylome analysis.

In summary, this dissertation provides a comprehensive exploration of epigenetic landscapes in canine mammary tumors, utilizing genome-wide analysis of methylome and transcriptome profiles both in tumor tissues and peripheral blood mononuclear cells. Cancer regulation through methylation of intronic motifs reveals intriguing similarities between humans and dogs, highlighting the value of companion dogs in advancing our understanding of cancer research. Moreover, the application of immune cell methylome data for predicting malignant tumors presents potential scalability in diagnosing malignancy across various types of cancer in humans as well as dogs. Although additional validation studies are needed for the clinical application of the diagnostic models, I believe these studies will be a crucial cornerstone for treating and diagnosing tumors.

Keywords: Methylome; Transcriptome; Canine Mammary Tumor; Human Breast Cancer; Comparative Medicine; PBMC; Machine Learning; Biomarker

Student Number: 2016-31835

Table of contents

Page

Abstract ..... **i**

Table of contents ..... **v**

List of figures ..... **viii**

List of tables ..... **xii**

Abbreviation ..... **xiv**

Background **1**

1. Epigenetics in cancer 1

    1.1. What is DNA methylation? 1

    1.2. DNA methylation in cancer 2

2. Comparative medicine 6

3. Genome-wide methylome technologies 9

4. Aims of the dissertation 13

Chapter 1. Alternative methylation of intron motifs is associated with cancer-related gene expression in both canine mammary tumor and human breast cancer 15

1. Introduction 16
2. Materials and methods 19
3. Results 30
4. Discussion 62

Chapter 2. The landscape of PBMC methylome in canine mammary tumors reveals the epigenetic regulation of immune marker genes and its potential application in predicting tumor malignancy 66

1. Introduction 67
2. Materials and methods 70
3. Results 88
4. Discussion 132

General conclusion 138

References 140

Abstract in Korean (국문초록) 149

List of figures

Background

Figure B.1 Schematic representation of DNA methylation and its preferential occurrence at CpG site.

Figure B.2 Dynamic changes of methylation across the genome-wide CpG region in cancer.

Figure B.3 Leveraging pet dogs as a translational model for human clinical trials in oncology.

Figure B.4 Comparison of genome-wide DNA methylation technologies.

Figure B.5 Conceptual scheme of the dissertation.

Chapter 1

Figure 1.1 Schematic presentation of genome wide methylation profiling in CMT.

Figure 1.2 Visualizing methylation peaks using processed MBD-seq from 11 pairs of CMT and adjacent normal tissues.

Figure 1.3 The CpG coverage of genome wide DNA sequence patterns.

Figure 1.4 Analytical strategies.

Figure 1.5 Identification of differentially methylated regions (DMRs) among the three CMT subtypes and between CMT and adjacent normal.

Figure 1.6 Functional annotation of CMT-DMGs.

Figure 1.7 The expression level of the top 4 orthologous genes ranked by OncoScore in canine mammary tumor.

Figure 1.8 Functional annotation of Subtype-DMGs.

Figure 1.9 Intron DMRs may associate with cancer-related genes.

Figure 1.10 Adjust thresholds to select distinguished CMT-DMRs for intronic motif analysis.

Figure 1.11 Kaplan-Meier plots showed PAX5 and PAX6 expression reversely effect on the survival rate of breast cancer patients..

Figure 1.12 PAX motifs are enriched in hyper- and hypo-methylated intron DMRs.

Figure 1.13 PAX motifs are enriched in hyper- and hypo-methylated intron DMRs. Validation of intron hypermethylation in the candidate genes, CDH5 and LRIG1.

Figure 1.14 Validation of individual CG methylation around PAX5 motif regions in CDH5 and LRIG1 genes.

Figure 1.15 Conservation of intron DMRs and associating RNA expression in the candidate genes between HBC and CMT.

Chapter 2

Figure 2.1 Pair-wise comparison for genome-wide PBMC methylome datasets from benign, carcinoma, and normal dogs.

Figure 2.2 Quality check and processing MBD-seq data.

Figure 2.3 Venn diagram for hyper- and hypo-methylated DMRs.

Figure 2.4 Unsupervised and supervised clustering between comparison groups.

Figure 2.5 Gene enrichment analysis for DMGs shows differential immune signatures between tumor and normal PBMCs.

Figure 2.6 Enriched terms ranked in the Top 3 by combined score according to comparison groups.

Figure 2.7 Immune cell markers involved in normal proliferation and activation of B-cells, T-cells, and NK cells are hypermethylated in tumor PBMCs.

Figure 2.8 Targeted CpG methylation and expression analysis in representative hypermethylated genes related to immune cell activation.

Figure 2.9 A machine learning-based diagnostic two-step classifier discriminating tumor from normal PBMCs followed by carcinoma from benign PBMCs.

Figure 2.10 Evaluating the accuracy and predictive performance of the two-step classifier.

Figure 2.11 PCA analysis using DMRs involved in the BC classifiers.

Figure 2.12 Permutation accuracy importance of DMRs used for modeling the final BC classifier.

Figure 2.13 The predictive performance of transcriptome-based two-step classifier.

List of Tables

Background

Table B.1 The advantages and disadvantages of representative genome-wide DNA methylation sequencing techniques.

Chapter 1

Table 1.1 Information for CMT tissue samples used for MBD-seq

Table 1.2 Information for CMT tissue samples used for BS-seq

Table 1.3 Primers designed for BS-conversion PCR

Table 1.4 Quality check for MBD-seq data

Chapter 2

Table 2.1 The information about dog donors providing blood samples used for MBD-seq

Table 2.2 The list of primers designed for targeted BS-sequencing

Table 2.3 The list of 127DMRs which have high feature importance in BC classifier

Table 2.4 The list of hypermethylated DMRs in immune cell type markers (Panglao DB)

Table 2.5 The information of unknown dog PBMC donors (used for validation sets of NT classifier)

Abbreviation

ADAM	A Disintegrin And Metalloproteinase
BACH2	BTB Domain and CNC Homolog 2
BP	Biological process
BRCA	Breast Invasive Carcinoma
CAR-T	Chimeric Antigen Receptor T cell
CC	Cellular component
CDH	Cadherin
CGI	CpG island
CMT	Canine Mammary Tumor
CpG	Cytosine and Guanine separated by phosphate
DMG	Differentially Methylated Gene
DMR	Differentially Methylated Region
DNMT	DNA methyl transferases
EGF	Epidermal growth factor

GO	Gene ontology
HBC	Human breast cancer
KEGG	Kyoto Encyclopedia of Genes and Genomes
LMM	Linearized mixed model
LRIG	Leucine Rich Repeats And Immunoglobulin Like Domains
MBD-seq	Methyl CpG Binding Domain Sequencing
MGT	Mammary Gland Tumor
ML	Machine Learning
PAX	Paired box
PBMC	Peripheral Blood Mononuclear Cell
PCA	Principal Component Analysis
ROC	Receiver Operating Characteristic Curve
SD	Standard variation
SH2D1A	SH2 Domain Containing 1A

TCGA	The Cancer Genome Atlas
TCR	T Cell Receptor
TF	Transcription factor
TIL	Tumor Infiltrating Lymphocyte
TME	Tumor Microenvironment
TSS	Transcription Start Site
TTS	Transcription Termination Site
TXK	Tyrosine Kinase
UHRF1	Ubiquitin Like with PHD and Ring Finger Domains 1

Background

1. Epigenetics in cancer

1-1. What is DNA methylation?

DNA methylation is a vital epigenetic modification that regulates gene expression and genome stability [1]. It involves adding a methyl group to cytosine at the carbon 5 position, mainly occurring in CpG dinucleotides (**Figure B.1**). During early development, DNA methylation patterns are established and faithfully maintained throughout cell divisions, contributing to cellular identity and differentiation. Abnormal DNA methylation patterns are linked to diseases such as cancer, cardiovascular disorders, neurological disorders, and imprinting disorders [2].

In cancer, alterations in DNA methylation patterns are prominent. Global hypomethylation, involving a reduction in genome-wide DNA methylation, is commonly observed in cancer cells and is associated with genomic instability and the activation of transposable elements. Simultaneously, localized hypermethylation at CpG islands in gene promoters can lead to the silencing of tumor suppressor genes, contributing to cancer development and progression [3]. Additionally, DNA methylation changes can occur in various genomic regions, including gene bodies

and enhancers, influencing gene expression and cellular functions in a context-dependent manner.

Understanding the dynamic nature of DNA methylation and its influence on gene regulation is pivotal in unraveling the molecular mechanisms that drive cancer development and progression. Such understanding holds great promise for identifying epigenetic alterations associated with cancer, which can serve as valuable insights for discovering potential biomarkers for early detection, prognosis, and targeted therapeutics.

1-2. DNA methylation in cancer

Genome-wide CpG methylation is a critical epigenetic modification that regulates gene expression and is closely linked to various biological processes, including development, differentiation, and disease progression. In cancer, aberrant CpG methylation patterns, as depicted in **Figure B.2**, are frequently observed, leading to the disruption of important genes involved in tumor initiation, progression, and metastasis. Integrating genome-wide CpG methylation data with gene expression data has provided valuable insights into the molecular mechanisms driving cancer development and has identified potential biomarkers and therapeutic targets [4,5].

The relationship between CpG methylation and gene expression is complex and dynamic. Hypermethylation of CpG sites located in promoter regions often leads to

gene silencing, while hypomethylation can result in increased gene expression [3,6]. However, recent studies have highlighted the importance of considering CpG methylation beyond promoter regions, such as intronic and intergenic regions, which have emerged as novel regulatory elements for gene expression [6]. Intriguingly, CpG methylation changes in these regions have been implicated in the regulation of critical cancer-related pathways and the balance between tumor suppressor and oncogene activities.

Furthermore, the investigation of genome-wide CpG methylation profiles in cancer has provided valuable insights into the identification of cancer subtypes, disease prognosis, and treatment response prediction. The development of high-throughput sequencing technologies and advanced bioinformatics tools has enabled the comprehensive characterization of CpG methylation landscapes, facilitating the discovery of novel CpG methylation markers with diagnostic and prognostic potential.

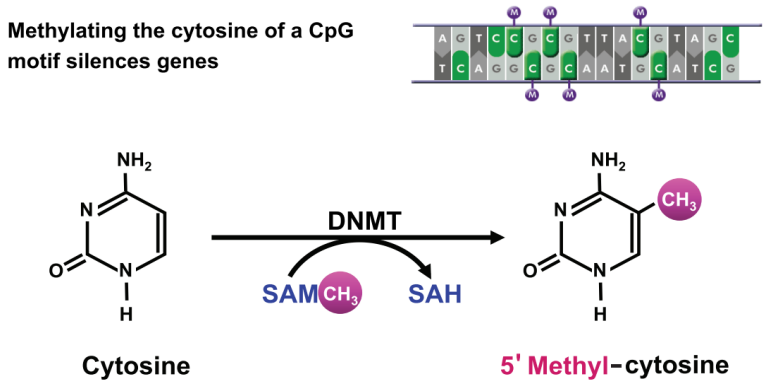


Figure B.1. Schematic representation of DNA methylation and its preferential occurrence at CpG site. DNA methylation, facilitated by DNA methyltransferase (DNMT), converts cytosine to 5'methyl-cytosine. This process primarily targets cytosines followed by a guanine, referred to as CpG sites, and plays a crucial role in establishing DNA methylation patterns [7]. (S-adenosylmethionine, SAM; S-adenosylhomocysteine, SAH)

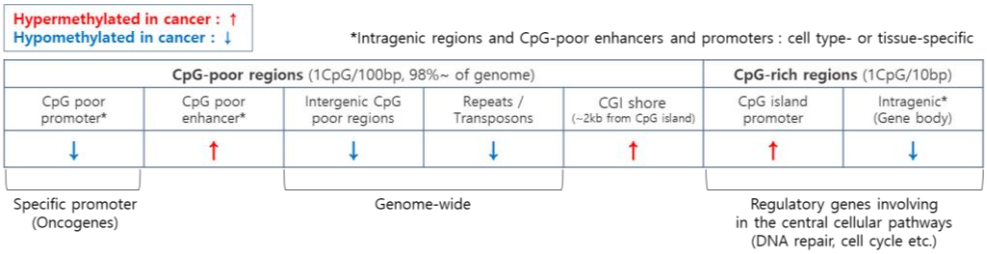


Figure B.2. Dynamic changes of methylation across the genome-wide CpG region in cancer. In cancer cells, hypermethylation in CpG islands leads to the silencing of genes, resulting in aberrant gene expression. Simultaneously, hypomethylation of intergenic regions and CpG-poor promoters contributes to genomic instability and abnormal gene expression, respectively, further complicating the disease [3].

2. Comparative medicine

In comparative medicine, the study of dog cancer offers several advantages that contribute to our understanding of cancer biology and the development of effective therapies (**Figure B.3**). Here are some key advantages of dog cancer research:

1) Spontaneous and Naturally Occurring Cancer: Dogs develop cancer spontaneously, similar to humans, and share many similarities in terms of tumor biology, genetics, and clinical behavior [8]. This makes dogs an excellent comparative model for studying cancer in a realistic and clinically relevant context. The occurrence of cancer in dogs is not artificially induced, providing insights into the natural progression and heterogeneity of tumors.

2) Similarities in Tumor Types and Pathways: Dogs develop a wide range of tumor types that closely resemble those found in humans, including breast, lung, skin, bone, and prostate cancer [9,10]. The similarities extend to the molecular pathways involved in tumor initiation, progression, and metastasis. Studying canine cancers allows for the investigation of shared mechanisms and potential therapeutic targets across species.

3) Genetic Diversity: Like humans, dogs exhibit genetic diversity within different breeds and populations [11,12]. This diversity offers an opportunity to investigate the influence of genetic factors on cancer susceptibility, tumor development, and treatment response. By studying cancer in different dog breeds, researchers can

identify genetic variants associated with specific cancer types, paving the way for personalized medicine approaches.

4) Comparative Therapeutic Evaluation: Dogs with cancer can benefit from treatment interventions similar to those used in human oncology. This allows for the evaluation of novel therapeutic strategies, such as targeted therapies, immunotherapies, and gene therapies, in a spontaneous tumor model [9,10]. The response to treatment and the assessment of adverse effects can provide valuable preclinical data to inform human clinical trials.

In summary, dog cancer research in comparative medicine offers unique advantages for understanding cancer biology, evaluating therapeutic interventions, and advancing translational medicine. The natural occurrence of cancer in dogs, shared tumor types and pathways, genetic diversity, and translational implications make dogs a valuable model for improving our understanding of cancer and developing more effective treatments for both dogs and humans.

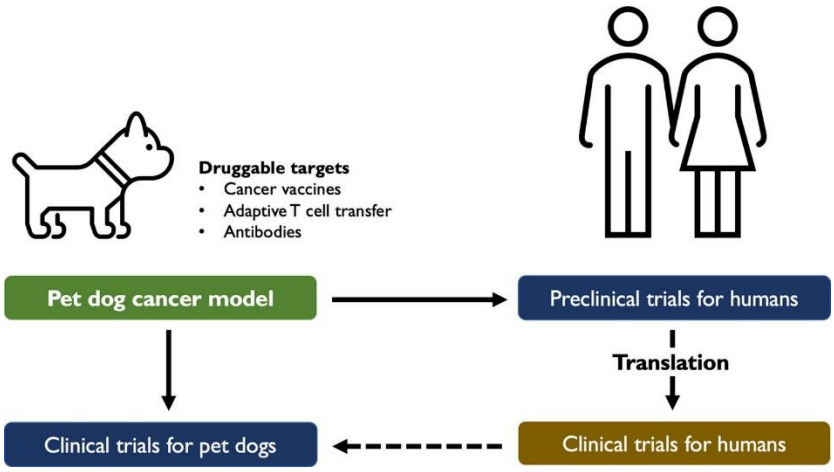


Figure B.3. Leveraging pet dogs as a translational model for human clinical trials in oncology. Spontaneous tumors in pet dogs offer a valuable opportunity to bridge the gap between preclinical rodent models and human clinical trials, overcoming limitations and providing a relevant model due to their natural occurrence, immune-competence, and genetic similarity to humans [10].

3. Genome-wide methylome technologies

Genome-wide methylome technologies have revolutionized the field of epigenetics by enabling comprehensive analysis of DNA methylation patterns across the entire genome. DNA methylation, an essential epigenetic modification, plays a critical role in gene regulation, development, and disease. Advancements in sequencing technologies and the development of innovative techniques have provided powerful tools for studying DNA methylation patterns at a global scale (**Figure B.4 & Table B.1**).

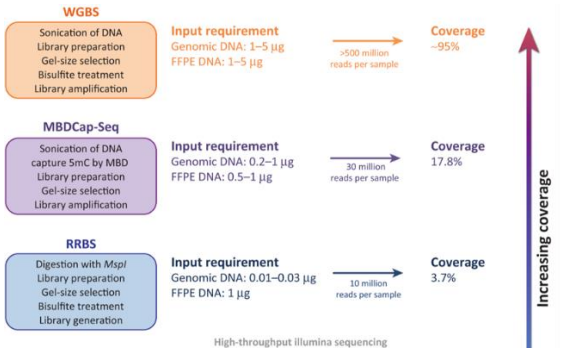
Whole-genome bisulfite sequencing (WGBS) is a widely used method for genome-wide DNA methylation analysis. WGBS involves treating genomic DNA with sodium bisulfite, which converts unmethylated cytosines to uracil while leaving methylated cytosines unchanged [3]. Subsequent high-throughput sequencing of the bisulfite-treated DNA allows for the determination of DNA methylation status at single-base resolution throughout the genome. WGBS provides comprehensive and accurate information about DNA methylation patterns, enabling the identification of differentially methylated regions (DMRs) and the exploration of their functional implications.

Array-based DNA methylation profiling is another commonly employed approach for genome-wide DNA methylation analysis. This method utilizes DNA methylation microarrays containing probes targeting specific CpG sites across the genome. By hybridizing bisulfite-converted DNA to these arrays, researchers can obtain

quantitative measurements of DNA methylation levels at thousands to millions of CpG sites simultaneously [13]. Array-based profiling offers a cost-effective and high-throughput method for large-scale DNA methylation studies.

Another notable technique in genome-wide methylome analysis is Methylated DNA Binding Domain sequencing (MBD-seq). MBD-seq capitalizes on the affinity of the MBD protein for methylated DNA, allowing for the enrichment and sequencing of methylated regions in the genome [3,13]. MBD-seq provides a targeted approach to investigate DNA methylation patterns by capturing methylated DNA fragments, offering an efficient and cost-effective alternative to whole-genome sequencing.

Collectively, these genome-wide methylome technologies, including WGBS, array-based profiling, and MBD-seq, have greatly advanced our understanding of DNA methylation dynamics and its functional implications. They have facilitated the identification of key regulatory regions, discovery of novel epigenetic marks, and identification of disease-associated DNA methylation patterns. These technologies have become indispensable tools in the field of epigenetics, paving the way for further insights into the complex relationship between DNA methylation, gene expression, and human health.



Genome wide

	WGBS	MBDCap	RRBS	HM450
Promoter	1 962 844	1 281 138	504 446	187 791
Genic	11 951 925	2 188 593	312 957	175 760
Intergenic	13 116 432	1 571 060	236 875	118 871
Total CpGs assayed	**27 031 201**	**5 040 791**	**1 054 278**	**482 422**

CpG-rich regions

CpG islands	2 019 500	1 572 591	641 182	150 253
CpG shores	1 936 549	902 062	127 090	111 988

Figure B.4. Comparison of genome-wide DNA methylation technologies. Despite its relatively low overall coverage of 17.8%, MBD-seq demonstrates a remarkably high coverage of CpG-rich regions, which is comparable to that of WGBS [3].

Table B.1. The advantages and disadvantages of representative genome-wide DNA methylation sequencing techniques.

	Method	Advantages	Disadvantages
WGBS	Sodium bisulfite treatment	Single-nucleotide resolution Whole-genome coverage	Requires computational expertise and higher costs
MeDIP-seq / MBD-seq	[MeDIP] Enrich mC with antibody : single-stranded DNA fragments [MBD] Enrich mCpG with MBD protein : double-stranded DNA fragments	Methylated regions of low CpG density (e.g., intergenic regions) CpG-dense regions (e.g., CpG islands) Can focus on CpG-rich regions and regulatory elements	Does not provide single-base resolution of methylation patterns Can have some bias in capturing methylated regions.
RRBS	MspI enzyme digestion Sodium bisulfite treatment	More cost-effective Provides intermediate resolution and coverage of CpG-dense regions	Biased towards CpG-rich regions, limiting the coverage of CpG-poor regions

4. Aims of the dissertation

The primary objective of this dissertation is to comprehensively explore the epigenetic landscapes in canine mammary tumors (CMT) and their potential implications for cancer prevention, treatment, and diagnosis. The study aims to investigate the genome-wide methylome and transcriptome profiles in both tumor tissues and peripheral blood mononuclear cells (PBMCs) of CMT. By analyzing DNA methylation patterns and gene expression alterations, the research seeks to identify key regulatory mechanisms and potential biomarkers associated with CMT. Additionally, the study aims to evaluate the similarities and conservation of epigenetic regulation between CMT and human breast cancer (HBC), highlighting the value of companion dog research in advancing our understanding of cancer in both species (**Figure B.5**). Furthermore, the dissertation explores the application of immune cell methylome data for predicting malignant tumors, with the goal of expanding its relevance to human oncology and providing insights for practical implementation in clinical settings. Ultimately, the research aims to contribute to the recognition of the vital role of methylome research in cancer and offer new insights for improving cancer prevention, treatment, and diagnosis strategies.

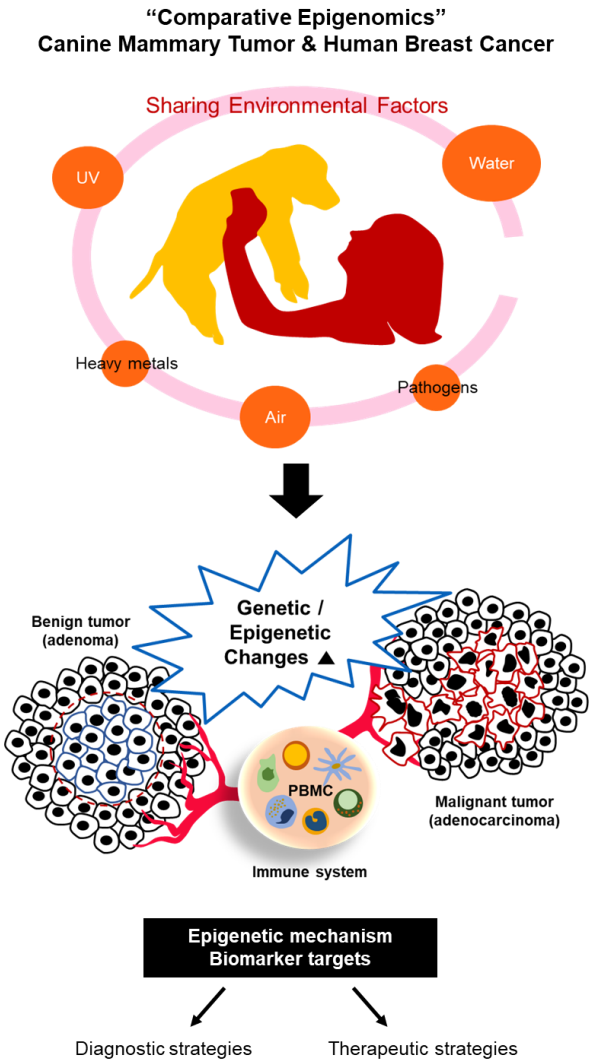


Figure B.5. Conceptual scheme of the dissertation. Humans and dogs, sharing diverse environments, provide a valuable framework for comparative medicine focused on naturally occurring cancers. Through a comparative analysis of the epigenome observed in dog mammary tumors and human breast cancer, profound insights can be obtained for the diagnosis and treatment of cancer.

CHAPTER I

Alternative methylation of intron motifs is associated with cancer-related gene expression in both canine mammary tumor and human breast cancer

Introduction

Breast cancer (BC) is the most frequently diagnosed and the second leading cause of cancer death in woman worldwide [14]. The comparison of 5-year survival rates between cancer stages 4 and 2, 27% vs. 99% in the USA, clearly shows that earlier diagnosis is crucial for increasing patient survival [15]. Many BC risk factors have been reported; some are uncontrollable, such as old age and gene mutations, while some are controllable, such as diet and smoking [16]. Only about 5-10% of BCs are thought to be hereditary [17]. Representatively, inherited mutations in *BRCA1 and BRCA2*, which have roles in DNA repair, have been known as the most common cause of hereditary BC [18]. In addition to inherited mutations, somatic mutations of dozens of genes, including *CCND1, ERBB2, PIK3CA, PTEN, etc.*, have been revealed as driver mutations that can lead to functional abnormalities and initiate breast tumorigenesis [19,20]. The fast-growing databases of various human cancers, such as COSMIC and TCGA, now provide researchers with access to genomic data to test their hypothesis in clinical samples (https://cancer.sanger.ac.uk/cosmic; https://www.cancer.gov/tcga) [21,22]. On the other hand, the molecular biological effects of environmental factors such as smoking, diet and exercise [16] are not readily accessible in BC and further approaches are needed to investigate epigenomic changes, including DNA methylation [23].

The association of CpG dinucleotide DNA methylation with cancer-related phenotypes [24] is well understood in various types and at all stages of cancer

progression [25,26] . Hypermethylation, which has been known to be associated with repressed gene expression of tumor suppressors, is one of the important paradigms of carcinogenesis [27] and is supported by the activated mutations of DNA methyltransferases (DNMTs) being oncogenic in several tissues [28]. In various human cancers, genome wide methylation has been profiled [27] and global DNA hypomethylation [29], along with local hyper- (tumor suppressors) and hypo- (oncogenes) methylations concomitant with the respective silencing and activating of gene expression [30,31] were reported and suggested as potential diagnostic and predictive biomarkers [32]. The use of methylation alteration as a biomarker has several obvious advantages, such as early detection and relative specimen stability, but only a few are currently clinically used (e.g., methylation of *MGMT* in glioblastoma, *SEPT9* in hepatocellular carcinoma, and *PITX2* in breast cancer) [33].

Very similar to BC in human, canine mammary tumor (CMT) is one of the most common cancers in female dogs [34]. Clinical and pathophysiological similarities existing between HBC and CMTs are well-documented, including the spontaneous tumor incidence, comparable onset age, hormonal etiology and the identical course of the disease [34]. Furthermore, CMT's molecular characteristics, including several subtype molecular markers such as steroid receptor, epidermal growth factor (EGF) and proliferation markers, are also similar to HBC [35]. Recently, we reported a transcriptome signature in CMT [36] and other high-throughput sequencing studies on the aspects of CMT have been reported [37,38]. However, no comprehensive genome

wide methylome profiles that are comparable to studies in HBC have been uncovered yet.

In the present study, we profiled the CMT-associated genome-wide methylation signature using methyl CpG binding domain (MBD) sequencing. In particular, altered DNA methylation in the intron region associated with CMT was comparatively investigated in both CMT and human breast cancer. Finally, we tried to show the putative function of differentially methylated regions (DMRs) in the intron region on gene expression using motif analysis with validation in additional samples.

Materials and methods

Tissue samples

Based on the methods reviewed and approved by the Seoul National University Institutional Review Board/Institutional Animal Care and Use Committee (IACUC SNU-170602-1), a total of 11 dog patients with clinically diagnosed CMT were enrolled in the present study. Tumor and adjacent normal tissue samples of spontaneously occurred canine mammary gland cancer were obtained and freshly frozen. The information for CMT dogs is provided in **Table 1.1**.

Genomic DNA isolation and MBD-sequencing

Genomic DNA was extracted from 11 pairs of CMT and adjacent normal tissues and sheared into 100-300 bp lengths using Bioruptor® Pico (Diagenode, Belgium). Methylated DNA fragments were captured by MBD-beads using the MethylMiner™ Methylated DNA Enrichment Kit (Cat# ME10025) from Invitrogen (CA, USA) according to the manufacturer's protocol (Invitrogen, Carlsbad, CA). To obtain more highly methylated DNA, MBD-captured DNA was eluted step-wise with different NaCl concentrations (200, 300, 400, 600 and 800 mM) and ethanol precipitated. After we confirmed that methylated DNA was highly enriched in the 600 and 800

Table 1.1. Information for CMT tissue samples used for MBD-seq

No.	Subtype	Cancer_ID	Normal_ID	Breeds	Sex	Age (years)
1	Simple	SC054	SN054	Miniature pinscher	FS	12
2	Simple	SC076	SN076	Cocker spaniel	FS	13
3	Simple	SC127	SN127	Poodle	F	14
4	Ductal	DC011	DN011	Bichon frise	F	12
5	Ductal	DC017	DN017	Cocker spaniel	FS	13
6	Ductal	DC070	DN070	Maltese	FS	12
7	Ductal	DC125	DN125	Schnauzer	FS	16
8	Complex	CC001	CN001	Great pyrenees	F	10
9	Complex	CC012	CN012	Dachshund	F	11
10	Complex	CC128	CN128	Shih-tzu	FS	14
11	Complex	CC132	CN132	Maltese	F	12

* F: Female / FS: Female Spayed

mM fractions using real-time PCR. We pooled the 600 and 800 mM fractions and then conducted paired-end sequencing (read length: 101bp) on the Illumina Hiseq 4000 next-generation sequencing platform (Illumina, CA, USA) after library construction using the TruSeq Nano DNA Sample Preparation Guide (Part # 15041110 Rev. D) as the manufacturer's guide.

MBD-sequencing data processing

Both per base sequence quality and per sequence quality scores were checked with FastQC v0.11.8 [39] and sequencing reads with low quality were trimmed using Trim Galore v0.5.0 [40]. Processed reads were mapped to the dog reference genome CanFam3.1 using Bowtie2 v2.3.4.3 [41] and complete BAM files were obtained after converting SAM to BAM and removing duplicated reads in Linux OS. Using MEDIPS v.1.38.0 (R Bioconductor) [42], MBD reads were calculated in every bin, dividing the whole genome into user-defined window sizes (500 bp, total 4,655,287 bins). Each read per bin was quantile normalized to reduce experimental difference, followed by an estimation of genomic CpG coverage by sequencing depth, sequencing reproducibility and enriched methylated fragments according to the number of CpGs in bins. Read counts across the total bins showed high correlation between each sample. The entire process is summarized in **Figure 1.1**.

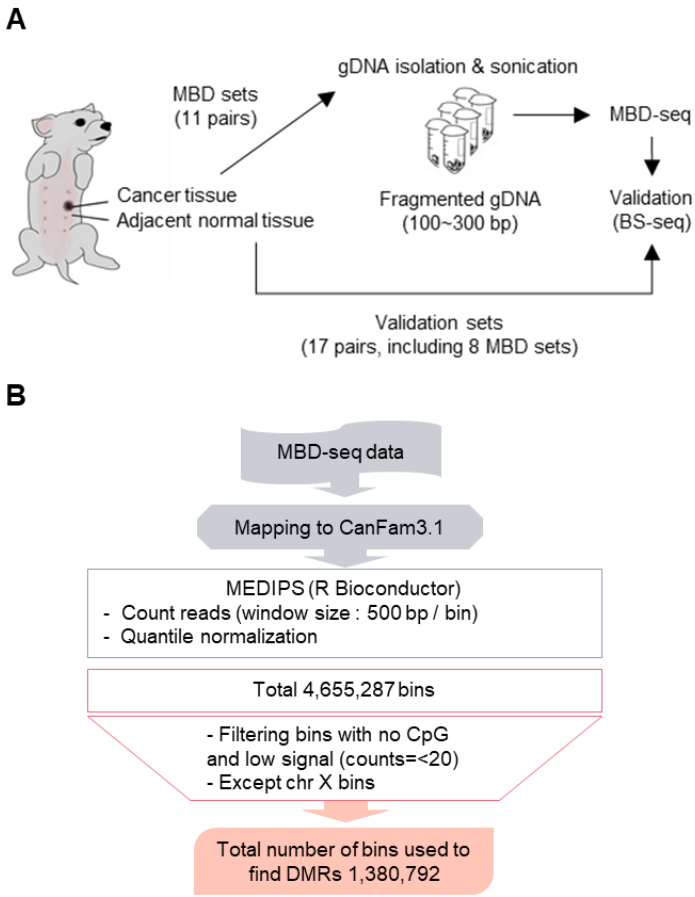


Figure 1.1. Schematic presentation of genome wide methylation profiling in CMT. A) Sample preparation for MBD-seq. B) Sequencing data preprocessing with major parameters (window size: 500bp, filtration: bins without any CG, low signal: counts ≤20, bins on Chr X).

DMR identification using LMM (Linear Mixed Model)

Bins located in chromosome X were excepted for downstream analysis, because some CMT patients were spayed females, which could affect the methylation difference on sex chromosome. Low signal bins with ~<20 counts throughout all samples and also bins with no CG dinucleotides had been removed to obtain only valuable signal peaks. Finally, a total of 1,380,792 bins were used for DMR identification. Covariance between 'CMT vs. adjacent normal' and 'between subtypes' respectively, were calculated for the entirety of the bins using R package `lme4` and we chose the upper 5% of the bins in each comparison group (between 'CMT vs. adjacent normal' and 'between subtypes') following prioritizing variance by descending order from 0 to 1. After this, we defined bins whose priority between CMT vs. adjacent normal was higher than that between subtypes as `CMT-DMRs`. Inversely, if the priority between subtypes was higher than that between CMT vs. adjacent normal, we called those bins `Subtype-DMRs`. This LMM analysis and further analyses were performed using a custom R script. *P*-values and fold changes for DMRs were obtained using `MEDIPS.meth` function based on the `edge.R` calculation method.

RNA expression

For 10 pairs of CMT dog tissues that we performed MBD-seq on in this study, RNA sequencing was also performed in a previous study and the data was obtained

from PRJNA527698 (SRA accession number: SRR8741587-SRR8741602) [36]. Data processing was conducted as mentioned above (`Materials and methods - MBD-sequencing data processing`). Using `CuffLinks`, a tool to quantitate RNA expression data and statistically identify differential expression between groups, we estimated expression levels for 32,218 genes and identified DEGs based on *p*-value ($p < 0.05$).

OncoScore

OncoScore is a tool that scores genes according to their association with cancer, based on text-mining technology using the available scientific literature in PubMed. OncoScore for DMGs with anti-correlated expression was obtained through the R package `OncoScore` (https://github.com/danro9685/OncoScore) [43].

Functional annotation

To investigate the disease enrichment analysis, we used the interactive web-based enrichment analysis tool, `Enrichr` (http://amp.pharm.mssm.edu/Enrichr/) [44,45]. Among 35 gene set libraries in Enrichr, a category of the Disease Perturbations from GEO (Gene Expression Omnibus) down was chosen to find the disease terms. We investigated the functional annotation of 7 DMG groups and searched for subtype-associated GO terms using `DAVID`, a web-based software for functional

annotation analysis [46]. Since the database of gene ontology in dog is not well established, we converted the dog Ensembl Gene IDs to human IDs using the table of human-dog gene orthologues provided by Ensembl BioMart (www.ensembl.org/biomart/martview) [47]. The functional mechanism studies for dog genes are poorly conducted. KEGG terms for CMT DMGs with *p*-values <0.05 were considered relevant.

Motif analysis

Highly enriched known motifs in hypermethylated and hypomethylated intron DMR sequences were respectively identified using the `HOMER – findMotifsGenome.pl` command. The CpG normalization option was used since genome-wide methylation changes in CMT usually occur in CpG-rich regions. The *p*-value for each motif was estimated by comparing the percentage of target sequence with motifs with the percentage of background sequence with motifs. We considered motifs relevant when the *p*-value was <0.01. After that, we found loci where the PAX5 and PAX6 motifs exist across the dog reference genome `CanFam3` (or `hg19` for human) using a motif scanning tool, `FIMO` (matched *p*-value <0.01) (http://meme-suite.org/doc/fimo.html).

Targeted BS-conversion sequencing

A total 17 pairs of CMT and adjacent normal tissue were used for validation, including the same 8 sets used in MBD-sequencing (**Table 1.2**). Bisulfite conversion was done on 500ng of genomic DNA using the EZ DNA Methylation-Lightning Kit (Zymo Research, USA). Primers were designed using MethPrimer (http://www.urogene.org/methprimer/index1.html) [48] and are listed in **Table 1.3.** After PCR, amplicons were purified from the agarose gels using the QIAquick Gel Extraction Kit (Qiagen, Germany) and directly sequenced at Macrogen Co. Ltd. (Macrogen Co. Ltd., Seoul, Korea).

Human TCGA (BRCA) expression and methylation data

RNA-sequencing and Infinium Human Methylation 450K BeadChip array were performed in various human cancer types, such as human invasive breast cancer patients, and in normal people. Wanderer (http://maplab.imppc.org/wanderer/) grants access to a large dataset and offers an interactive viewer to show expression and methylation levels for interesting genes in BRCA (data for other cancer types also provided) [49]. We could thus obtain the methylation beta value for the interesting CGs near PAX motif regions of target genes (*CDH5, LRIG1, CDH2* and *ADMA19*) and their transcription level changes in BRCA patients (wilcoxon's test).

Table 1.2. Information for CMT tissue samples used for BS-seq

No.	Subtype	Cancer_ID	Normal_ID	Breeds	Sex	Age (years)
M1	Simple	SC054	SN054	Miniature pinscher	FS	12
M2	Simple	SC076	SN076	Cocker spaniel	FS	13
M3	Ductal	DC011	DN011	Bichon frise	F	12
M4	Ductal	DC017	DN017	Cocker spaniel	FS	13
M5	Ductal	DC125	DN125	Schnauzer	FS	16
M6	Complex	CC001	CN001	Great pyrenees	F	10
M7	Complex	CC128	CN128	Shih-tzu	FS	14
M8	Complex	CC132	CN132	Maltese	F	12
V1	Simple	SC094	SN094	Shih-tzu	FS	10
V2	Simple	SC165	SN165	Poodle	F	5
V3	Simple	SC200	SN200	Spitz	F	9
V4	Simple	SC205	SN205	Shih-tzu	FS	12
V5	Complex	CC088	CN088	Schnauzer	FS	9
V6	Complex	CC149	CN149	Cocker spaniel	F	11
V7	Complex	CC151	CN151	Dachshund	FS	14
V8	Complex	CC166	CN166	Dachshund	F	~<1
V9	Complex	CC221	CN221	Bichon frise	F	12

* M1~M8 are overlapped samples used for MBD-sequencing

* V1~V9 are samples used only for BS-seq (Validation set)

* F: Female / FS: Female Spayed

Table 1.3. Primers designed for BS-conversion PCR

Target Gene	Target Locus	Strand	Sequences (5'→3')
LRIG1	chr20:25007595-25007877	Forward	GAAGGGTGGGTGATTTTTATTAGATA
		Reverse	ACCAAAACTTTTCTCTTCTTTCTAACTC
CDH5	chr5:82863641-82864024	Forward	GGTTTGTTTTTTAAGAATGGTTTTT
		Reverse	CCACCACAAAACCTACCTATCTAC
ADAM19	chr4:52717697-52718030	Forward	GTATTAGGTATTAAAGTGGGGG
		Reverse	AAAAAACAATCAATATCTCAAATACCCT
CDH2	chr7:60865092-60865487	Forward	ACTTAAGGTTTATGAGTGAAGA
		Reverse	CAAAACTACTAATTTCATTTAACA

Statistical Analysis

To estimate the methylated CpG level between CMT and adjacent normal tissues, we calculated the ratio of C/(C+T) from the BS-sequencing data. For validating methylation changes between them in the target motif DMR regions, statistical significance was assessed on *p*-values obtained by paired t-test using R basic command.

Results

Genome-wide methylation was profiled in 11 pairs of CMT and adjacent normal tissues via MBD-sequencing

Eleven pairs of CMT and adjacent normal tissues consisting of three subtypes, simple, ductal and complex carcinoma, were subjected to MBD-sequencing (**Figure 1.1A and Table 1.1**). The statistic information, including the number of reads, Q20 and 30 scores for all the raw sequence data and enrichment scores, and the CG coverage for the processed sequence data generated in this study showed good quality (**Table 1.4**). From a total of 4,655,287 bins (500 bp in size), 1,380,792 high quality bins were obtained by filtration of no CpG, low signals (counts ≥20) and bins on the X chromosome (**Figure 1.1B**). Even signal distribution across CMT and adjacent normal in the 11 samples was representatively depicted within the genomic region (Chr 1:18,286,500-19,222,630, ~100 Kb) by integrative genomic viewer (IGV) [50] with peak and annotation files. Differentially methylated regions (DMRs), shown in yellow, were distributed similarly on CpG islands and tended to be enriched in gene regions (**Figure 1.2**). The quality of MBD-enrichment was checked according to the coverage of CpGs in the dog genome. Bins with high signal depth (>5X) covered 45~55% of the dog genome, indicating that methylated DNA was successfully enriched by MBD not only from promoter regions

Table 1.4. Quality check for MBD-seq data

	Raw data QC					MBD-seq QC	
Sample ID	Total read bases (bp)	Total reads	GC (%)	Q20 (%)	Q30 (%)	enrichment. score.relH*	enrichment. score.GoGe**
SN054	6,266,888,624	62,078,552	54.13	97.34	93.16	3.643870699	2.024206695
SN076	8,074,844,152	79,948,952	55.11	97.07	92.62	4.097720852	2.120520181
SN127	8,297,964,868	82,158,068	54.87	97.31	93.28	3.776880884	2.029195266
SC054	6,162,333,745	61,042,716	57.38	97.36	93.07	4.621383356	2.242049714
SC076	7,055,536,194	69,856,794	53.44	97.36	93.22	3.438118038	1.951425376
SC127	6,406,478,884	63,430,484	53.67	97.48	93.68	3.412195461	1.948910435
DN011	6,979,519,920	69,136,566	51.79	97.31	93.3	3.121841121	1.89731437
DN017	7,717,562,106	76,411,506	53.48	97.37	93.29	3.369929515	1.92720443
DN070	7,297,507,348	72,252,548	52.96	97.13	92.72	3.273069456	1.915837506
DN125	7,626,912,586	75,513,986	53.38	97.32	93.2	3.44385977	1.962578118
DC011	6,592,741,947	65,307,016	52.49	97.1	92.81	3.204819395	1.896182073
DC017	7,645,841,400	75,701,400	52.81	97.2	93.01	3.328160347	1.923816406
DC070	7,390,602,280	73,174,280	52.29	97	92.56	3.1279579	1.868868146
DC125	7,898,553,904	78,203,504	52.94	97.38	93.34	3.406331475	1.969939524
CN001	7,805,241,418	77,279,618	56.23	97.41	93.31	4.122384585	2.11373002
CN012	7,605,706,626	75,304,026	53.17	96.98	92.64	3.408393805	1.942141539
CN128	7,703,398,068	76,271,268	52.22	97.19	93.07	3.194818603	1.8869285

CN132	6,843,561,636	67,758,036	56.45	97.56	93.81	4.313466538	2.153527831
CC001	8,186,238,870	81,051,870	54.11	96.52	91.16	3.65260293	2.010762082
CC012	8,764,742,428	86,779,628	51.07	96.8	92.3	2.883893082	1.804174017
CC128	7,143,367,612	70,726,412	52.4	97.44	93.54	3.183262184	1.886515542
CC132	7,043,265,502	69,735,302	55.35	97.24	93.11	3.923643991	2.049587669

*enrichment.score.relH
: the relative frequency of CpGs within the regions / the relative frequency of CpGs within the reference genome
**enrichment.score.GoGe
: the o/e ratio of CpGs within the regions / the o/e ratio of CpGs within the reference genome

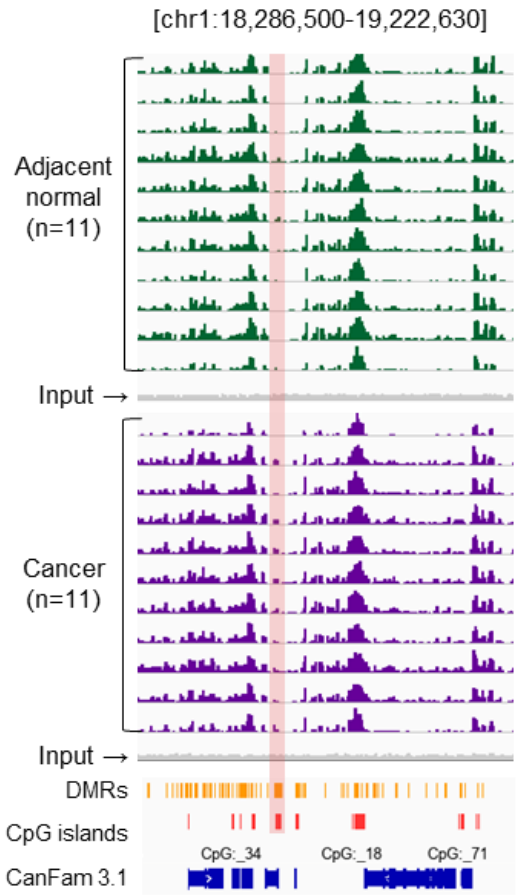


Figure 1.2. Visualizing methylation peaks using processed MBD-seq from 11 pairs of CMT and adjacent normal tissues. Overall sequencing quality is visualized by IGV showing DMRs (yellow) CGI (red) and Reference genes (blue). Methylation peaks are colored in 11 cancer tissues as purple and adjacent normal tissues as green. The region with high density of DMRs is highlighted by the red box.

but also from various regulatory regions, including both genic and intergenic regions (**Figure 1.3**). The methylation profiles were analyzed further by focusing on the DMRs in intergenic regions for the tissue origin of CMT subtypes and the DMRs in genic regions for CMT-enriched methylation. Gene ontology (GO) enrichment analysis and OncoScore [43] were employed to elucidate the functional linkage between differential methylation and gene regulation. Additionally, the transcription factor (TF) binding motifs on the subtype-enriched DMRs were investigated. The CMT-enriched methylation signatures and putative regulation were furthermore comparatively investigated in HBC datasets to show how epigenetically similar these two diseases are. The analytical scheme was depicted in **Figure 1.4.**

Linearized mixed model (LMM) successfully clustered DMRs between CMT and adjacent normal, and among subtypes

To determine differential methylated bins as variables that respond to CMT as well as each subtype, linearized mixed model (LMM) was employed and two different thresholds, top 5% and top 10% bins based on standard variation (SD) that corresponds to p-value <0.01 and p-value <0.05, respectively, were used to obtain DMRs. A total of 137,755 bins (68,741 for CMT DMRs and 69,014 for subtype DMRs) were determined as strict DMRs (5%) of either CMT or across subtypes (**Figure 1.5A**). Principal component analysis (PCA) using the DMRs successfully

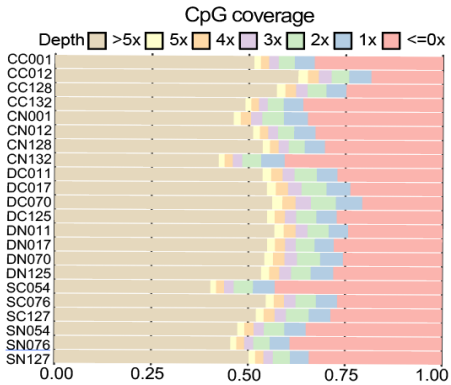


Figure 1.3. The CpG coverage of genome wide DNA sequence patterns. High quality signals (depth >5X) cover more than 50% of the canine genome in 22 samples.

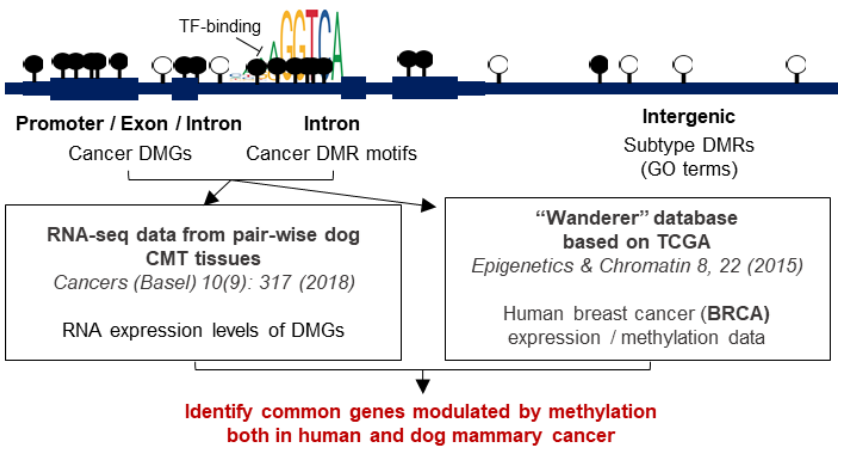


Figure 1.4. Analytical strategies. Investigating both intergenic and genic regions where subtype-DMRs and CMT-DMRs exist. Additionally, CMT transcriptome data set and BRCA expression and methylation data in TCGA were accompanied for further analysis.

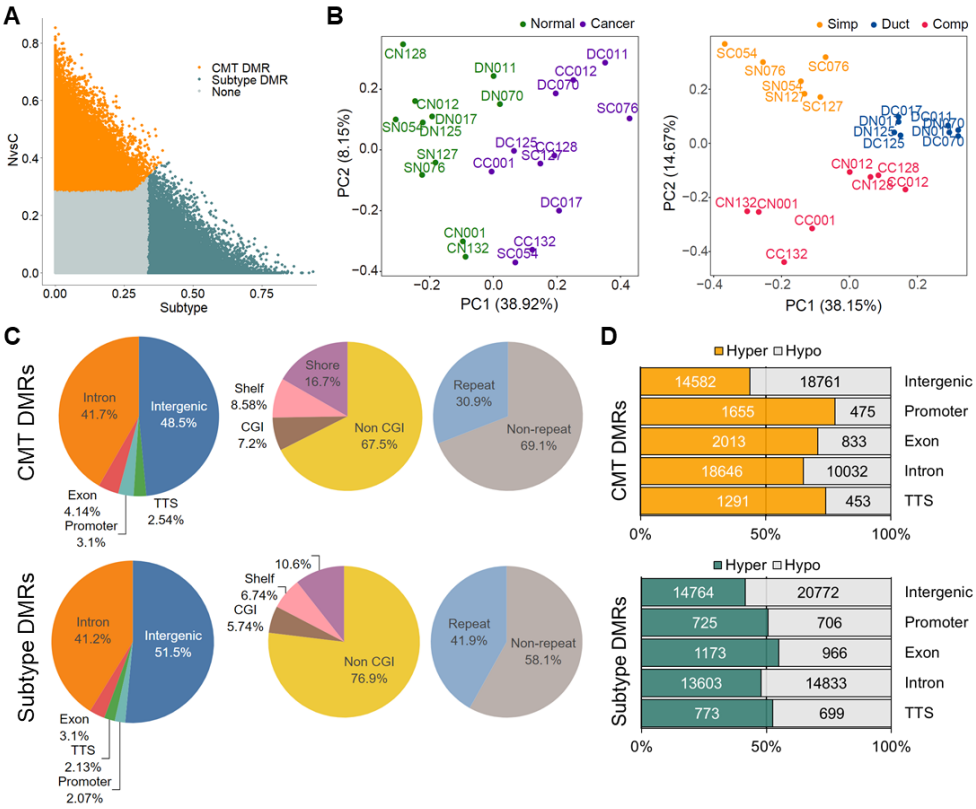


Figure 1.5. Identification of differentially methylated regions (DMRs) among the three CMT subtypes and between CMT and adjacent normal. A) LMM separated CMT-DMRs and Subtype-DMRs. B) PCA analysis using CMT-DMRs and Subtype-DMRs. CMT-DMRs successfully divides adjacent normal and CMT and also Subtype-DMRs into simple, ductal and complex types. C) Genomic distribution of CMT-DMRs and Subtype-DMRs. Distribution between genic and intergenic regions, CGI and non-CGI, and repeat and non-repeat. D) Hyper- and hypo methylation profiles in CMT-DMRs and Subtype-DMRs. Colored region (orange and blue green): hypermethylation, gray: hypomethylation.

separated 22 specimens with multiple variances (CMT and adjacent normal and three different subtypes: simple, ductal and complex) into corresponding groups (**Figure 1.5B**). The sum of PC1 and PC2 in both CMT- and Subtype-DMRs represented more than 50% of the total DMRs. Although no clear difference was found in the comparison of genic features consisting of CMT- and Subtype-DMRs, the Non-CGI (CpG island) region showed a clear difference between CMT (67.5%)- and Subtype (76.9%)- DMRs that might occur in the alteration of repeat element regions (30.9% in CMT-DMR/ 41.9% in Subtype-DMR). On the contrary, the proportion of CGI (7.2%) and Shore (16.7%) regions encompassed in CMT-DMRs was higher than in Subtype-DMRs (CGI (5.74%) and Shore (10.6%)) (**Figure 1.5C**). Interestingly, methylation profiles (hyper- and hypo-methylation) showed a distinct difference between CMT- and Subtype -DMRs, although, no significant difference was seen in genome wide methylation distribution. Of note, methylation patterns were clearly biased in genic regions of CMT-DMRs. Approximately 66% of CMT-DMRs in the genetic regions were hypermethylated, while only 45% of DMRs in the intergenic region were hypermethylated. This bias was not seen in Subtype-DMRs, which indicates that the bias is not due to the MBD-sequencing (**Figure 1.5D**). This biased genic hypermethylation in CMT fits the general features of higher methylation of genic region in cancer tissues and is similar to a previous report in human BC by Ball et al. (Ball et al., 2009)

Gene ontology (GO) enrichment and pathway analysis using DMRs on both genic and intergenic regions-fittingly represented the functional relationship between DMRs and CMT as well as subtypes

Extraordinary hypermethylation throughout genic regions including promoter, exon, intron and TTS in CMT was shown (**Figure 1.5D**). On the other hand, differential methylation on intergenic regions where enhancers or silencers exist contributes to the tissue-type specificity [51]. We first performed hierarchical clustering and heatmap plotting using the genic regions of CMT-DMRs (**Figure 1.6A**). Hypermethylation was more enriched in CMT than adjacent normal, parallel to **Figure 1.5D** and what was previously known (**Figure 1.6A**). Subsequently, OncoScore [43], Functional annotations and Gene ontology (GO) [45] enrichment analysis were performed with the list of CMT-DMGs (**Figure 1.6B, D**) to investigate the functional linkage between DMGs and the molecular pathophysiology of CMT. As expected, many DMGs that were hypermethylated and down-regulated in CMT including *TP63, LIFR, PLA2G16, LRIG1, STAT5A* and *AKAP12* and has been known as tumor suppressors, were identified from high scoring (OncoScore >50) CMT-DMRs (**Figure 1.6B**). On the contrary, some oncogenes including *WT1, TFPI2* and *ETV1* were also found from hypomethylated and up-regulated DMGs. The methylation of 4 representative canine genes and their orthologous human genes, identified as three hypermethylated tumor suppressors (*TP63, LIFR* and *FOLH1*) and one

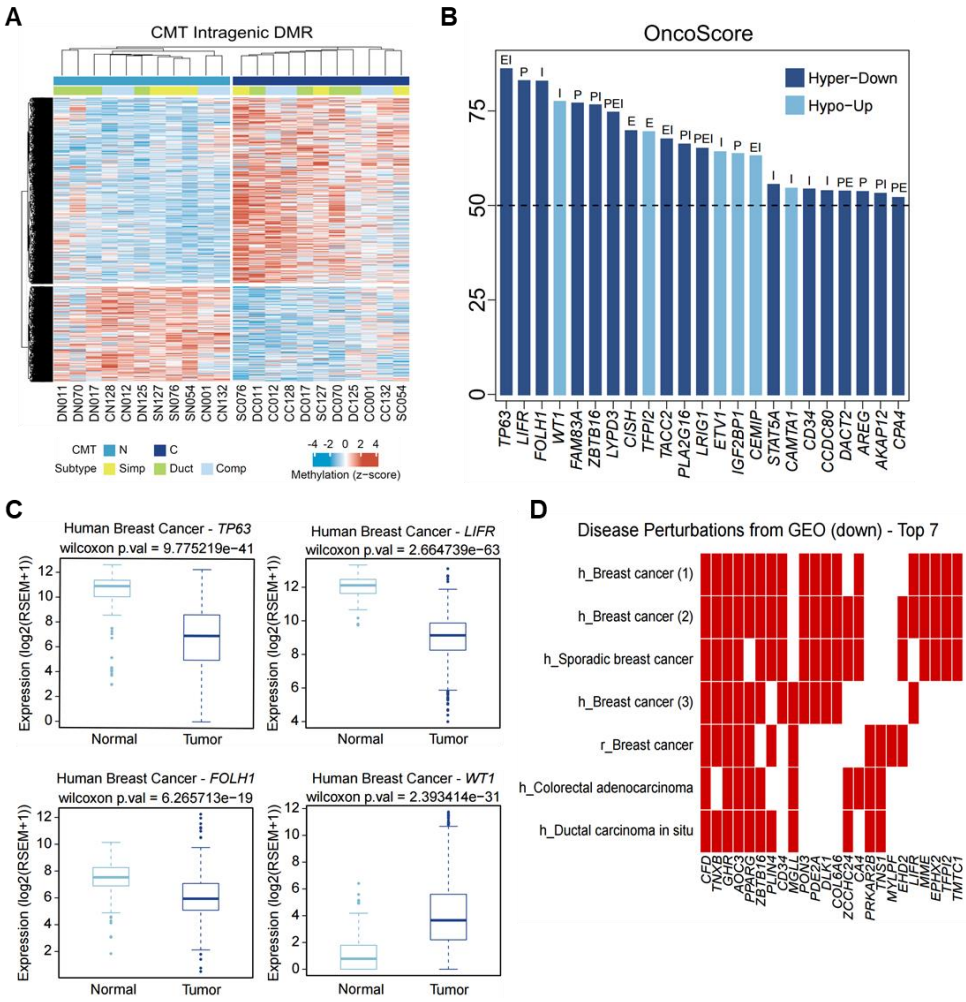


Figure 1.6. Functional annotation of CMT-DMGs. A) Hierarchical clustering of CMT-DMGs separates 11 adjacent normal (light blue) and 11 CMT (dark blue) independent of subtypes (simple- yellow, ductal- green and complex- blue). Methylation levels were z-scored and are indicated by blue (hypo) and red (hyper) scale. B) OncoScore of 224 CMT-DMGs were measured and those with a score greater than 50 are depicted. Dark blue indicates hypermethylated DMGs and down-

regulated in RNA-seq data and light blue indicates hypomethylated DMGs and up-regulated in RNA-seq data. C) Box plot shows the expression level of the top 4 orthologous genes from the TCGA database ranked by OncoScore in normal (light blue) and human invasive breast cancer (dark blue). D) CMT-DMGs were clustered into the library of Disease perturbations from GEO (down). The top 7 terms are composed of breast cancer related terms. h: human, r: rat, m: mouse, (1) Breast cancer C0006142 rat GSE1872, (2) Breast cancer DOID-1612 human GSE26910, (3) Sporadic breast cancer DOID-8029 human GSE3744, (4) Colorectal adenocarcinoma DOID-0050861 human GSE24514, (5) Tendonopathy 971 human GSE26051, (6) Neurological pain disorder C0423704 rat GSE15041 and (7) Ductal carcinoma in situ DOID-0060074 human GSE21422.

hypomethylated oncogene (*WT1*) in CMT, showed an anti-correlation with gene expression between normal and cancer in both dogs and humans (**Figure 1.6C** and **Figure 1.7**). In addition, GO analysis with the Disease Perturbations from the GEO library revealed that CMT-DMGs were frequently enriched in the list of down-regulated genes from various types of cancers including BC (Breast Cancer C0006142 rat GSE1872 sample 63 (*p*-value = 1.4E-16), breast cancer DOID-1612 human GSE26910 sample 602 (*p*-value = 9.81E-13), sporadic breast cancer DOID-8029 human GSE3744 sample 979 (*p*-value = 2.49E-11)) (**Figure 1.6D**). Furthermore, based on the methylation profiles in the intergenic regions of Subtype-DMRs, the ductal subtype was distinctively separated from the simple subtype, while the complex subtype was located in between (**Figure 1.8A**). This result may indicate that the cell type components are shared by the simple and complex subtypes of CMT but not by the ductal subtype. Hierarchical clustering was performed using the intergenic Subtype-DMRs (**Figure 1.8A**) and the nearest genes from the intergenic DMRs were found and processed with GO analysis. The list of genes near intergenic Subtype-DMRs were presented in **Table S8**. The top 5 GO_biological process (BP) and GO_cellular component (CC) terms found in Subtype-DMRs indicated that diverse processes were enriched in each subtype. Of note, simple and complex subtypes shared some biological processes, such as extracellular matrix organization (GO:0030198, *p*-value = 6.79E-04 (simple), *p*-value = 2.32E-03 (complex)) and cellular response to tumor necrosis factor (GO:0071356, *p*-value = 1.25E-03 (simple), *p*-value = 4.56E-03 (complex)), but all terms were unique in the ductal

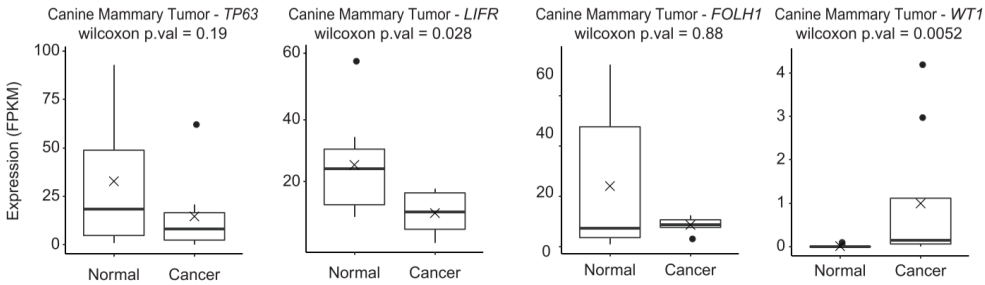


Figure 1.7. The expression level of the top 4 orthologous genes ranked by OncoScore in canine mammary tumor. Box plots showed the expression level of four genes (TP63, LIFR, FOLH1 and WT1) in adjacent normal (n=8) and paired CMT tissues (n=8). Expression values are presented by FPKM calculated from RNA-sequencing data. Statistical *p*-value was calculated by Wilcoxon's test.Functional association of DMGs.

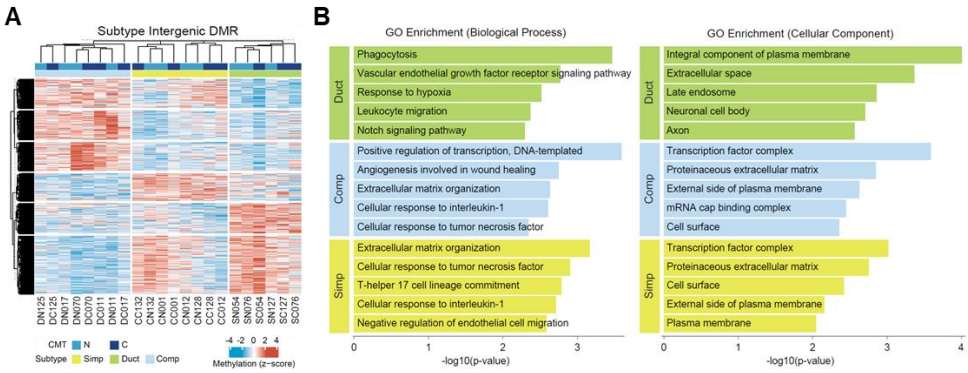


Figure 1.8. Functional annotation of Subtype-DMGs. A) Hierarchical clustering of Subtype-DMGs. B) GO enrichment analysis in biological process (left) and cellular component (right). Duct: ductal, Comp: complex and Simp: simple subtype. Length of bar represents -log(*p*-value).

subtype, such as vascular endothelial growth factor receptor signaling pathway (GO:0048010, *p*-value = 1.69E-03). Similarly, in GO_CC, 4 out of 5 terms were also common in simple and complex subtypes whereas all 5 terms in the ductal subtype were unique (**Figure 1.8B**). This coincides with the Hierarchical clustering in **Figure 1.8A**. Substantial GO analysis using the nearest gene from intergenic CMT-DMRs as well as genic Subtype-DMGs and pathway analysis using intergenic Subtype-DMRs were performed. In brief, no relevant terms to either cell-types or cancer were retrieved.

Aberration in intron methylation is associated with cancer

A total of 10,583 CMT-DMGs were divided into 7 sub-groups based on the distribution of DMRs (**Figure 1.9A**). More than 60% of DMGs, consisting of 6,745 genes, harbored DMRs only in the intron region, whereas 977 and 819 genes were identified with DMRs in only promoter and exon regions, respectively. A greater amount of intronic DMRs than either exonic or promoter DMRs could have been expected due to the large discrepancy in chromosomal coverage among the intron (26%), exon (1.5%) and promoter (<1%) regions. Indeed, CMT-DMRs in the exon and promoter regions account for 22% and 17% of the total DMRs, respectively. This is higher than expected based on the coverage of the exon and promoter regions in the genomic sequence (less than 2%). This may mean that more CpG enrichment was done by MBD-seq in these areas (**Figure 1.9A**).

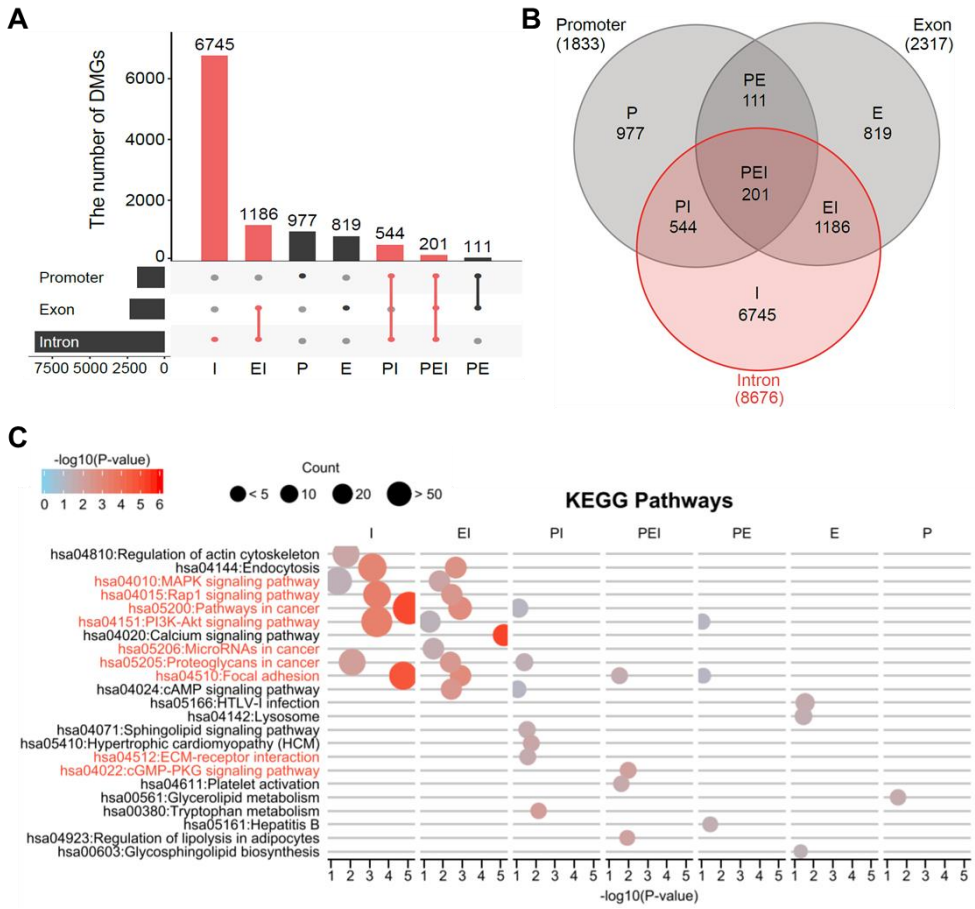


Figure 1.9. Intron DMRs may associate with cancer-related genes. A) The DMGs are catagorized into 7 groups based on the combination of the DMR's genic loci. I: intron only, EI: exon+intron, P: promoter only, E: exon only, PI: promoter+intron, PEI: promoter+exon+intron, PE: promoter+exon. Red color indicates DMGs containing intron DMRs. B) Venn diagram differentially presents intron DMRs (red) in 7 groups. C) KEGG pathway analysis with intron DMRs shows cancer-related pathways are highly enriched in I and EI group.

The most interesting finding was that all terms associated with cancer in the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis were enriched in DMRs that included intron DMRs such as intron only (I), exon+intron (EI), promoter+exon+intron (PEI), and promoter+intron (PI) (**Figure 1.9B, C**). Not only the term of 'pathways in cancer (hsa05200)' but also 'microRNAs in cancer (hsa05206)', 'proteoglycans in cancer (hsa05205)', 'PI3K-Akt signaling pathway (hsa04151)', *etc.*, which are associated with cancer and cancer pathophysiological characteristics, were highly enriched in intron only DMGs followed by EI and PI groups (**Figure 1.9C**). However, KEGG terms such as 'HTLV-1 infection (hsa05166)', 'Neuroactive ligand-receptor interaction (hsa04080)' and 'Lysosome (has04142)' that are extrinsic to cancer and CMT were enriched in DMGs that excluded intron DMGs such as the promoter only (P), exon only (E), and promoter+exon (PE) groups (**Figure 1.9C**). Considering that intronic regions comprise a large portion of the genome, we counted the number of genes enriched in the 'hsa05205: Pathways in cancer' term from a group of 530 genes, and the proportion for each group was calculated (Data not shown). The percentage of cancer-related DMGs containing intron DMRs was 22.85% (I: 5.34%, EI: 7.80%, PI: 6.70%, PEI: 3.01%), which is higher than 17.27%, the percentage of cancer-related DMGs with promoter DMRs (P: 3.51%, PI: 6.70%, PE: 4.05%, PEI: 3.01%). Consequently, these results indicate that intron methylation may have important regulatory functions that are associated with CMT. It has been reported that intron CpG methylation might be associated with gene expression in human cancer. For

instance, the methylation of the first intron of the *EGR2* gene, known as a tumor suppressor, affects the recruitment of proteins required for transcription [52] and anti-tumorigenic *PMP24* gene is silenced by the intronic single CpG methylation in prostate cancer cells [53].

Altered CG methylation surrounding transcription factor binding motifs is an important epigenetic regulation in CMT

To investigate enriched CMT-responsible transcription factor (TF) binding motifs, intron DMRs were leniently extracted from the upper 10% of covariance in an LMM analysis (mean *p*-value <0.05, **Figure 1.10A**). The list of the top 10% of CMT-DMRs was also able to separately group cancer and adjacent normal (**Figure 1.10B**). According to the alteration of methylation, a total of 56,253 intron-DMRs were obtained and subsequently divided into hyper- (36,401) and hypo- (19,852) methylated intron DMRs in CMT, then subjected to motif analysis using HOMER v4.11 [54]. Motif analysis revealed that 10 putative motifs, including PAX5, USF1, ZFX and SREBF1, were enriched in hypermethylated intron DMRs, while 6 motifs, including CREB1, ELK1, PAX6 and ELK4 motifs, were enriched in hypomethylated intron DMRs. These motifs harbor CG nucleotides the methylation of which may influence protein binding activity [6]. We indeed focused on two PAX motifs, PAX5 and PAX6, that have been known as tumor suppressive and oncogenic, respectively [55-58]. Additionally, Kaplan-Meier plot [59,60] showed breast cancer patients

with lower PAX5 expression live shorter than those with higher, while the survival rate of patients with higher PAX6 expression decreased compared to those with lower expression (**Figure 1.11**). It was expected that these two genes would have reverse effects in breast cancer. PAX5 and PAX6 motifs, respectively designated by 16 bp and 20 bp consensus nucleotide sequences (PAX5 – GCAGCCAAGCGTGACC, PAX6 – NGTGTTCAVTSAAGCGKAAA), were significantly enriched in each DMR group (PAX5 *p*-value:1E-9, PAX6 *p*-value: 1E-3) (**Figure 1.12A, B**). An enriched heatmap successfully visualized the enrichment of hyper- and hypo-methylation signals in the 5 kb surrounding PAX5 and PAX6 motifs, respectively (**Figure 1.12C, D**). We then investigated putative target genes that harbor hypermethylated PAX5 and PAX6 motifs in their intron regions (Data are not shown). Hypermethylation in the intron DMRs of the PAX5 motifs of CMT, relative to that in adjacent normal, was visualized in the representative genes, *CDH5* and *LRIG1*, by IGV (**Figure 1.12E**). On the other hand, hypomethylation related to PAX6 was found in the *CDH2* and *ADAM19* genes (**Figure 1.12F**). All of these target genes, hyper- and hypomethylated in CMT, were reversely correlated to gene expression. RNA expression levels of the candidate genes, were obtained from our previous transcriptome data [36] and an anti-correlation was shown by box plot (**Figure 1.12G, H**).

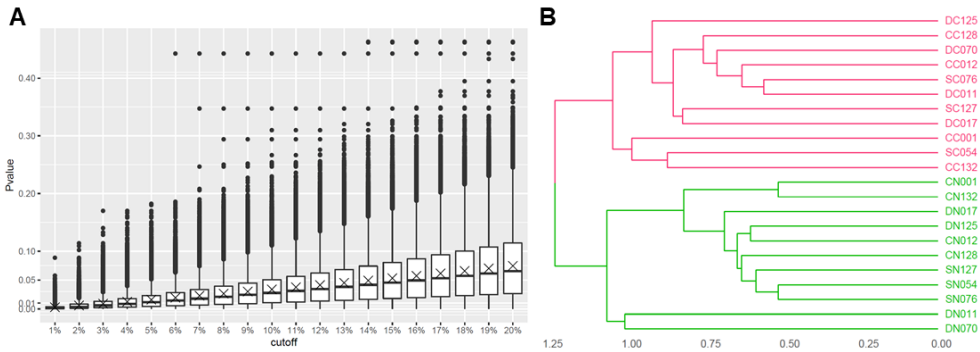


Figure 1.10. Adjust thresholds to select distinguished CMT-DMRs for intronic motif analysis. A) *P*-values for each DMRs extracted using serial cutoff manner (upper 1~20%), B) Dendrogram for 22 cancer and adjacent normal tissue samples supervised cancer groups from normal when we identify CMT-DMRs at cutoff 10% in linear mixed model (LMM).

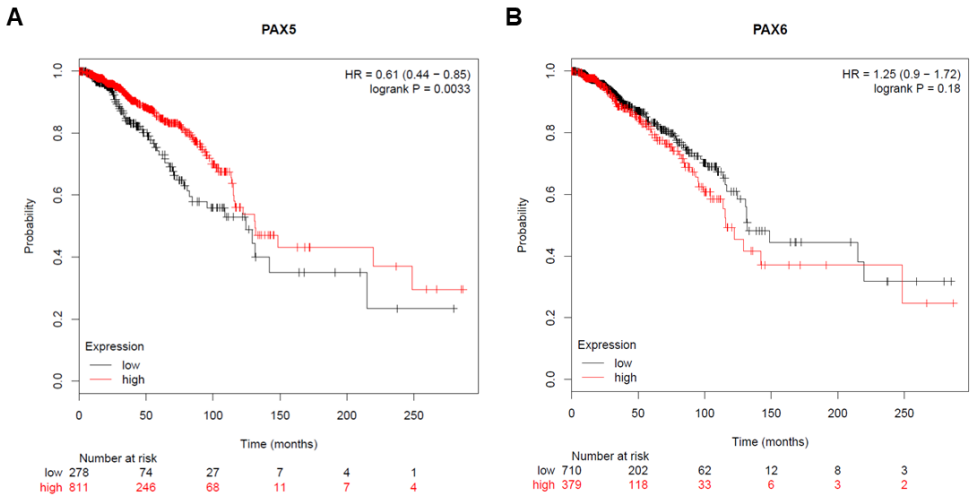
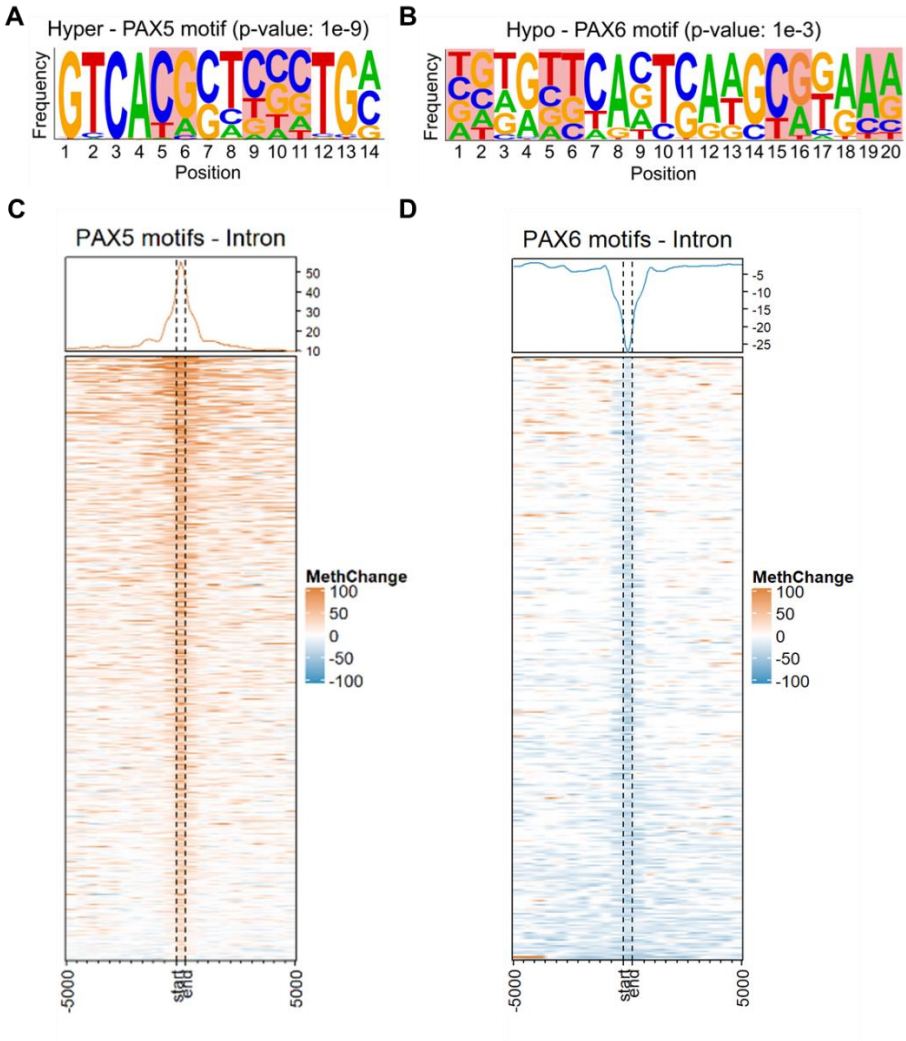


Figure 1.11. Kaplan-Meier plots showed PAX5 and PAX6 expression reversely effect on the survival rate of breast cancer patients. Survival rates depends on A) PAX5 and B) PAX6 expression are shown.


A
Hyper - PAX5 motif (p-value: 1e-9)
Frequency
Position
B
Hypo - PAX6 motif (p-value: 1e-3)
Frequency
Position
C
PAX5 motifs - Intron
MethChange
100
50
0
-50
-100
-5000
start
end
5000
D
PAX6 motifs - Intron
MethChange
100
50
0
-50
-100
-5000
start
end
5000

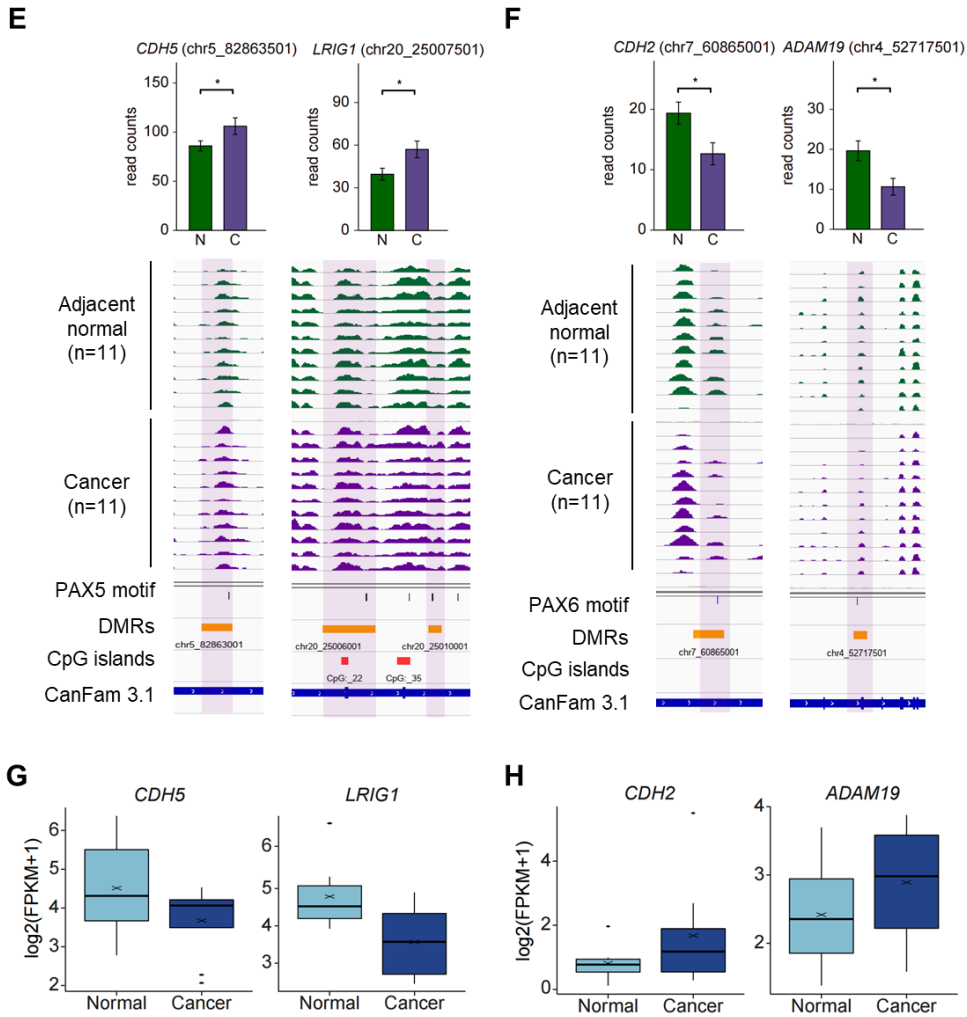


Figure 1.12. PAX motifs are enriched in hyper- and hypo-methylated intron DMRs. Consensus motif sequence and sequence frequency of A) PAX5 and B) PAX6 motif. CGs on the motifs are highlighted with red. Accumulated heatmaps present 5kb up-and downstream regions of C) PAX5 and D) PAX6 motifs. Hyper- (orange) and hypo- (blue) methylation. E and F) Differential methylation peaks between 11 adjacent normal (green) and 11 cancer (purple) samples visualized with

motif loci, DMRs, CGI and gene structure annotations. *P*-values for each DMR were generated by paired t-test. The level of candidate gene expression (log_2(FPKM+1) of G) *CDH5* and *LRIG1*, and H) *CDH2* and *ADAM19* in adjacent normal (light blue) and cancer (dark blue),

Validation of intron DMRs and their anti-correlation to gene expression

The methylome signature in CMT identified by MBD-sequencing was validated in both the 8 pairs of specimens originally subjected to high-throughput sequencing and 9 additional validation sets. Bisulfite genomic DNA conversion followed by PCR was performed in the pairs of CMT and adjacent normal samples to obtain a fine map of intron methylation surrounding PAX5 motif regions of candidate genes. Primers used in BS-conversion PCR and sequencing are listed in **Table 1.3.** Overall, a hypermethylated intron was confirmed in two candidate genes that included the PAX5 motif, *CDH5* and *LRIG1*, with box plots showing the DNA methylation profiles of the intron DMRs of genes (**Figure 1.13**). As for the *CDH5* and *LRIG1* genes, respectively, a total of 16 CGs and 7 CGs surrounding PAX5 motifs, were tested in 14 and 17 pairs of CMT and adjacent normal samples. Of the 16 CGs tested in the 1st intron region of *CDH5*, 12 showed significant hypermethylation (**Figure 1.13A**, upper panel). Unexpectedly, the PAX5 motif was located on the 14th and 15th CGs where no significant difference was found (**Figure 1.14A**). Pairwise comparison of each CG's methylation between CMT and adjacent normal showed significant hypermethylation. In the intron-DMR tested region of *LRIG1*, all CG loci tended to show hypermethylation in CMT and one CG locus (1st CG, *p*-value = 0.019, **Figure 1.14B**) among them showed a significant difference (**Figure 1.13A**, lower panel). In addition, differential intron methylation of *CDH5* was clear in all three CMT subtypes but showed the best result in the ductal subtype (*p*-value = 3.9E-13).

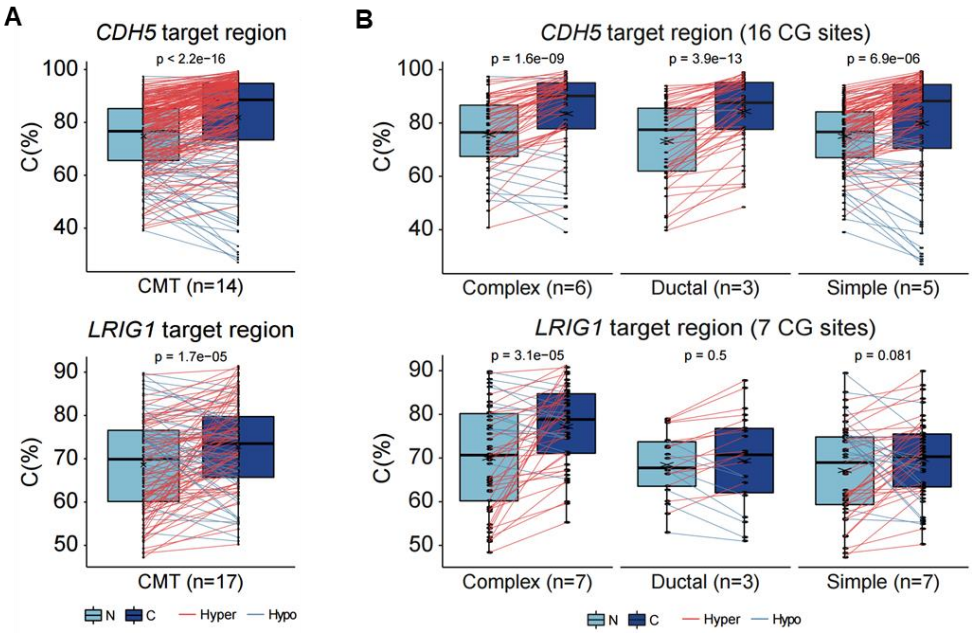


Figure 1.13. PAX motifs are enriched in hyper- and hypo-methylated intron DMRs. Validation of intron hypermethylation in the candidate genes, CDH5 and LRIG1. A) Comparison of overall methylation states in the surrounding regions of the intronic PAX5 motif in CDH5 and LRIG1 genes. Methylation was measured by the ratio of cytosine on each CG site. Red lines between CMT and adjacent normal indicate hypermethylation, while blue lines indicate hypomethylation. N: adjacent normal, C: CMT. Statistical *p*-value was calculated by paired t-test. B) Differential methylation is depicted in three separated CMT subtypes.

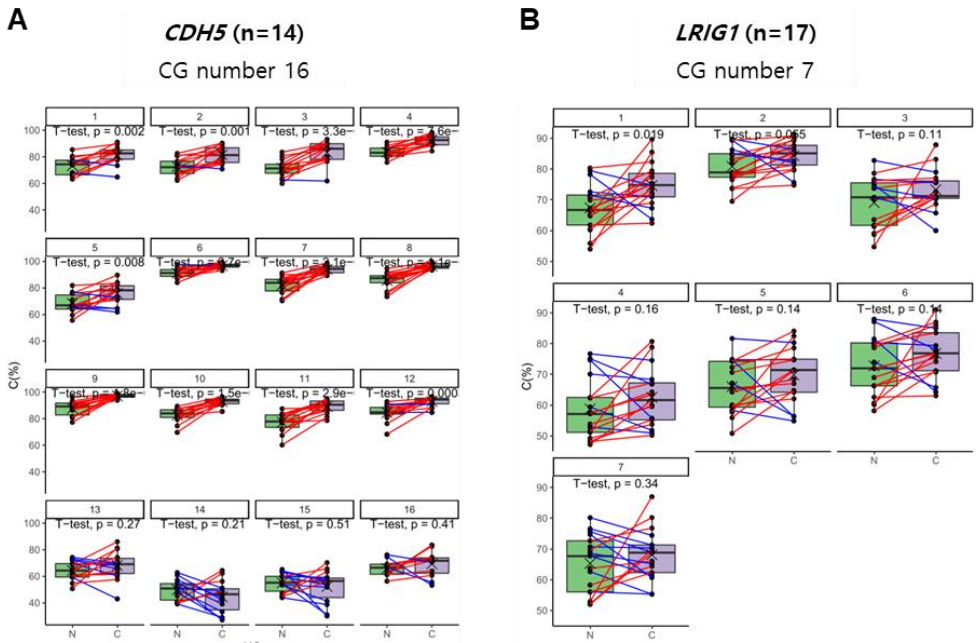
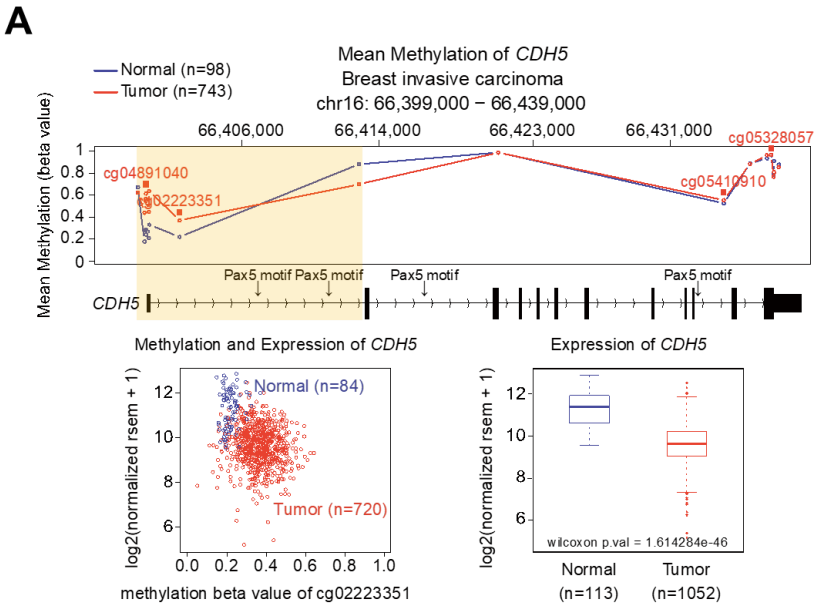


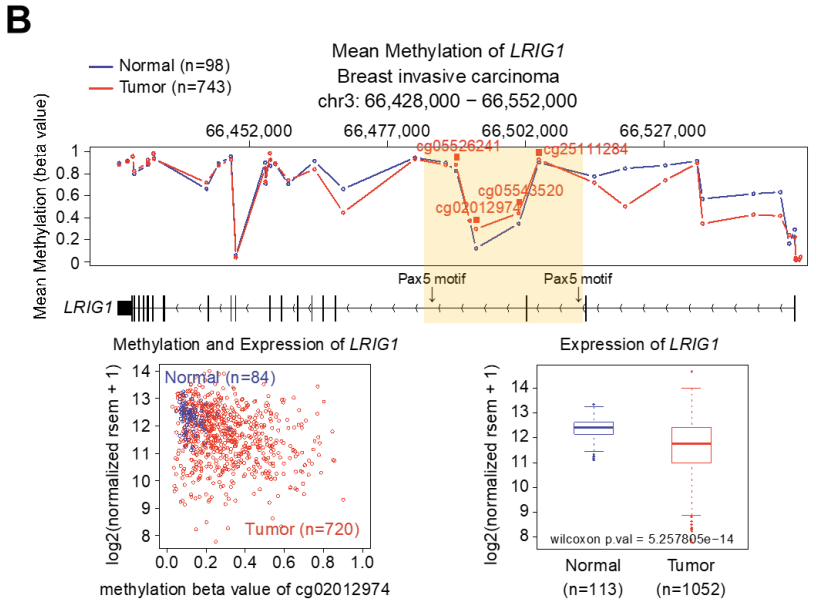
Figure 1.14. Validation of individual CG methylation around PAX5 motif regions in CDH5 and LRIG1 genes. Paired t-test for individual CG in A) CDH5 and B) LRIG1 intronic PAX5 motif region. Percentage of methylated cytosine (C (%)) is represented by (C/C+T) * 100. Red lines between CMT and adjacent normal indicate hypermethylation, while blue lines indicate hypomethylation (N: adjacent normal, C: CMT). Statistical *p*-value was calculated by paired t-test.

The differences in LRIG1 intron methylation was more distinct in the complex subtype (*p*-value = 3.1E-05) than in the other subtypes (**Figure. 1.13B**). These results suggest that hypermethylation of these two intron regions can be useful candidate epigenetic markers for CMT as well as subtypes.

CMT-enriched differential intron methylation and its anti-correlation with gene expression was conserved in human breast cancer

To validate our CMT-enriched methylome signature findings to human breast cancer (HBC), we investigated the consistency of the aberrations of candidate gene methylation and RNA expression between CMT and HBC. The methylation status and expression profiles of 4 representative candidate genes in HBC was surveyed using the Wanderer database [49]. We determined locally corresponding CG sites and introns of the human orthologous genes from the breast cancer methylome data. Methylation levels were regionally dynamic within a target gene and there were some CGs differentially methylated between normal and HBC populations (**Figure 1.15**, top panels of mean methylation). The scatter plots for *CDH5* and *LRIG1* consisting of hypermethylated intron motifs depicted the trend of increased methylation and decreased gene expression in HBC when compared to normal and thus resulted in normal being represented by the blue dots located in the top-left and HBC being represented by the red dots located in the bottom-right (**Figure 1.15A, B**). On the contrary, *CDH2* and *ADAM19* showed the opposite


A
Mean Methylation of CDH5
Breast invasive carcinoma
chr16: 66,399,000 − 66,439,000
Normal (n=98)
Tumor (n=743)
Mean Methylation (beta value)
66,406,000
66,414,000
66,423,000
66,431,000
cg04891040
cg02223351
cg05410910
cg05328057
Pax5 motif
Pax5 motif
Pax5 motif
Pax5 motif
CDH5
Methylation and Expression of CDH5
Normal (n=84)
Tumor (n=720)
log2(normalized rsem + 1)
methylation beta value of cg02223351
Expression of CDH5
wilcoxon p.val = 1.614284e-46
Normal (n=113)
Tumor (n=1052)


B
Mean Methylation of LRIG1
Breast invasive carcinoma
chr3: 66,428,000 − 66,552,000
Normal (n=98)
Tumor (n=743)
Mean Methylation (beta value)
66,452,000
66,477,000
66,502,000
66,527,000
cg05526241
cg25111284
cg05543520
cg02012974
Pax5 motif
Pax5 motif
LRIG1
Methylation and Expression of LRIG1
Normal (n=84)
Tumor (n=720)
log2(normalized rsem + 1)
methylation beta value of cg02012974
Expression of LRIG1
wilcoxon p.val = 5.257805e−14
Normal (n=113)
Tumor (n=1052)

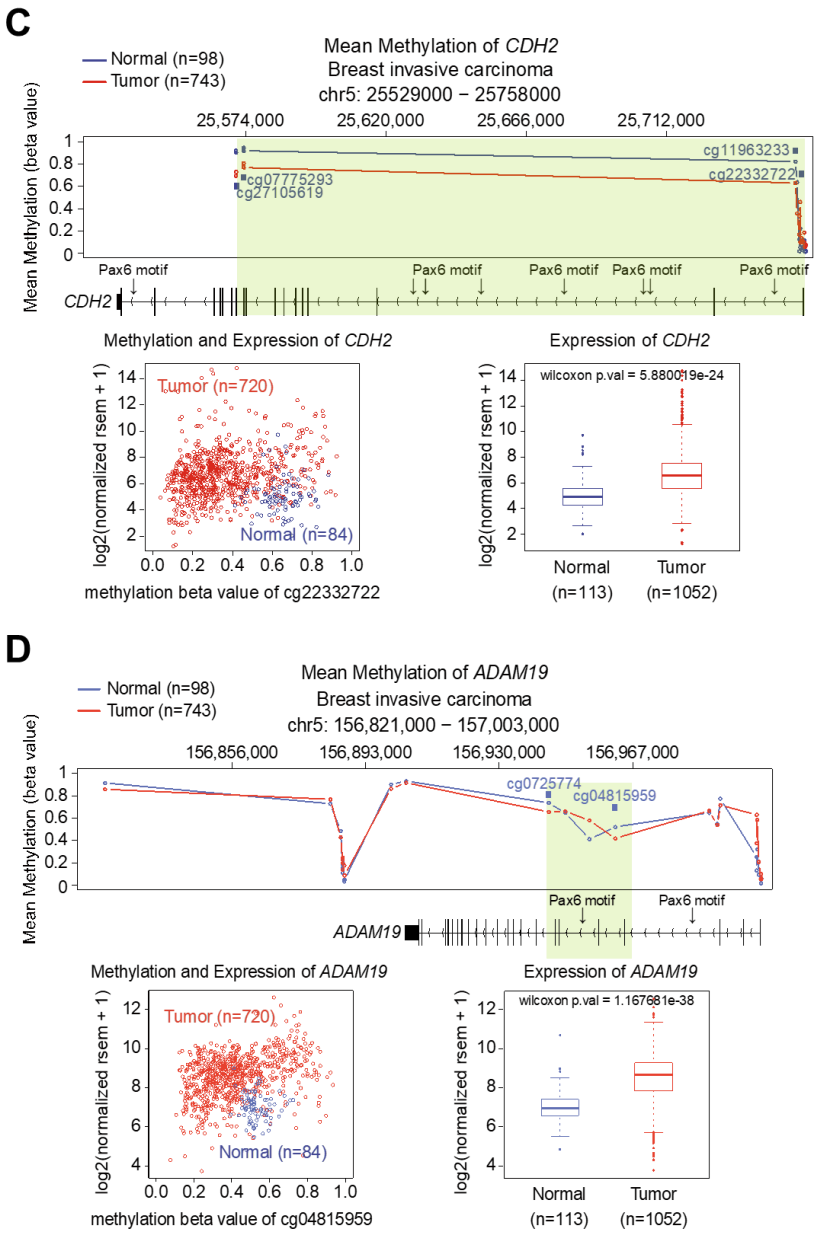


Figure 1.15. Conservation of intron DMRs and associating RNA expression in the candidate genes between HBC and CMT. Hypermethylated candidate genes, A) CDH5 and B) LRIG1. Hypomethylated candidate genes, C) CDH2 and D)

ADAM19. Human gene structures are line-drawn with intron PAX5 and PAX6 motifs (arrows). Wanderer database provided CG methylation levels in normal (blue line) and cancer (red line). CGs surrounding PAX motifs are labeled in red (hypermethylation) or in blue (hypomethylation). Scatter plot presents anti-correlation between methylation level in selected CG and gene expression; normal: blue, cancer: red. Box plot shows overall gene expression levels of normal (blue) and cancer (red) in TCGA database.

pattern of methylation profiles and gene expression between normal and HBC (**Figure 1.15C, D**). Methylation profiles and gene expression of two *CDH* genes (hypermethylation in *CDH5*, hypomethylation in *CDH2*) were well-conserved in normal and HBC populations. The 1st intron of *CDH5* harboring the hypermethylated PAX5 motif in CMT was also hypermethylated and down-regulated in HBC (**Figure 1.15A**). Moreover, the 2nd intron of *CDH2* which harbors a hypomethylated PAX6 motif in CMT was also hypomethylated and up-regulated in HBC (**Figure 1.15B**). Of note, *LRIG1* has somewhat different gene structures in human and dog, such as different number of exons (22 in human, 25 in dog), and thus the hypermethylated intron with the PAX5 motif that has anti-correlation with gene expression (**Figure 1.15C**) was found in the 3rd and 5th introns in human and dog, respectively. Similarly, hypomethylated PAX6 motifs in *ADAM19* have an anti-correlation with the gene expression even though the hypomethylated intronic PAX6 motifs are located on different introns in dog and human (13th intron in dog and 5th intron in human) (**Figure 1.15D**).

As a whole, our date revealed that the orthologous intron regions of PAX5 and PAX6 binding motifs between human and dog have similar CG methylation alterations in breast cancers. These results thus suggest that the molecular similarity between CMT and HBC exists not only at the genomic and transcriptomic levels but also the epigenomic level.

Discussion

This study of CMT has gained increasing importance not only for animal welfare but also for better understanding of HBC. Over the past decade, comparative studies of CMT and HBC have been conducted at the genome and transcriptome levels using high-throughput sequencing data and have presented similarities and discrepancies existing between CMT and HBC [36,38]. However, a comprehensive analysis of the genome-wide methylome in CMT and its comparison with the HBC methylome had not been studied yet.

We employed a linearized mixed model to classify DMRs with multiple variances and successfully determined CMT- and Subtype-DMRs. Our methylome data showed that DMRs were biased towards hypermethylation on the genic regions represented by promoter, exon, intron and TTS in CMT. This is consistent with the previous knowledge that the general cancer methylation pattern is represented by intergenic hypomethylation and gene body hypermethylation [61]. In addition, each DMR (CMT- and subtype-) as a methylation signature could separate either normal from CMT or among the three subtypes in principal component analysis. The OncoScore and the GO enrichment analysis results demonstrated that CMT- and subtype-DMRs are functionally linked to CMT and subtypes.

Of further note in the present study was that most of the enriched cancer-associated pathways were from DMRs that included intron regions. Recently, the regulatory

role of the intron region has been proposed in certain gene expressions, particularly the first intron closely located to the promoter [52,62,63]. Some studies proposed enhancer sequences in introns and showed the transcription factor (TF) binding to the sequences [64]. Although, some studies also proposed alternative splicing in RNA causing intron retention as putative roles of intron DNA methylation, this needs to be further elucidated [62,65,66]. Furthermore, the role of TFs and DNA methylation in intron regions also needs to be elucidated because, although DNA methylation is generally associated with transcriptional silencing, the effect of methylation on binding affinity for most TFs is still unknown [67,68]. Yet, Yin et. al, measured the TF binding affinity to the methylated motif in about half of human TFs using modified high-throughput sequencing and suggested that the affinity of individual TFs can either be increased or decreased on methylation, depending on the different positions within the binding site [6]. In this study, we identified PAX5 and PAX6 motifs, known to be tumor suppressive and oncogenic TFs, that are enriched in hyper- and hypo-methylated intron DMRs of CMT, respectively. Nine members are known in the Paired box (PAX) gene family and some members [69] particularly PAX5 and PAX6 are known to have similar binding sites based on their crystal structure [70]. However, recent studies provided enough evidence that PAX5 and PAX6 work independently [56-58]. For instance, they are clustered in different groups (PAX5 in group 2, PAX6 in group 4) [71] and bind to different genomic loci in ChIP-seq analysis [72]. It is also known that only PAX genes from the same group are capable of complementing the loss of function in others [71]. We also identified a list of motifs, such as NR2F1, RORA,

HNF4G, NR3C, MYB and RUNX, that were enriched in intron DMRs but of which the motifs lacked a CG nucleotide inside their recognition sites. The substantial putative target genes reversely regulated by intron methylation around motifs were investigated. These are also meaningful to study further since these motifs without a CG sequence in their recognition site can still be influenced by the surrounding CG methylation levels [6].

There exists some limitation in directly comparing our CMT methylation profile to the HBC methylome database since the methylation profiling for HBC provided by TCGA was generated from an Infinium Human Methylation450 BeadChip array (Illumina, USA), not MBD-sequencing. Nonetheless, the result showing the correlation between methylation in the intron region and gene expression may support the importance of intron methylation, at least in regard to these candidate genes, *CDH5* and *LRIG1* with PAX5 motifs and *CDH2* and *ADAM19* with PAX6 motif in both CMT and HBC.

This chapter was published as:

Alternative methylation of intron motifs is associated with cancer-related gene expression in both canine mammary tumor and human breast cancer

A-Reum Nam[1], Kang-Hoon Lee[1], Hyeon-Ji Hwang[1], Johannes J. Schabort[1], Jae-Hoon An[2], Sung-Ho Won[2] and Je-Yoel Cho[1]* (2020)

[1] Department of Biochemistry, BK21 Plus and Research Institute for Veterinary Science, School of Veterinary Medicine, Seoul National University, Seoul, Korea.

[2] Department of Public Health Sciences, Graduate School of Public Health, Seoul National University, Seoul, Korea.

Clinical Epigenetics 12, 110 (2020). https://doi.org/10.1186/s13148-020-00888-4

CHAPTER II

The landscape of PBMC methylome in canine mammary tumors reveals the epigenetic regulation of immune marker genes and its potential application in predicting tumor malignancy

Introduction

Immune cells interact with the tumor and are involved in tumor invasion, metastasis, and systemic immune cell exhaustion in the tumor environment [73]. Accordingly, cancer treatments have been developed using immune checkpoint inhibitor (ICI) that interferes with the signal between immunity and tumor and adoptive cell therapy that allows immune cells to attack tumor cells (e.g., CAR-T, TILs, *etc.*). In numerous clinical trials, the effectiveness of immunotherapy on tumors depends on the cancer type and the cancer patient's immune status [74]. Peripheral blood mononuclear cells (PBMCs) containing a variety of cell types such as T- and B- lymphocytes, natural killer cells (NK cells), dendritic cells (DCs), and monocytes actively respond to tumor cells [75]. Though PBMC is a valuable source for monitoring immune-relevant tumor mechanisms and diagnosing tumor status [75], a comprehensive omics analysis in PBMCs from tumor patients has not been performed. Here, we generated a primary dataset suitable for understanding epigenetic regulation circulating immune cells respond to tumors using PBMCs derived from dog mammary gland tumors.

Epigenetic modification is an essential factor that enhances the effectiveness of cancer treatment by immune cells [76]. Recently, clinical trials have been underway on the combination therapy of ICI with epigenetic drugs such as HDAC inhibitors (HDACi), 5-aza-2-deoxycytidine (5-Aza), and decitabine [77]. DNA methylation is a reversible change and a valuable target that can be modulated and quickly detected

[78]. Promoter methylation of checkpoints such as CTLA-4, PD-1, and CD28 has been reported to be associated with systemic suppression of immune cells in the tumor microenvironment (TME) [79]. In addition, methylation of peripheral blood immune cells is a strong candidate for diagnosing solid tumors such as head and neck squamous cell carcinoma [80], liver cancer [81], bladder cancer [82], and ovarian cancer [83]. Understanding epigenetic regulation in circulating immune cells provides valuable information to diagnose tumor type, grade, and prognosis and treat tumors with immune remodeling therapy (e.g., CAR-T therapy) [84]. Nevertheless, many studies in human cancer methylome have focused on tumor-infiltrating immune cells and immune checkpoints. Epigenetic information of PBMC has advantages in providing diagnostic, prognostic, and therapeutic information based on easily accessible liquid biopsy modality.

Since epigenetic responses to environmental factors occur actively in dogs as in humans, comparative medical studies using dogs have been conducted on aging, tumor biogenesis, and inflammatory diseases [8]. It has been reported that dogs might be helpful animal models for immunotherapy studies because they are immune-competent, and their tumor biology is similar to that of humans [85]. Indeed, several recent studies have evaluated the cross-reactivity of immunotherapy against human and canine cancers [9].

We identified epigenetic signatures in circulating immune cells of CMT through a genome-wide methylation study of PBMCs in normal, benign tumors and malignant

tumors (carcinoma). We investigated abnormal methylation patterns in immune regulatory genes associated with the proliferation and normal differentiation of various immune cells. This result suggests that immune cell activity is affected by CpG methylation not only in the tumor microenvironment but also in peripheral blood. Furthermore, we modeled a two-step classifier that can distinguish benign and malignant tumors from normal through machine learning (ML) algorithms using the PBMC methylome datasets.

Materials and methods

Clinical samples

The protocol was approved by the Institutional Review Board (IRB) of Seoul National University (IACUC SNU-170602-1) and the Institutional Animal Care and Use Committee (IACUC). Blood samples from healthy dogs and dogs with clinically diagnosed mammary tumors were collected in EDTA tubes. For PBMC isolation, 1-2ml of blood was carefully transferred to a 2X volume of Ficoll-Paque PLUS (GE Healthcare, 17144002) and centrifuge at 400 *g*. After washing with phosphate-buffered saline (PBS), obtained PBMCs were fresh-frozen for storage or used for following MBD sequencing, target BS sequencing, and total RNA sequencing. Clinical information for normal and mammary tumor dogs is presented in **Table 2.1**.

Methyl-binding domain (MBD) sequencing

MBD sequencing was performed as previously reported by our group [86]. Briefly, genomic DNA has been isolated from dog-derived PBMCs using the DNeasy DNA Extraction Kit (QIAGEN, 69504). After 3 μg of genomic DNA was sonicated, MBD-biotin was incubated with Dynabeads-streptavidin and bound to 500 ng of dsDNA. MBD-enriched DNA was obtained from 600 and 800 mM elutes which contain highly methylated DNA fragments. MBD-enriched DNA was subjected to library

construction and sequenced by Illumina Hiseq 4000 next-generation sequencing platform (Illumina, CA, USA).

Genome-wide methylome profiling

Quality check, trimming, alignment, and quantitation processes for MBD-seq data were executed as detailed in our previous methylome study [86]. We calculated raw counts for bins (called 'Bins_used' in **Figure 2.1C-E**) excluding low signal bins and zero CpG bins using the '*MEDIPS.createROIset'* function of MEDIPS R Bioconductor [42]. We performed pairwise DMR analysis for the Bins_used by applying the '*MEDIPS.meth*' function of MEDIPS. We set specific parameters (p.adj = "fdr", diff.method = "edge R", minRowSum = 1000, diffnorm = "quantile"), the bins with FDR-adjusted *p*-value <0.1 and $|\log_2 FC| \geq 0.585$ (same as fold change upper 1.5) were defined as significant DMRs. Quantile normalized counts and $\log_2$ transformed CPM values were used for plotting and quantitative analysis. In addition, we counted reads in every 50 bp across the whole genome using the source code of MethylAction (https://github.com/jeffbhasin/methylaction) to generate high-resolution 'bigwig' files for visualizing methylation peaks in the Integrative Genome Viewer (IGV v.2.8.0) [87].

Table 2.1. The information about dog donors providing blood samples used for MBD-seq

Donor ID	Type	Subtype	Hospital	Sex	Age (years)	Breeds	RNA-seq
N102	N	Normal	HMR	MC	10	Schnauzer	O
N163	N	Normal	HMR	MC	3	Dachshund	
N169	N	Normal	SNU	F	7	Cocker Spaniel	
N171	N	Normal	SNU	FS	2	Cocker Spaniel	
N172	N	Normal	SNU	F	1	Maltese	
N173	N	Normal	SNU	F	2	Maltese	
N174	N	Normal	SNU	F	7	Maltese	
N178	N	Normal	SNU	FS	5	Maltese	
N181	N	Normal	SNU	FS	9	Poodle	
N182	N	Normal	SNU	FS	11	Shih-tzu	
N183	N	Normal	SNU	FS	10	Maltese	
N187	N	Normal	SNU	FS	6	Maltese	
N188	N	Normal	SNU	FS	5	Poodle	
N189	N	Normal	SNU	FS	11	Maltese	
N190	N	Normal	SNU	F	12	Maltese	
B004	B	Ductal	SNU	F	10	Schnauzer	
B006	B	Simple	HMR	FS	12	Poodle	
B007	B	Mixed	HMR	F	12	Maltese	
B013	B	Simple	HMR	F	9	Maltese	
B018	B	Complex	HMR	F	10	Mixed	O
B019	B	Complex	HMR	F	10	Cocker Spaniel	
B020	B	Ductal	SNU	F	12	Chihuahua	

B022	B	Mixed	SNU	F	9	Maltese	
B024	B	Complex	SNU	F	11	Maltese	O
B029	B	Complex	SNU	F	10	Maltese	O
B034	B	Complex	HMR	FS	10	Shih-tzu	O
B062	B	Complex	HMR	FS	14	Schnauzer	
B063	B	Complex	HMR	FS	14	Schnauzer	
B066	B	Complex	HMR	FS	12	Yorkshire Terrier	
B072	B	Complex/Mixed	SNU	FS	11	Maltese	O
B073	B	Complex	SNU	F	8	Maltese	O
B084	B	Complex	HMR	F	10	Shih-tzu	O
B085	B	Mixed	HMR	F	14	Maltese	
B086	B	Complex	HMR	FS	15	Maltese	O
B087	B	Simple	HMR	F	10	Maltese	O
B091	B	Mixed	SNU	F	16	Maltese	O
B096	B	Simple	HMR	F	5	Yorkshire Terrier	O
B099	B	Ductal	SNU	FS	14	Cocker Spaniel	O
B101	B	Complex	SNU	F	10	Maltese	O
B104	B	Complex	HMR	FS	10	Maltese	O
B109	B	Complex	SNU	F	9	Maltese	O
B110	B	Complex	SNU	FS	10	Bichon Frise	O
B137	B	Mixed	HMR	F	10	Maltese	O
B141	B	Complex	HMR	F	5	Mixed	
B145	B	Complex	BON	FS	7	Old English Sheepdog	O
B155	B	Complex	SNU	FS	13	Mixed	
C009	C	Mixed	HMR	F	11	Maltese	O
C010	C	Simple	SNU	FS	10	Cocker Spaniel	O
C012	C	Simple	HMR	F	11	Dachshund	O

C021	C	Simple	SNU	FS	13	Pomeranian	O
C028	C	Inflammatory	HMR	FS	13	Schnauzer	
C052	C	Simple	SNU	F	12	Maltese	O
C053	C	Inflammatory	SNU	FS	13	Jindo	O
C064	C	Inflammatory	HMR	F	11	Maltese	
C065	C	Comedo	HMR	F	10	Great Pyrenees	
C067	C	Simple	HMR	F	10	Maltese	
C068	C	Simple	HMR	FS	13	Siberian Husky	
C071	C	Inflammatory	HMR	FS	14	Cocker Spaniel	
C078	C	Simple	HMR	FS	14	Cocker Spaniel	
C079	C	Simple	HMR	F	12	Poodle	
C081	C	Simple	HMR	FS	10	Poodle	
C083	C	Simple	HMR	FS	11	Shih-tzu	
C094	C	Simple	HMR	FS	10	Shih-tzu	O
C095	C	Simple	HMR	FS	11	Maltese	O
C105	C	Inflammatory	HMR	FS	13	Jindo	O
C106	C	Simple	SNU	F	15	Shih-tzu	
C107	C	Simple	SNU	F	14	Cocker Spaniel	O
C128	C	Complex	SNU	FS	14	Shih-tzu	O
C132	C	Complex	SNU	F	12	Maltese	O
C138	C	Simple	HMR	FS	9	Mixed	O
C143	C	Simple	HMR	F	15	Shih-tzu	O
C148	C	Simple	SNU	F	10	Chihuahua	
C149	C	Complex	SNU	F	11	Cocker Spaniel	O
C151	C	Complex	SNU	FS	14	Dachshund	O
C152	C	Simple	BON	FS	14	Dachshund	
C157	C	Unknown	SNU	F	13	Samoyed	O

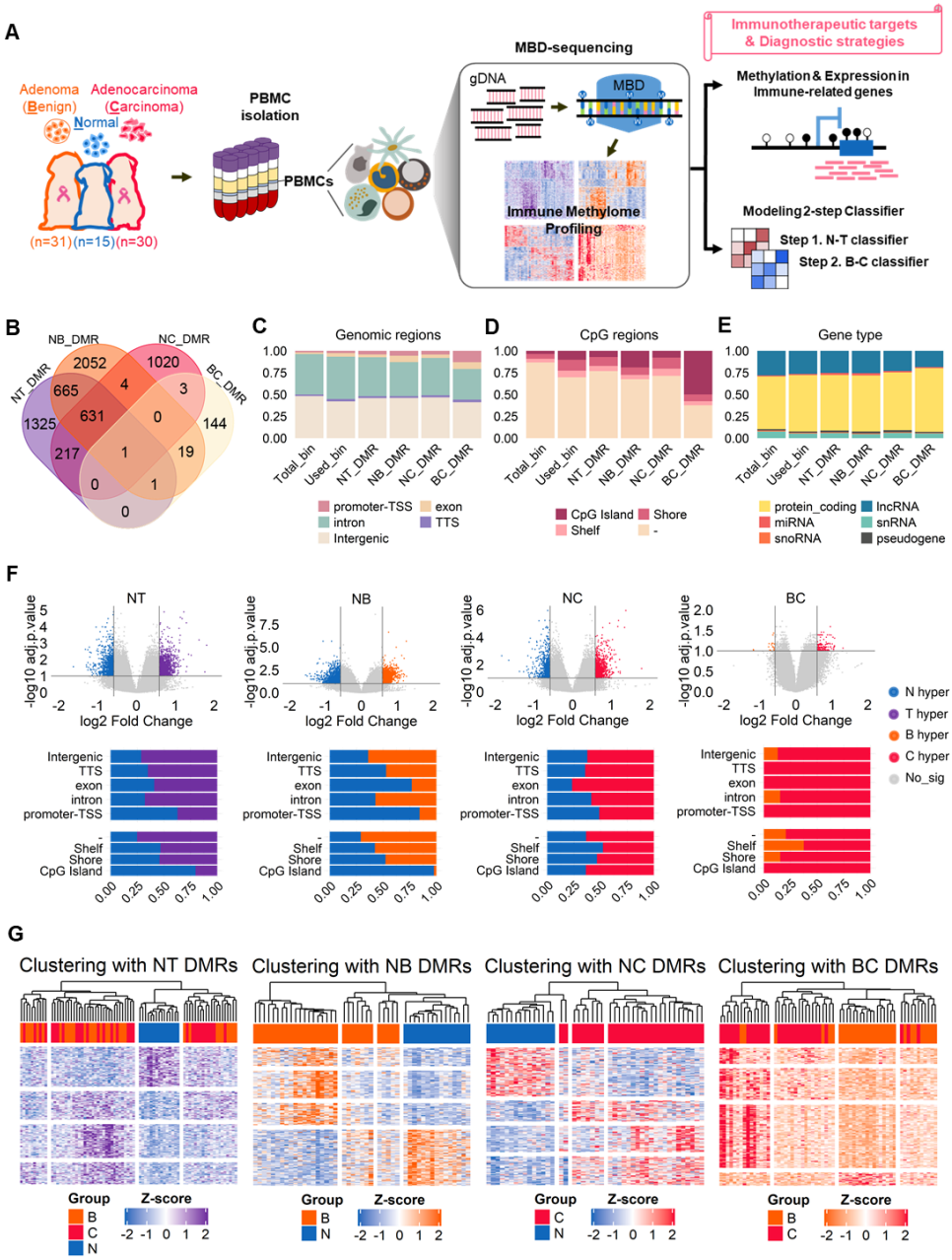


Figure 2.1. Pair-wise comparison for genome-wide PBMC methylome datasets from benign, carcinoma, and normal dogs. A) Synopsis of genome-wide PBMC

methylome study. B) A Venn diagram shows the number of common and unique DMRs identified in each comparison (FDR-adjusted *p*-value <0.1 and $|\log_2 FC| \geq 0.585$). C-E) The distributions of genomic features in Total bins, Bins_used, and each DMR to see pronounced regions. 'Bins_used' regarded signal peaks used for DMR analysis, excluding noise bins (both low signal bins and zero CpG bins) from 'Total bins'. F) Volcano plots and 100%-scaled stacked bar plots with the frequency and genomic profile of hypo- and hyper- methylated bins. The x-axis is the '$\log_2$ methylation fold change', and the y-axis means the statistical significance. Hypermethylated in 'N' is expressed as blue, 'T' as purple, 'B' as orange, and 'C' as red. G) Heatmap Clustering of 'N and T with NT_DMR (2840 DMRs)', 'N and B with NB_DMR (3373 DMRs)', 'N and C with NC_DMR (1876 DMRs)', 'B and C with BC DMRs (168 DMRs)'. The clustering distance between samples (columns) followed Pearson's correlation, and the 'complete' method was used.

Annotation of methylation peaks

Information on genomic features of CanFam3.1 (v99), a dog reference genome, was obtained in a GTF format from Ensembl Genome Browser (release 104, May 2021). `Promoter-TSS` means extended regions around TSS from -1000 bp to + 100 bp, while `TTS` indicates extended regions around TTS from -100 bp to +1000 bp. We downloaded the genomic location of CpG islands from the UCSC Genome Browser and named the region extending ±2 kb from the CpG island as 'CpG shore' and the region extending from ±2 kb to ±4 kb from the CpG island as 'CpG shelf'. Total bins, Bins_used, and DMRs were annotated to the prepared genomic information using the '*annotatePeaks.pl*' function provided in HOMER v4.11.1.

Functional enrichment analysis

We investigated the enriched terms for DMGs using EnrichR (a web server for the comprehensive gene set enrichment analysis: maayanlab.cloud/Enrichr/) [44] to elucidate the function of genes undergoing aberrant methylation. Because most functional terms are derived from human and mouse, we converted dog Ensembl IDs into human orthologous gene symbols using multiple species datasets downloaded from the Ensembl Biomart (Ensembl Genes 104). Finally, we found significant functional terms in various libraries such as Gene Ontology (GO), KEGG pathway

(2021), MGI Mammalian Phenotype (Level 4, 2021), and Human Gene Atlas. Panglao DB is a web database that shares single-cell RNA sequencing data conducted on human and mouse [88]. We extracted a list of marker genes for 11 immune cell types corresponding to the composition of PBMC included in the immune system from the Panglao DB. This list was used to identify methylation changes in cell marker genes.

Targeted Bisulfite-sequencing (BS-seq)

Targeted BS-seq was performed using genomic DNA from 9 PBMC samples, including PBMCs used for MBD-seq (n=3 in normal (N), benign (B), and carcinoma (C), respectively). We designed bisulfite primers using the Bisulfite Primer Seeker (https://www.zymoresearch.com/pages/bisulfite-primer-seeker). The overall process of targeted BS-seq was conducted as previously described [89]. The primer sequences are listed in **Table 2.2**. Subsequently, the sequences were aligned to the reference sequence in the amplified region using MEGA v11.0.11 [90]. The methylation (%) for the whole CpGs in each region was calculated and visualized as violin plots. To compare the methylation levels between different groups each other, the *t*-test was employed.

Table 2.2. The list of primers designed for targeted BS-sequencing

Target Gene	Direction	Sequence (5' → 3')
BACH2	Forward	ATTTGTGTGTTTGTTTATTATTTAGAAA
BACH2	Reverse	TTAAAATTAACTTTCTCTAACCTAAACC
SH2D1A	Forward	TGGTTTTAATTAGGTATTAYGTTTTTTA
SH2D1A	Reverse	CCTTAAATTACCATCACTTAAAACTATT
TXK	Forward	AGAAATTAAAATTTGGTTTTTTAGTTTT
TXK	Reverse	ATTCTTTCCACCTATAAATAAAATAACT
UHRF1	Forward	GTTTGGATTAGGTAAGAATAAAGGT
UHRF1	Reverse	CACCRTTATTAATCATTAATACACTAAT

Classifier modeling and evaluation

We calculated the log (CPM + 1) values for the entire bins to generate the methylome-based classifiers, while log (TPM + 1) was used for modeling transcriptome-based classifiers. Five ML algorithms; 1) Support vector machine (SVM) with linear kernel, 2) SVM with the radial kernel, 3) Random Forest (RF), 4) Gradient Boosting Machines (GBM), and 5) K-Nearest Neighbor (KNN) were compared to construct an optimal classifier. We estimated the performance of the ML algorithms through the 10-fold cross-validation (10-fold CV) to reduce the overfitting of models. In this process, the hyperparameters in each model were selected by default because we chose an ML algorithm to find DMRs that generally classified the groups well using R package caret (v6.0.85) [91]. The two-step classifier consists of an NT classifier that distinguishes tumors from normal and a BC classifier that distinguishes carcinoma from benign tumors using PBMC methylome. Although both classifiers were constructed through the same computational modeling process, there was an additional modeling step based on feature importance to enhance the performance of the BC classifier. The optimal BC classifier was designed with 127 DMRs, which had high feature importance from the GBM classifier with the highest accuracy among the primary models (**Table 2.3**). Feature importance was calculated based on nested cross-validation using the R package gbm (v2.1.8) [92]. We evaluate multiple classifiers using the prediction accuracy and

Table 2.3. The list of 127DMRs which have high feature importance in BC classifier

DMR ID	DMR Group	Hyper_in	Importance	Gene Name	Annotation	CpG_annotation
chr28_28069001	BC_DMR	B-hyper	8.938291598	-	Intergenic	-
chr20_53056501	BC_DMR	C-hyper	6.933393228	MYO1F	intron	CpG Island
chr8_65136001	BC_DMR	C-hyper	4.335951851	AK7	promoter-TSS	CpG Island
chr26_10793501	BC_DMR	C-hyper	3.775880139	-	Intergenic	CpG Island
chr9_49248501	BC_DMR	C-hyper	3.024101558	LHX3	intron	CpG Island
chr1_630001	BC_DMR	C-hyper	2.067359261	ADNP2	intron	Shelf
chr9_56626501	BC_DMR	C-hyper	0.673368967	-	Intergenic	CpG Island
chr6_55672501	BC_DMR	B-hyper	0.580970748	-	Intergenic	-
chr33_14546501	BC_DMR	C-hyper	0.049627896	-	Intergenic	-
chr18_11566001	BC_DMR	C-hyper	9.73E-04	-	intron	-
chr27_45316001	NB_DMR	B-hyper	11.7308111	BCL2L13	intron	-
chr14_58652001	NB_DMR	N-hyper	11.48877426	-	intron	-
chr13_37124501	NB_DMR	N-hyper	2.077878102	MAFA	promoter-TSS	CpG Island
chr18_54047001	NB_DMR	N-hyper	2.061007414	-	Intergenic	CpG Island
chr20_57532501	NB_DMR	N-hyper	1.436104355	SBNO2	intron	CpG Island
chr16_53991501	NB_DMR	N-hyper	1.321658453	-	promoter-TSS	CpG Island
chr1_29205001	NB_DMR	B-hyper	1.22912013	-	exon	-
chr21_45191501	NB_DMR	B-hyper	1.133891561	-	intron	-
chr10_67061501	NB_DMR	B-hyper	0.871292278	-	Intergenic	-

chr15_50750001	NB_DMR	B-hyper	0.427678749	ARFIP1	intron	-
chr8_4259001	NB_DMR	N-hyper	0.340928288	-	Intergenic	-
chrX_18832001	NB_DMR	B-hyper	0.228576134	-	Intergenic	-
chr6_7117501	NB_DMR	N-hyper	0.145512841	-	Intergenic	Shore
chr6_70470001	NB_DMR	N-hyper	0.136077045	ST6GALNAC3	intron	-
chr14_20757501	NB_DMR	B-hyper	0.133282151	ASB4	intron	-
chr1_97003501	NB_DMR	N-hyper	0.117417436	-	Intergenic	CpG Island
chrX_7181001	NB_DMR	B-hyper	0.109534584	MID1	intron	-
chrX_120970001	NB_DMR	B-hyper	0.087101028	-	Intergenic	-
chr22_19693501	NB_DMR	B-hyper	0.084687271	-	Intergenic	-
chr17_60196501	NB_DMR	N-hyper	0.083005999	SEMA6C	exon	CpG Island
chr12_61573501	NB_DMR	N-hyper	0.072993916	-	Intergenic	-
chrX_32247501	NB_DMR	B-hyper	0.043477542	-	Intergenic	-
chrX_43196501	NB_DMR	B-hyper	0.038029931	DGKK	intron	-
chr27_18679501	NB_DMR	B-hyper	0.036513912	FAR2	intron	-
chr18_31993001	NB_DMR	B-hyper	0.032511661	LDLRAD3	intron	-
chr5_61583501	NB_DMR	B-hyper	0.02876518	PARK7	intron	-
chr4_64336001	NB_DMR	B-hyper	0.026342531	-	Intergenic	-
chr9_32953001	NB_DMR	N-hyper	0.024922147	TSPOAP1	intron	-
chr6_34874001	NB_DMR	B-hyper	0.021878134	RBFOX1	intron	-
chr15_38104501	NB_DMR	B-hyper	0.017460966	-	TTS	-
chr38_22235001	NB_DMR	N-hyper	0.016958292	-	Intergenic	Shelf
chr36_6205001	NB_DMR	N-hyper	0.016113485	-	exon	-

chr20_52911001	NB_DMR	N-hyper	0.014547905	KANK3	exon	CpG Island
chrX_113723501	NB_DMR	B-hyper	0.013371289	-	Intergenic	-
chr21_50661001	NB_DMR	N-hyper	0.01335164	MS4A7	intron	-
chr37_23042001	NB_DMR	B-hyper	0.01100321	PECR	intron	-
chr13_50751501	NB_DMR	B-hyper	0.009132059	-	intron	-
chr30_27422501	NB_DMR	B-hyper	0.008082213	TLN2	intron	-
chr34_41892501	NB_DMR	B-hyper	0.008074174	-	Intergenic	CpG Island
chr26_25258001	NB_DMR	B-hyper	0.006884661	-	promoter-TSS	-
chr8_45573501	NB_DMR	B-hyper	0.004908331	RGS6	intron	-
chr7_9447001	NB_DMR	B-hyper	0.003427105	KCNH1	intron	-
chr4_70971501	NB_DMR	N-hyper	0.003336939	GDNF	intron	CpG Island
chr9_24848001	NB_DMR	N-hyper	0.002674867	-	Intergenic	CpG Island
chr5_30026001	NB_DMR	B-hyper	0.002237331	-	Intergenic	-
chr10_51118501	NB_DMR	N-hyper	0.002202345	-	intron	-
chr17_62057001	NB_DMR	B-hyper	0.001139659	MAGI3	TTS	-
chrX_9220001	NB_DMR	B-hyper	7.62E-04	FRMPD4	intron	-
chr5_32680501	NB_DMR	N-hyper	3.97E-04	DNAH2	intron	-
chr8_68475501	NB_DMR	N-hyper	3.20E-04	DEGS2	intron	Shore
chr2_11338001	NB_DMR	N-hyper	2.71E-04	-	Intergenic	CpG Island
chr2_4605501	NB_DMR	B-hyper	2.70E-04	-	Intergenic	Shelf
chr23_8153501	NB_DMR	B-hyper	6.88E-05	XYLB	intron	-
chr4_142501	NC_DMR	N-hyper	7.107012158	-	Intergenic	-
chr10_11437501	NC_DMR	C-hyper	4.248686749	-	Intergenic	-

chr3_27401001	NC_DMR	C-hyper	3.963346636	-	promoter-TSS	Shore
chr27_37350501	NC_DMR	N-hyper	2.395593107	SLC2A3	intron	-
chr10_910501	NC_DMR	N-hyper	1.873185325	-	Intergenic	Shore
chr8_2817001	NC_DMR	C-hyper	1.79969214	-	intron	-
chr2_77318001	NC_DMR	N-hyper	1.701805015	HSPG2	intron	CpG Island
chr27_43047001	NC_DMR	N-hyper	1.384637872	ERC1	intron	-
chr10_9261501	NC_DMR	C-hyper	1.223971759	-	Intergenic	Shelf
chr2_36696001	NC_DMR	N-hyper	1.085250354	-	Intergenic	-
chr31_24657001	NC_DMR	N-hyper	1.019872619	GRIK1	intron	-
chr9_50428001	NC_DMR	N-hyper	0.807627757	-	Intergenic	-
chr17_29263001	NC_DMR	C-hyper	0.679763372	-	Intergenic	-
chr9_24082001	NC_DMR	C-hyper	0.519813452	TBX21	intron	Shelf
chr1_21222001	NC_DMR	C-hyper	0.451198724	-	Intergenic	Shore
chr17_57218501	NC_DMR	C-hyper	0.43712045	PDE4DIP	intron	-
chr9_16000001	NC_DMR	C-hyper	0.374481999	-	Intergenic	-
chr2_32645501	NC_DMR	C-hyper	0.348228369	-	Intergenic	-
chr25_5055501	NC_DMR	C-hyper	0.330430139	MAB21L1	promoter-TSS	CpG Island
chr10_41317501	NC_DMR	N-hyper	0.322992905	-	TTS	-
chrX_51426501	NC_DMR	C-hyper	0.212449079	-	Intergenic	-
chr21_47489501	NC_DMR	C-hyper	0.178018583	-	Intergenic	-
chr7_27483501	NC_DMR	N-hyper	0.167991078	-	Intergenic	-
chr33_13136001	NC_DMR	C-hyper	0.161766794	-	Intergenic	-
chr12_68395001	NC_DMR	C-hyper	0.159269639	-	Intergenic	Shelf

chr22_1077501	NC_DMR	N-hyper	0.118465453	-	Intergenic	-
chr1_69608001	NC_DMR	C-hyper	0.110420087	-	Intergenic	-
chr37_4719001	NC_DMR	C-hyper	0.104329829	-	intron	-
chr15_2084501	NC_DMR	C-hyper	0.087735364	CTPS1	intron	-
chr23_50064501	NC_DMR	C-hyper	0.071507163	KCNAB1	exon	Shelf
chr3_49343501	NC_DMR	N-hyper	0.069011587	-	Intergenic	-
chr8_45100001	NC_DMR	N-hyper	0.059384184	SIPA1L1	intron	-
chr11_62869501	NC_DMR	C-hyper	0.049798542	-	intron	CpG Island
chr1_99648001	NC_DMR	N-hyper	0.043905775	ZNF8	intron	-
chr9_50427501	NC_DMR	N-hyper	0.039854523	-	Intergenic	-
chr25_38162001	NC_DMR	C-hyper	0.038411981	DOCK10	promoter-TSS	-
chr1_111974501	NC_DMR	N-hyper	0.032059185	CXCL17	intron	-
chr17_35808001	NC_DMR	N-hyper	0.025486887	-	intron	-
chr21_32086501	NC_DMR	N-hyper	0.025086979	TRIM66	intron	-
chr12_67126001	NC_DMR	C-hyper	0.014644766	SLC22A16	intron	-
chr16_36268001	NC_DMR	C-hyper	0.009128027	-	Intergenic	-
chr5_60358501	NC_DMR	N-hyper	0.004007098	-	Intergenic	-
chr37_16851001	NC_DMR	C-hyper	0.003824876	-	Intergenic	-
chr18_48696001	NC_DMR	N-hyper	0.003499961	-	Intergenic	-
chr1_95807001	NC_DMR	N-hyper	0.003389147	-	exon	-
chr10_13761001	NC_DMR	C-hyper	0.00304049	TRHDE	intron	-
chr20_51355001	NC_DMR	C-hyper	0.002936002	-	Intergenic	CpG Island
chr32_21296001	NC_DMR	N-hyper	0.002473558	ADH4	intron	-

chr2_11784001	NC_DMR	N-hyper	0.00245975	-	Intergenic	-
chrX_16516001	NC_DMR	C-hyper	0.002297134	-	Intergenic	-
chr3_179001	NC_DMR	C-hyper	0.001712125	-	Intergenic	-
chr11_73554501	NC_DMR	C-hyper	0.001612541	CDK5RAP2	intron	Shore
chr7_38802001	NC_DMR	C-hyper	0.001401762	H3-3A	exon	Shelf
chr4_5357001	NC_DMR	C-hyper	0.001280147	-	Intergenic	Shore
chr16_20238501	NC_DMR	N-hyper	0.001224167	PTPRN2	intron	Shelf
chr14_3159001	NC_DMR	N-hyper	8.94E-04	-	Intergenic	-
chr1_17409001	NC_DMR	C-hyper	8.18E-04	-	Intergenic	Shore
chr7_20661501	NC_DMR	N-hyper	7.44E-04	-	Intergenic	-
chr27_16380501	NC_DMR	N-hyper	6.48E-04	FGD4	intron	-
chr10_34483001	NC_DMR	N-hyper	4.77E-04	SH3RF3	intron	Shelf
chr7_62550001	NC_DMR	C-hyper	4.31E-04	SS18	intron	-
chr15_63693501	NC_DMR	C-hyper	4.14E-04	DDX60	intron	-
chr20_42568501	NC_DMR	C-hyper	1.45E-04	FYCO1	intron	-
chr5_80367501	NC_DMR	N-hyper	1.44E-04	SNTB2	intron	-

area under the ROC curve (AUC) using the R package pROC (v1.18.0) [93].

Statistics

Statistics and statistical tools for each analysis have been described above. The correlation coefficient between DMR methylation and gene expression was calculated by Pearson correlation and regression analysis. Comparison for the expression between The t-test was implemented to compare gene expression between groups. The number of asterisks between the two groups indicates the degree of statistical significance. If there was no statistical difference between the two groups, it was expressed as 'ns (not significant)' without an asterisk. We exploited Rex (v3.6.1) [94] and R (v4.0.2) in NGS data quantification, statistical analyses, and classifier modeling.

Results

Profiling differential methylation of peripheral blood mononuclear cells in canine mammary gland tumor

We first made genome-wide differential methylation profiles of PBMCs in CMT. To evaluate the genome-wide effects of mammary tumors on PBMC DNA methylation, PBMCs were collected from 15 healthy dogs (Normal; N), 31 dogs with mammary adenoma (Benign; B), and 30 dogs with mammary carcinoma (Carcinoma; C) (**Figure 2.1A**). The donor's information is listed in **Table 2.1.** The healthy samples consist of six dog breeds, aged 1 to 12, and 13 females, including eight spayed females and two neutered males. Patient specimens comprise 16 dog breeds aged 5 to 16 and six significant subtypes of canine mammary tumors (ductal, simple, complex, mixed, inflammatory, and comedo). All patient dogs were females or spayed females.

Global CpG methylomes have enriched and analyzed by methyl-CpG-binding domain sequencing (MBD-seq) that has high coverage in highly methylated CpG and CpG-rich regions (**Figure 2.1A**). The quality check for NGS data has also been performed. Sequencing reads more than 5X depth (considered as signal peaks) show about 50% CpG coverage, indicating that the MBD-seq data was successfully produced and informative (**Figure 2.2A**). The R Bioconductor MEDIPS (v.1.46.0) [42] was mainly employed to calculate methylation

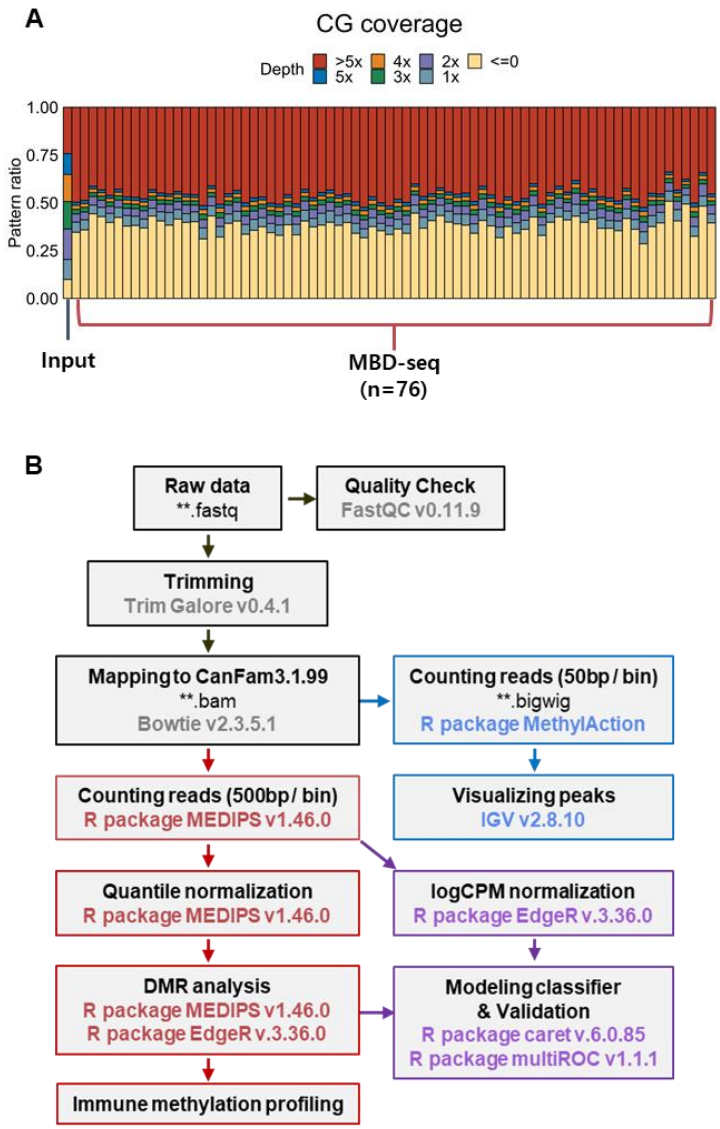


Figure 2.2. Quality check and processing MBD-seq data. A) The CpG coverage according to the read depth is shown as a 100% stacked bar plot. Compared to Input (the first bar), about half of genome CpGs have been covered by reads with high depth (>5x) in MBD-seq. It states that MBD-seq data has been successfully enriched

in CpG regions across the 76 PBMC samples. B) The workflow of the MBD-seq data processing. After trimming and mapping to CanFam3.1, MBD-seq data were quantified for DMR analysis, peak visualization, and classifier modeling. The black box indicates data pre-processing (from raw data to mapped reads), the blue box shows data processing for peak visualization, the red box exhibits the process of quantitation and normalization for DMR analysis, and the purple box shows the normalizing counts for the classifier modeling.

levels and identify differentially methylated regions (DMRs) (**Figure 2.2B**). DMRs were further subjected to ML for modeling an immune classifier for CMT. Of the total, 4,655,287 bins (referred to as 'Total bins' in **Figure 2.1C-E**) were generated at 500 bp size, and 1,220,164 bins (referred to as 'Bins_used' in **Figure 2.1C-E**) with reading counts of 25 or more were used for further analysis.

Together with pair-wise comparisons Normal vs. Benign (NB), Normal vs. Carcinoma (NC), and Benign vs. Carcinoma (BC), we also compared Normal vs. Tumor (NT), in which tumor includes benign and carcinoma. From each comparison, 2840, 3373, 1876, and 168 DMRs were identified with significance ($|\log_2 FC| \geq 0.585$, which is equal to |Fold Change| ≥ 1.5, and adjusted *p*-value (FDR) <0.1) for NT, NB, NC, BC, respectively (**Figure 2.1B**). Interestingly, the NB comparison shows the highest number of DMRs, followed by NT. As expected, NT comparison shares more than half of DMRs (1514) with NB and NC comparisons. Of note, DMRs from NB and NC comparisons share 636 DMRs and methylation directions (that is, B-hyper = C-hyper, B-hypo = C-hypo), indicating the methylation status of immune cells against tumors are similar in benign and carcinoma (**Figure 2.3**). Most of all, we focused if DMR profiles of PBMC can distinguish corresponding tumor types (benign or carcinoma) as well as Normal. However, only a small number of DMRs were identified from BC, and most BC_DMRs were unique across all DMRs, indicating that they are not explicitly associated with tumor states.

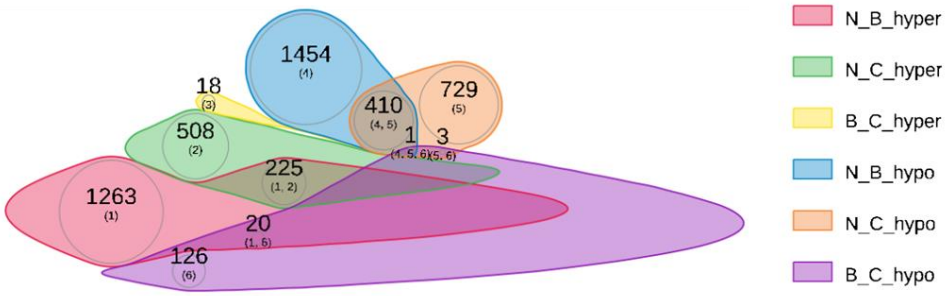


Figure 2.3. Venn diagram for hyper- and hypo-methylated DMRs. A Venn diagram shows the number of common and unique DMRs identified in each comparison according to the direction of methylation (FDR-adjusted *p*-value <0.1 and $|\log_2 FC| \geq 0.585$). There is no common DMR between 'NB_hyper and NC_hypo' OR 'NB_hypo and NC_hyper', which suggests that the methylation pattern in Benign is similar in Carcinoma PBMCs.

The uniqueness of BC_DMRs was shown in the genomic and CpG regional distribution and gene types linked to DMRs (**Figure 2.1C-E**). Total bins consist of five genomic regions. Compared with the 'Total bins', the intron region was increased when the intergenic region was decreased in the 'Bins used'. Moreover, more numbers of the CpG island, Shore, and Shelf regions were enriched in the 'Bins used' compared to the 'Total bins'. Interestingly, BC_DMRs were enriched in the promoter and exon regions and the CpG island regions, which are more associated with the protein-coding region.

We then analyzed the direction of DMRs using volcano plots and 100% stacked bar charts in eight genomic regions (**Figure 2.1F**). Overall, methylation increased in tumors compared to Normal. In BC, the Carcinoma group was more methylated than the Benign group. Regionally, changes in methylation status were highly dynamic according to the comparison group. In the NB comparison, there were more hypomethylated DMRs in CpG islands, promoter, and exon compared to other regions. Although these characteristics were similarly shown in the NT comparison, hypermethylated DMRs are prominent across all eight regions in the NC comparison.

Nevertheless, exon, promoter, and CpG island regions were highly hypomethylated in the BC comparison. Most of BC_DMRs, indeed, were hypermethylated in carcinoma. It is an essential feature because hypermethylation of certain groups of genes and DMRs might be a cancer-specific signature.

We then tested whether DMRs separate each comparison group. The pair-wise hierarchical clustering separated the Normal group from the Benign, Carcinoma, and Tumors groups (**Figure 2.1G**, **Figure 2.4A-B**). However, the Benign and Carcinoma groups were not entirely separated from each other, suggesting a new clustering algorithm for PBMC methylome classification for these group differentiation. The PBMC samples used in this study were obtained from dogs with diverse characteristics, including age, gender (neutered or not), tumor subtype, hospital where the blood was collected, and tumor features, among others. To investigate the potential effects of these variables, we performed hierarchical clustering using the NT_DMRs that we identified, to examine their influence (**Figure 2.4C**). These results show that the clustering of normal PBMC and tumor PBMC samples using NT_DMRs was not influenced by the diverse variables between the samples.

Differential methylation accompanies changes in immune cell populations and proliferation in malignant tumor patients.

Several studies have investigated the methylation patterns of blood immune cells, limited to specific target genes and not on a genome-wide scale [95-98]. Since PBMC is a mixture of a wide variety of immune cells, there is a limit to the regulation or role of various immune cells. To this end, single-cell bisulfite sequencing

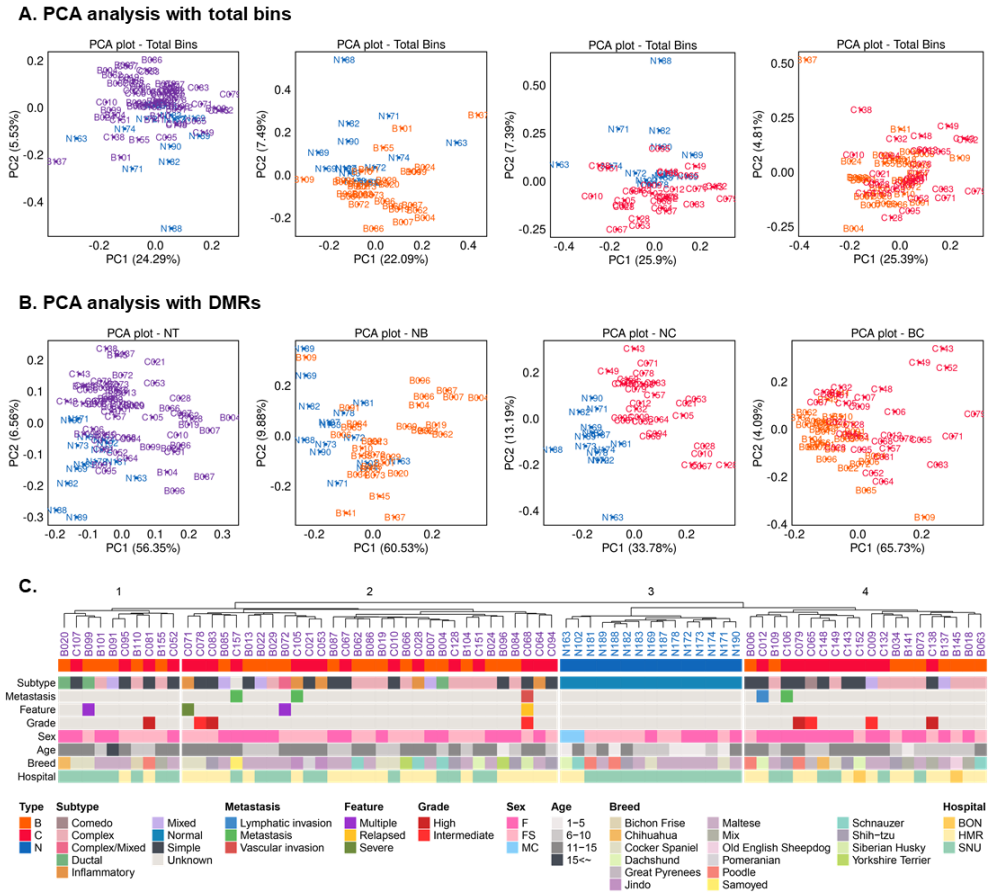


Figure 2.4. Unsupervised and supervised clustering between comparison groups. A) PCA clustering comparison groups using total bins. B) PCA clustering comparison groups using corresponding DMRs. Unlike unsupervised clustering using total bins, supervised clustering using DMRs distinguishes two groups. (N: blue, T: purple, B: orange, C: red) C) A total of 2,840 DMRs were identified through a comparison of normal and tumor PBMC samples (|Fold Change| ≥ 1.5 and adjusted *p*-value (FDR) <0.1), and subjected to hierarchical clustering to examine

the effects of various sample variables. Each column represents a different sample, while each row represents a variable of the samples, including subtype of tumor, metastasis, tumor feature, grade, sex, age, and others.

technology has been attempted, but several limitations exist in diagnosing cancer or defining the immune status. We analyzed the whole genome-wide methylation profile obtained from bulk PBMC samples and attempted to confirm various immune status changes in different tumors.

We defined DMGs using DMRs existing in promotor, exon, intron, and TTS and performed gene set enforcement analysis (**Figure 2.5 and Figure 2.6**). **Figure 2.5** shows that the immunocyte-related terms are significantly enriched in Gene Ontology (GO), Mammalian Phenotype Ontology in Mouse Genome Informatics (MGI), and Human Gene Atlas (HGA) databases [99-101]. In all comparative groups, genes involved in signal pathways directly related to cell activity, receptor activity, and cytokine modulation are hypomethylated in tumors (both benign and carcinoma), whereas there is no significant term or pathway found in hypermethylated in carcinoma (Part of 'GO' and 'KEGG' in **Figure 2.5**).

The MGI and HGA databases, which focus on the function of immune cells, provide clues to infer the immune status in the blood (Part of 'MGI Mammalian Phenotype' and 'Human Gene Atlas' in **Figure 2.5**). Comparing the normal with the overall tumor, the terms associated with the increase or abnormal function of T-cells, B-cells, and NK cells were high. The comparison between normal and cancer showed that the gene group with higher methylation in cancer PBMC was involved in the increasing or decreasing of B-cells or T-cells. Among T-cell types, the genes

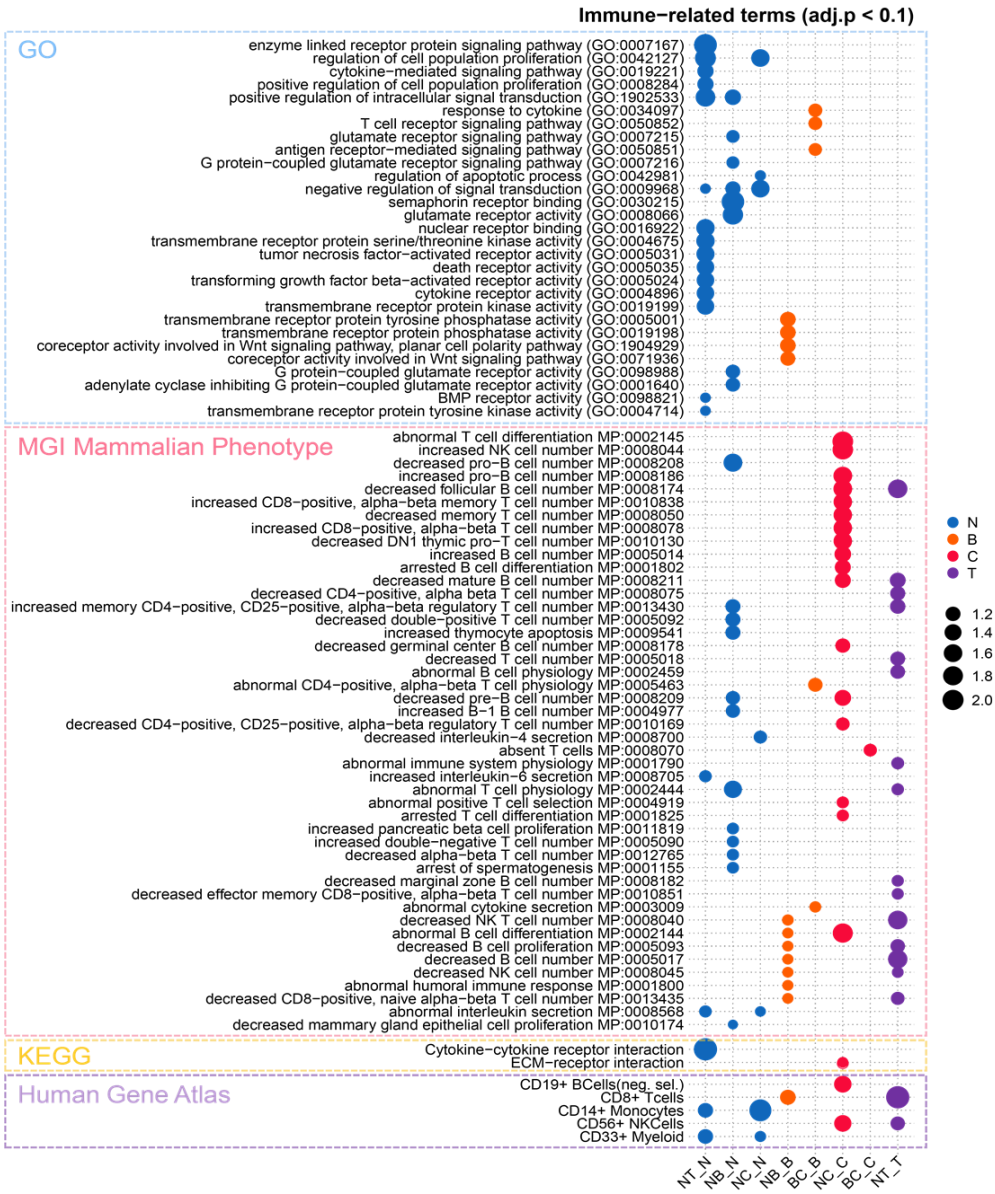


Figure 2.5. Gene enrichment analysis for DMGs shows differential immune signatures between tumor and normal PBMCs. Immune-related terms significantly enriched in the Gene Ontology (blue box), the MGI Mammalian

Phenotype (pink box), the KEGG pathway (yellow box), and the Human Gene Atlas (purple box) are shown. The color of dots means which group is hypermethylated ('N-hyper' is expressed as blue, 'T-hyper' as purple, 'B-hyper' as orange, and 'C-hyper' as red. The size of the dots indicates the statistical importance (according to -log10 adjusted *p*-value).

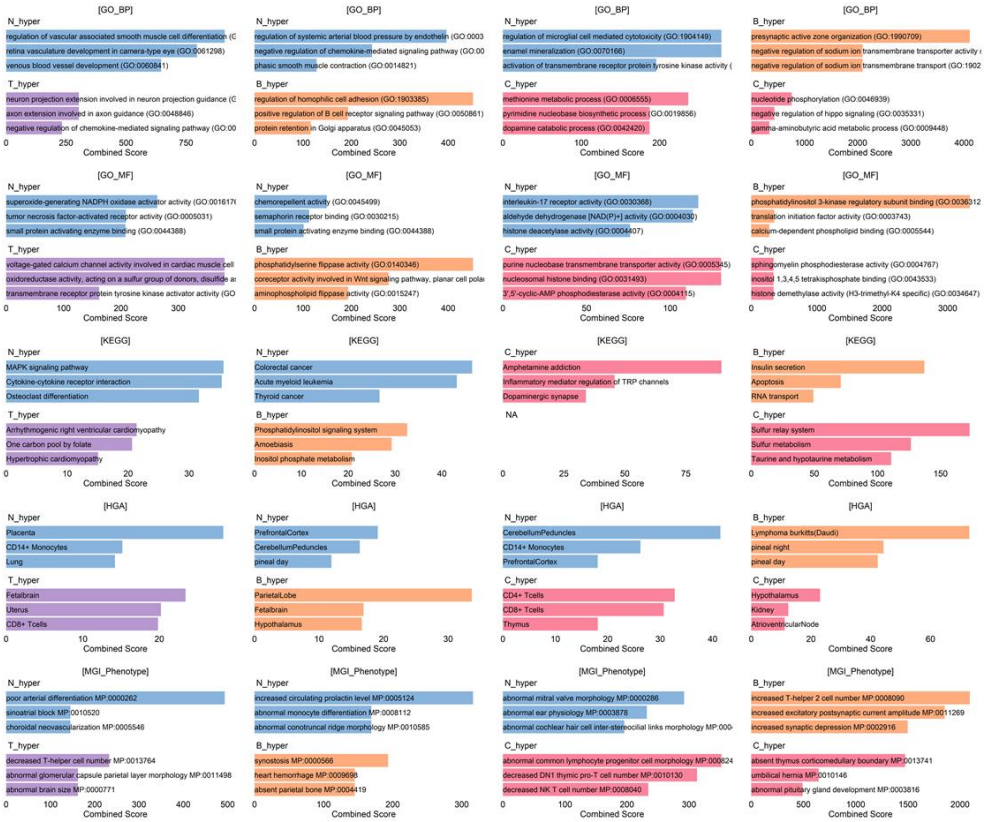


Figure 2.6. Enriched terms ranked in the Top 3 by combined score according to comparison groups. Enriched terms ranked in the Top 3 by combined score according to comparison groups. The top three terms are shown based on the combined score, the unit used in EnrichR. Terms enriched in Gene Ontology (both Biological Process and Molecular Function), KEGG pathway, Human Gene Atlas (HGA), and MGI mammalian phenotypes (MGI_Phenotype) are shown.

associated with the increase in CD8+ T-cells were most highly associated. On the other hand, compared with benign and normal the highly methylated genes in the benign group showed abnormalities in NK and B-cells. The primary immune cell types responding to benign and carcinoma differ. As for the DMR of BC comparison, there was no significant difference in the gene enrichment analysis, as the number was minimal, as shown in **Figure 2.1B**. Through the PBMC DMRs associated with immune responses to tumors, it is expected to find methylation biomarkers that can distinguish the presence or absence of tumors and the malignancy of tumors.

Immune cell markers functionally involved in cell proliferation and activation of B, T, and NK cells are hypermethylated in tumor PBMCs.

Through gene enrichment analysis (**Figure 2.5**), we could expect that methylation of immune cells in tumor patient dogs is involved in the population or activity of specific cell types. The gene enrichment analysis mapped the highest terms. Using text mining for meaningful GO terms in adj. p <0.1, words containing 'receptor', 'signal', 'activity', 'pathway', 'T cell', and 'B cell' were prominent in all comparisons (**Figure 2.7A**). These enrichments suggest that hypermethylation occurs in immune cells responding to tumors and is involved in signal transduction of immune cells. To confirm whether the methylation change in PBMC is due to the alteration of immune cell populations and or the cell activity, we investigated the

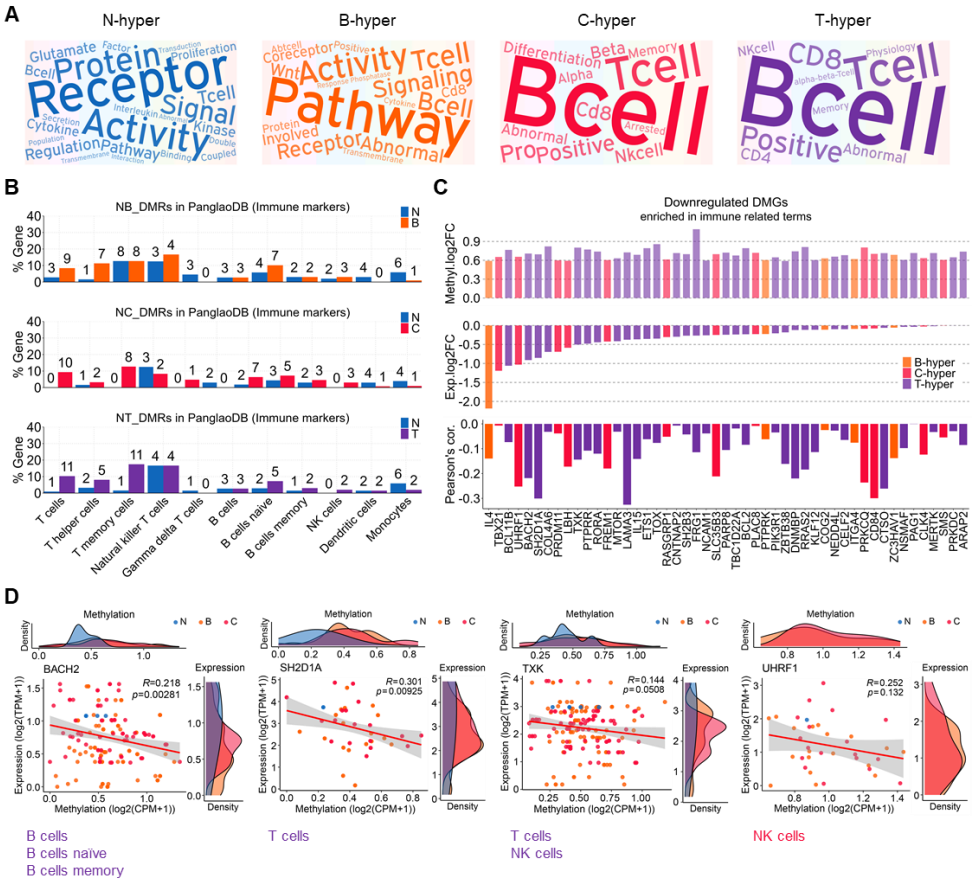


Figure 2.7. Immune cell markers involved in normal proliferation and activation of B-cells, T-cells, and NK cells are hypermethylated in tumor PBMCs. A) Text clouds intuitively show the frequency of words enriched in immune-related terms. The color of the text indicates which group is hypermethylated ('N-hyper' is expressed as blue, 'T-hyper' as purple, 'B-hyper' as orange, and 'C-hyper' as red). The meaning of the four colors (blue, purple, orange, and red) was applied equally to the following graphs in this figure. B) The number of hypermethylated genes included in immune cell type markers is expressed as a

percentage (%) of total genes in the corresponding cell type. The number of matched genes is displayed on the top of each bar. The list of marker genes for 11 types of immune cells was downloaded from Panglao DB. C) Among genes enriched in significant immune-associated terms, hypermethylated DMGs that reversely correlate with expression are shown. The y-axis of the bar graph on top means $\log_2$ fold change of methylation values, and that of the middle one means $\log_2$ fold change calculated using TPM values derived from RNA-seq. The y-axis of the bottom one shows the degree of inverse correlation between methylation and expression by Pearson's correlation. Hypermethylated genes included in Panglao DB and its genomic features are listed in **Table 2.4**. D) The scatter plots with linear regression (red line) in 4 representative genes among 49 genes listed in (C).

Table 2.4. The list of hypermethylated DMRs in immune cell type markers (Panglao DB)

Gene Name	Cell Type	Hyper_in	DMR ID	adj.*p*-value	log_2FC
ADGRG1	Gamma delta T cells	N-hyper	chr2_58841501	0.00313083	0.676948196
ADGRG1	Gamma delta T cells	N-hyper	chr2_58835501	0.077801732	0.646550585
ADGRG1	Gamma delta T cells	N-hyper	chr2_58841501	0.012071425	0.61497088
ADGRG1	Gamma delta T cells	N-hyper	chr2_58841001	0.000881389	0.615462186
ADGRG1	Gamma delta T cells	N-hyper	chr2_58841501	0.00337687	0.762624348
ARHGAP45	T memory cells	N-hyper	chr20_57786501	0.077736441	0.665636421
ARHGAP45	T memory cells	N-hyper	chr20_57787001	0.088862344	0.670393428
ARHGAP45	T memory cells	N-hyper	chr20_57786001	0.048979573	0.759269837
ARHGAP45	T memory cells	N-hyper	chr20_57786501	0.025360276	0.776182674
BACH2	B cells memory	N-hyper	chr12_49379501	0.064990035	0.638999182
BACH2	B cells	N-hyper	chr12_49379501	0.064990035	0.638999182
BACH2	B cells naive	N-hyper	chr12_49379501	0.064990035	0.638999182
BCL2	T memory cells	N-hyper	chr1_13875001	0.018474991	0.654214079
BCL2	T cells	N-hyper	chr1_13875001	0.018474991	0.654214079
CD44	Natural killer T cells	N-hyper	chr18_32793501	0.096986368	0.619997986
CD44	Monocytes	N-hyper	chr18_32793501	0.096986368	0.619997986
CD44	Natural killer T cells	N-hyper	chr18_32793501	0.088971726	0.712461803
CD44	Monocytes	N-hyper	chr18_32793501	0.088971726	0.712461803

CHSY1	NK cells	N-hyper	chr3_39851501	0.084892972	0.6035341
CLEC10A	Dendritic cells	N-hyper	chr5_32093001	0.006530008	0.736185195
CSF1R	Monocytes	N-hyper	chr4_58980501	8.96E-02	0.628570853
CX3CR1	Dendritic cells	N-hyper	chr23_8953501	0.065729997	0.601648294
CX3CR1	Monocytes	N-hyper	chr23_8953501	0.065729997	0.601648294
CX3CR1	Dendritic cells	N-hyper	chr23_8953501	0.072895147	0.653970226
CX3CR1	Monocytes	N-hyper	chr23_8953501	0.072895147	0.653970226
CYTIP	T memory cells	N-hyper	chr36_3456001	0.021800899	0.714411794
FCER1A	Dendritic cells	N-hyper	chr38_22688501	0.088287119	0.647878937
FCER1A	Dendritic cells	N-hyper	chr38_22688501	0.056950663	0.739279794
FCER2	B cells	N-hyper	chr20_52456001	0.009486532	0.730327583
FCER2	B cells naive	N-hyper	chr20_52456001	0.009486532	0.730327583
FCER2	B cells	N-hyper	chr20_52456501	0.003046101	0.805020388
FCER2	B cells naive	N-hyper	chr20_52456501	0.003046101	0.805020388
FCER2	B cells	N-hyper	chr20_52456501	0.012656403	0.793595381
FCER2	B cells naive	N-hyper	chr20_52456501	0.012656403	0.793595381
FCER2	B cells	N-hyper	chr20_52456001	0.03843911	0.694722337
FCER2	B cells naive	N-hyper	chr20_52456001	0.03843911	0.694722337
FCER2	B cells naive	N-hyper	chr20_52456001	0.001690093	0.791344084
FCER2	B cells	N-hyper	chr20_52456001	0.001690093	0.791344084
FCER2	B cells naive	N-hyper	chr20_52456501	0.002244865	0.835270952
FCER2	B cells	N-hyper	chr20_52456501	0.002244865	0.835270952
FLT3	Dendritic cells	N-hyper	chr25_11685001	0.020741486	0.596621498
GIMAP1	T memory cells	N-hyper	chr16_14794001	0.019621956	0.986586728

GNG7	B cells memory	N-hyper	chr20_56558001	0.027327207	0.616333427
GRAP2	T memory cells	N-hyper	chr10_25202001	0.072728152	0.631888785
HHEX	B cells memory	N-hyper	chr28_7063001	0.04331472	0.753088556
HHEX	B cells naive	N-hyper	chr28_7063001	0.04331472	0.753088556
IL17RA	Natural killer T cells	N-hyper	chr27_44808501	0.015025715	0.609522337
IL17RA	Natural killer T cells	N-hyper	chr27_44808501	0.026114003	0.666507466
IL1R2	T helper cells	N-hyper	chr10_41000501	0.025483805	0.64314538
IL1RN	Monocytes	N-hyper	chr17_37245001	0.043581459	0.661129035
IL1RN	Monocytes	N-hyper	chr17_37245001	0.054954839	0.636986022
IL1RN	Monocytes	N-hyper	chr17_37245001	0.040972246	0.707125318
IRF8	B cells naive	N-hyper	chr5_66796501	0.074488273	0.623631462
IRF8	B cells memory	N-hyper	chr5_66796501	0.074488273	0.623631462
IRF8	Dendritic cells	N-hyper	chr5_66796501	0.074488273	0.623631462
ITGAM	Monocytes	N-hyper	chr6_16855501	0.047296385	0.648316821
ITGAM	Dendritic cells	N-hyper	chr6_16855501	0.047296385	0.648316821
ITGAM	NK cells	N-hyper	chr6_16855501	0.047296385	0.648316821
LYN	Monocytes	N-hyper	chr29_7380501	0.02819836	0.607660328
LYN	Monocytes	N-hyper	chr29_7380501	0.023101019	0.585612756
LYN	Monocytes	N-hyper	chr29_7380501	0.046422882	0.646553763
MAFF	T cells	N-hyper	chr10_26483001	0.03398335	0.591260482
MAFF	T cells	N-hyper	chr10_26483001	4.33E-02	0.610506536
MS4A7	Monocytes	N-hyper	chr21_50661001	0.077220741	0.614799143
MS4A7	Monocytes	N-hyper	chr21_50661001	0.01962606	0.830081583
NFATC2	T helper cells	N-hyper	chr24_37672501	0.028863903	0.896844366

NFATC2	T helper cells	N-hyper	chr24_37672501	0.018695174	1.058895044
NFATC2	T helper cells	N-hyper	chr24_37672501	0.071177826	0.783472659
NOTCH3	T cells	N-hyper	chr20_46963501	0.003961028	0.610745512
NR4A1	Natural killer T cells	N-hyper	chr27_2899501	0.004143632	0.730659911
NR4A1	Natural killer T cells	N-hyper	chr27_2900001	0.000400681	0.700477974
NR4A1	Natural killer T cells	N-hyper	chr27_2900001	0.001973569	0.699916896
NR4A1	Natural killer T cells	N-hyper	chr27_2899501	0.004167384	0.776400617
NR4A1	Natural killer T cells	N-hyper	chr27_2899501	0.015229302	0.701241641
NR4A1	Natural killer T cells	N-hyper	chr27_2900001	0.000529735	0.708558486
PAX5	B cells memory	N-hyper	chr11_53342501	0.034524669	0.816040109
PAX5	B cells	N-hyper	chr11_53342501	0.034524669	0.816040109
PPL	Dendritic cells	N-hyper	chr6_36534001	0.024653736	0.623580811
PTGDS	Gamma delta T cells	N-hyper	chr9_48670001	0.074131546	0.658674298
SCIMP	Dendritic cells	N-hyper	chr5_31534001	0.016251029	0.689895705
SP100	T memory cells	N-hyper	chr25_42610501	1.76E-02	0.61548208
SPI1	Monocytes	N-hyper	chr18_42253501	0.065242288	0.663515559
SPI1	Monocytes	N-hyper	chr18_42254501	0.081585201	0.626830509
ADAMTS14	NK cells	B-hyper	chr4_21541001	0.018875982	0.641171176
ARHGAP15	T memory cells	B-hyper	chr19_45576001	0.057169695	0.706521255
ARHGAP15	T memory cells	B-hyper	chr19_45761001	0.035632986	0.640142785
BACH2	B cells memory	B-hyper	chr12_49247001	0.062080559	0.625552285
BACH2	B cells	B-hyper	chr12_49247001	0.062080559	0.625552285
BACH2	B cells naive	B-hyper	chr12_49247001	0.062080559	0.625552285
BCL2	T memory cells	B-hyper	chr1_13860001	0.029257888	0.612552592

BCL2	T cells	B-hyper	chr1_13860001	0.029257888	0.612552592
BCL2	T memory cells	B-hyper	chr1_13750001	0.027410084	0.624605891
BCL2	T cells	B-hyper	chr1_13750001	0.027410084	0.624605891
BCL2	T memory cells	B-hyper	chr1_13746001	0.004980075	0.792894308
BCL2	T cells	B-hyper	chr1_13746001	0.004980075	0.792894308
CYTIP	T memory cells	B-hyper	chr36_3473501	0.014692478	0.754361237
CYTIP	T memory cells	B-hyper	chr36_3485001	0.012122645	0.73521525
CYTIP	T memory cells	B-hyper	chr36_3450001	0.009003416	0.63869298
DOCK2	NK cells	B-hyper	chr4_42229001	0.059717155	0.65963723
FLI1	B cells	B-hyper	chr5_5853501	0.017541472	0.616215678
FOXP1	B cells naive	B-hyper	chr20_20878001	0.012600775	0.687075154
FOXP1	B cells naive	B-hyper	chr20_20966501	0.012397454	0.665932809
FOXP1	B cells naive	B-hyper	chr20_20809501	0.017748067	0.629669878
FOXP1	B cells naive	B-hyper	chr20_20805501	0.008248443	0.725372132
FOXP1	B cells naive	B-hyper	chr20_20805001	0.006404841	0.731390333
IFIT3	T cells	B-hyper	chr4_100001	0.008797588	0.773218594
IFIT3	Monocytes	B-hyper	chr4_100001	0.008797588	0.773218594
IFIT3	B cells	B-hyper	chr4_100001	0.008797588	0.773218594
IL17RB	T helper cells	B-hyper	chr20_36144001	0.018458025	0.663533837
IL4	T helper cells	B-hyper	chr11_20973001	0.025997809	0.590530151
IL4R	B cells naive	B-hyper	chr6_19266001	0.017954467	0.609615651
NFKB1	T helper cells	B-hyper	chr32_23983001	0.005390736	0.590831974
RORA	T helper cells	B-hyper	chr30_25755001	0.075985813	0.591881789
RORA	T cells	B-hyper	chr30_25755001	0.075985813	0.591881789

RORA	Natural killer T cells	B-hyper	chr30_25755001	0.075985813	0.591881789
RORA	T helper cells	B-hyper	chr30_25702501	0.074142641	0.608536326
RORA	T cells	B-hyper	chr30_25702501	0.074142641	0.608536326
RORA	Natural killer T cells	B-hyper	chr30_25702501	0.074142641	0.608536326
RORA	T helper cells	B-hyper	chr30_25605501	0.032570371	0.722975957
RORA	T cells	B-hyper	chr30_25605501	0.032570371	0.722975957
RORA	Natural killer T cells	B-hyper	chr30_25605501	0.032570371	0.722975957
RORA	T helper cells	B-hyper	chr30_25668001	0.019576872	0.621820977
RORA	T cells	B-hyper	chr30_25668001	0.019576872	0.621820977
RORA	Natural killer T cells	B-hyper	chr30_25668001	0.019576872	0.621820977
TMEM156	B cells memory	B-hyper	chr3_73398001	0.058696423	0.733628237
TXK	NK cells	B-hyper	chr13_43959001	0.037400777	0.588824234
TXK	T cells	B-hyper	chr13_43959001	0.037400777	0.588824234
ADAM28	B cells memory	C-hyper	chr25_33234001	0.097963537	0.685999253
ADAM28	B cells naive	C-hyper	chr25_33234001	0.097963537	0.685999253
APBB1IP	T memory cells	C-hyper	chr2_7152001	0.098757585	0.643802461
BACH2	B cells	C-hyper	chr12_49247001	0.033769949	0.70773131
BACH2	B cells naive	C-hyper	chr12_49247001	0.033769949	0.70773131
BACH2	B cells memory	C-hyper	chr12_49247001	0.033769949	0.70773131
BACH2	B cells	C-hyper	chr12_49246001	0.074117926	0.638363262
BACH2	B cells naive	C-hyper	chr12_49246001	0.074117926	0.638363262
BACH2	B cells memory	C-hyper	chr12_49246001	7.41E-02	0.638363262
BCL11A	B cells	C-hyper	chr10_60673501	0.067718872	0.587200473
BCL11A	B cells	C-hyper	chr10_60611001	0.011981926	0.594509417

BCL11A	B cells	C-hyper	chr10_60611501	0.000162282	0.670194427
BCL2	T cells	C-hyper	chr1_13878001	0.062950323	0.67997672
BCL2	T memory cells	C-hyper	chr1_13878001	0.062950323	0.67997672
BCL2	T cells	C-hyper	chr1_13746001	0.012949546	0.719343756
BCL2	T memory cells	C-hyper	chr1_13746001	0.012949546	0.719343756
BCL2	T cells	C-hyper	chr1_13860001	0.007644407	0.7489536
BCL2	T memory cells	C-hyper	chr1_13860001	0.007644407	0.7489536
CD84	B cells memory	C-hyper	chr38_21714501	0.0492882	0.702278504
CYTIP	T memory cells	C-hyper	chr36_3485001	0.022044971	0.708354999
CYTIP	T memory cells	C-hyper	chr36_3450001	0.018202776	0.628144247
FLI1	B cells	C-hyper	chr5_5853501	0.015835401	0.680938168
FOXP1	B cells naive	C-hyper	chr20_20878001	0.074396205	0.619159036
FOXP1	B cells naive	C-hyper	chr20_20805001	0.011410933	0.812952118
IKZF1	T memory cells	C-hyper	chr18_1721501	0.033531403	0.603541894
ITGB8	T regulatory cells	C-hyper	chr14_34373001	0.025529716	0.738108217
LY6E	Monocytes	C-hyper	chr13_36933501	0.02350524	0.693101151
MME	B cells naive	C-hyper	chr23_49000501	0.004007966	0.600300531
MME	B cells	C-hyper	chr23_49000501	0.004007966	0.600300531
MYB	T cells	C-hyper	chr1_27984001	0.01165958	0.591015459
NAPSA	Dendritic cells	C-hyper	chr1_106370001	0.085346249	0.754792217
PLAC8	Dendritic cells	C-hyper	chr32_7024501	0.060091174	0.718832839
RORA	T helper cells	C-hyper	chr30_25605501	0.030288129	0.738966894
RORA	T cells	C-hyper	chr30_25605501	0.030288129	0.738966894
RORA	Natural killer T cells	C-hyper	chr30_25605501	0.030288129	0.738966894

SATB1	T memory cells	C-hyper	chr23_24688501	0.038231313	0.654693318
SATB1	T cells	C-hyper	chr23_24688501	0.038231313	0.654693318
SATB1	T memory cells	C-hyper	chr23_24608501	0.023309981	0.69917531
SATB1	T cells	C-hyper	chr23_24608501	0.023309981	0.69917531
SH2D1A	T cells	C-hyper	chrX_95761001	0.04902597	0.695094935
TBX21	NK cells	C-hyper	chr9_24082001	0.019852365	0.653606809
TBX21	Natural killer T cells	C-hyper	chr9_24082001	0.019852365	0.653606809
TBX21	T helper cells	C-hyper	chr9_24082001	0.019852365	0.653606809
TXK	NK cells	C-hyper	chr13_43976501	0.013669561	0.715530065
TXK	T cells	C-hyper	chr13_43976501	0.013669561	0.715530065
TXK	NK cells	C-hyper	chr13_43976001	0.019396734	0.682103308
TXK	T cells	C-hyper	chr13_43976001	0.019396734	0.682103308
ARHGAP15	T memory cells	T-hyper	chr19_45797001	0.094930157	0.651522402
ARHGAP15	T memory cells	T-hyper	chr19_45672001	0.009607427	0.706836311
ARHGAP15	T memory cells	T-hyper	chr19_45509001	0.099258185	0.623842953
BACH2	B cells memory	T-hyper	chr12_49247001	0.04079359	0.653930385
BACH2	B cells naive	T-hyper	chr12_49247001	0.04079359	0.653930385
BACH2	B cells	T-hyper	chr12_49247001	0.04079359	0.653930385
BCL2	T cells	T-hyper	chr1_13860001	0.012425893	0.656767371
BCL2	T memory cells	T-hyper	chr1_13860001	0.012425893	0.656767371
BCL2	T cells	T-hyper	chr1_13746001	0.007245604	0.742065007
BCL2	T memory cells	T-hyper	chr1_13746001	0.007245604	0.742065007
CYTIP	T memory cells	T-hyper	chr36_3473501	0.027493938	0.685577318
CYTIP	T memory cells	T-hyper	chr36_3485001	0.016562977	0.705792302

CYTIP	T memory cells	T-hyper	chr36_3450001	0.007637485	0.607184036
FLI1	B cells	T-hyper	chr5_5853501	0.011823426	0.637224456
FOXP1	B cells naive	T-hyper	chr20_20878001	0.02819836	0.645958177
FOXP1	B cells naive	T-hyper	chr20_20805001	0.005745927	0.758683679
FOXP1	B cells naive	T-hyper	chr20_20805501	0.019613318	0.635652242
FOXP1	B cells naive	T-hyper	chr20_20808501	0.020050608	0.723497999
IFIT3	Monocytes	T-hyper	chr4_100001	0.041987007	0.694856739
IFIT3	T cells	T-hyper	chr4_100001	0.041987007	0.694856739
IFIT3	B cells	T-hyper	chr4_100001	0.041987007	0.694856739
IL17RB	T helper cells	T-hyper	chr20_36144001	0.051201034	0.630689492
IL6	T helper cells	T-hyper	chr14_36478001	0.040675109	0.657277776
IL6	Dendritic cells	T-hyper	chr14_36478001	0.040675109	0.657277776
MRC1	Monocytes	T-hyper	chr2_19128501	0.099219911	0.613722807
NCAM1	Natural killer T cells	T-hyper	chr5_19920001	0.05430983	0.596158979
NFATC2	T helper cells	T-hyper	chr24_37690501	0.033216081	0.604093172
RORA	Natural killer T cells	T-hyper	chr30_25825001	0.051956643	0.726234798
RORA	T helper cells	T-hyper	chr30_25825001	0.051956643	0.726234798
RORA	T cells	T-hyper	chr30_25825001	0.051956643	0.726234798
RORA	Natural killer T cells	T-hyper	chr30_25605501	0.029253382	0.719611869
RORA	T helper cells	T-hyper	chr30_25605501	0.029253382	0.719611869
RORA	T cells	T-hyper	chr30_25605501	0.029253382	0.719611869
RYR1	Dendritic cells	T-hyper	chr1_114487501	0.042923777	0.697831202
SATB1	T cells	T-hyper	chr23_24622501	0.061272856	0.596953596
SATB1	T memory cells	T-hyper	chr23_24622501	0.061272856	0.596953596

SATB1	T cells	T-hyper	chr23_24688501	0.06698576	0.590731815
SATB1	T memory cells	T-hyper	chr23_24688501	0.06698576	0.590731815
SH2D1A	T cells	T-hyper	chrX_95761001	0.061534081	0.608538367
TCF7	T memory cells	T-hyper	chr11_22319501	0.056825387	0.665960997
TCF7	Natural killer T cells	T-hyper	chr11_22319501	0.056825387	0.665960997
TCF7	T cells	T-hyper	chr11_22319501	0.056825387	0.665960997
TMEM156	B cells memory	T-hyper	chr3_73373501	0.022237735	0.695074816
TXK	NK cells	T-hyper	chr13_43977001	0.01308402	0.801660694
TXK	T cells	T-hyper	chr13_43977001	0.01308402	0.801660694
TXK	NK cells	T-hyper	chr13_43976501	0.019534804	0.61820186
TXK	T cells	T-hyper	chr13_43976501	0.019534804	0.61820186

DMR distribution on the immune cell type marker genes in PanglaoDB (**Figure 2.7B**). DMGs included in 11 types of immune cell markers are listed in **Table 2.4**. First, NB_DMRs was found increasingly on the marker genes of naive B-cells, T-cells, and T helper (Th) cells. Instead, NC_DMRs were found more in B-cells, NK cells, and many subtypes of T-cells. NT_DMRs were found more in naive B-cells, NK cells, and T, Th, and T memory cells, combined with NB and NC. On the contrary, it is of note that myeloid lineage cells, such as monocytes are decreased in tumors.

We then identified the most influenced genes by altered methylation among the cell type markers. **Figure 2.7C** shows the cell type marker genes highly enriched in the immune-related GO terms considering the gene expression levels. IL4 was most frequently altered in the GO terms, and the expression decreased significantly. The list of genes, including TBX21, BCL11B, UHRF1, BACH2, SH2D1A, COL4A6, PRDM11, LBH, and TXK, showed tumor-associated hypermethylation and a significant negative correlation to gene expression. We integrated RNA-seq data to show an association between methylation and gene expression in representative marker genes (**Figure 2.7D**). Among them, BACH2, a B-cell marker; SH2D1A, a T-cell marker; TXK, an NK cell marker; and UHRF1, known to be related to NK cell number, showed a significant negative correlation between the RNA expression and overall gene methylation. These results showed that the well-enriched immune cell markers in the genome-wide methylation changes are closely linked to gene expression and affect overall tumor immune cell activity.

Bisulfite-sequencing validated the tumor-associated differential methylation in immune cell marker genes.

We showed that hypermethylation and gene expression of cell-specific gene markers are inversely correlated with integration analysis of MBD-seq and RNA-seq (**Figure 2.7C-D**). Representative DMRs, which have a reverse correlation with the gene expression, verified the methylation status *in vitro* by the targeted bisulfite-sequencing (BS-seq). BACH2, an active marker gene of B cells, has hypermethylated DMRs consisting of 11 CpGs on the second intron out of six introns in tumors (benign and carcinoma). The SH2D1A gene, a T-cell activity-related marker, has a hypermethylated DMR consisting of seven CpGs in the TTS region in tumors. A representative carcinoma-related hypermethylated DMR was identified from the CpG shore location, consisting of nine CpG promotor-TSS regions of the TXK gene. A DMR harboring 22 CpGs, which were hypermethylated in carcinoma, was identified from the CpG shore region located in the second exon among 17 exons of the UHRF1 gene (**Figure 2.8A**). The four pairs of primers targeting the flanking regions of DMRs used for BS-seq are described in **Table 2.2**.

Overall, the DMRs from the MBD-seq analysis were confirmed in the targeted BS-seq. However, the methylation frequency varied from each CpG (**Figure 2.8B**). The targeted DMR of BACH2 was the most hypermethylated in benign, followed by carcinoma. DMR on the UHRF1 was most highly methylated in carcinoma, followed

by benign. The methylation levels of TXK were similarly high in benign and carcinoma. In the case of SH2D1A gene sites, only the 5th CpG site was a differentially methylated CpG in tumors. This can still be sufficiently meaningful because studies have reported that even the presence or absence of methylation of a single CpG can affect transcription level and cell type specificity [102]. **Figure 2.8C** shows the distribution of methylation percentage across samples calculated as the number of methylated CpGs/total number of clones * 100 (%). The RNA-sequencing results performed on PBMCs of CMTs and normal dogs showed a significant decrease in the expression of these four genes (**Figure 2.8D**). When compared between the methylation (**Figure 2.8C**) and gene expression (**Figure 2.8D**), overall methylation levels on the targeted regions by BS-seq were significantly opposite to RNA expression data. Targeted BS-seq results confirmed that the high-throughput sequencing analysis after methylated CpG enrichment showed relevant genome-wide methylation status in PBMC samples. It then identified DMRs that may directly link to gene expressions that have crucial roles in cell activity and populations in PBMCs. Validation of MBD-seq results through BS-seq increases the likelihood that they can be developed for clinical tumor diagnosis.

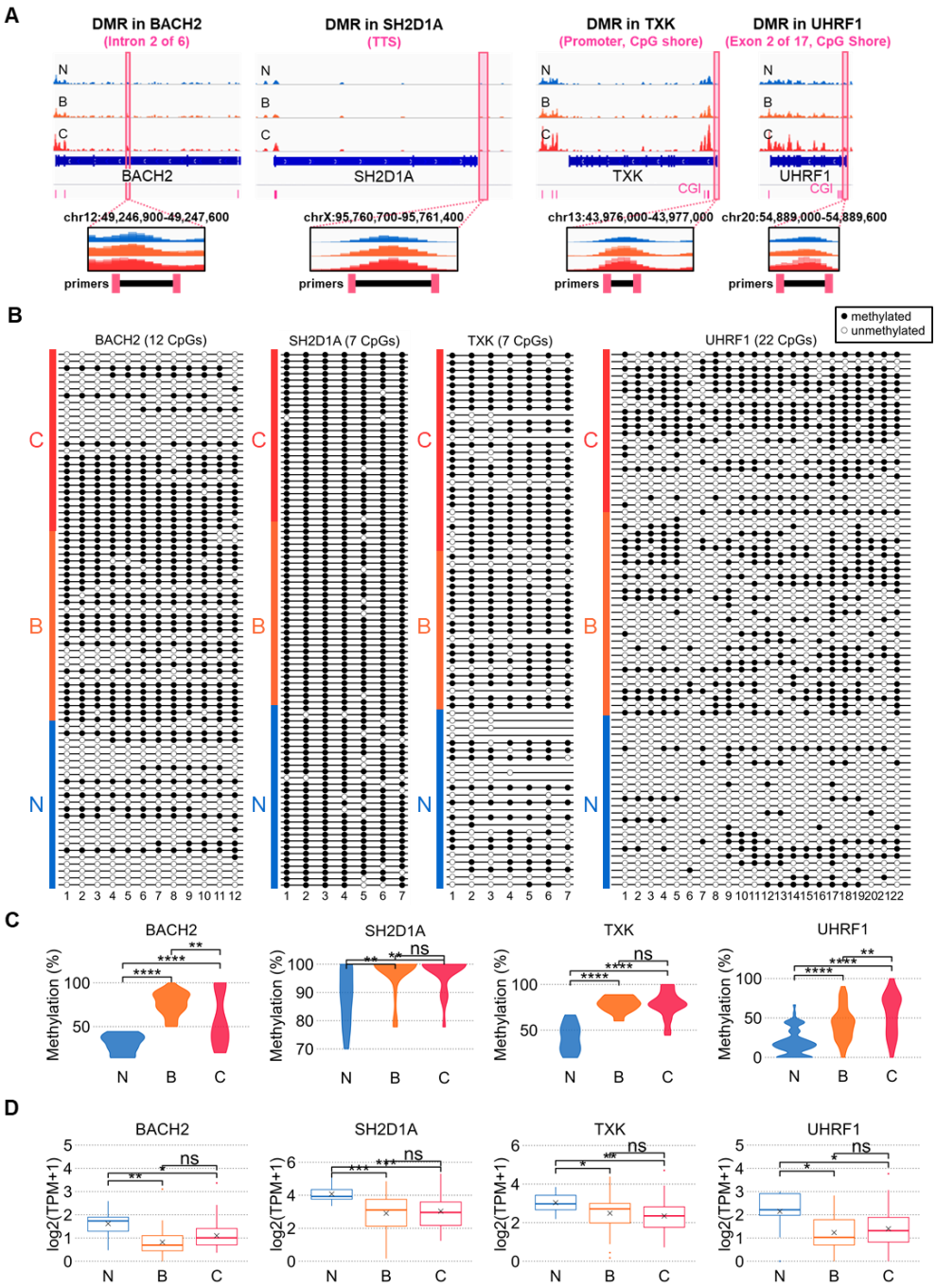

A
DMR in BACH2
(Intron 2 of 6)
DMR in SH2D1A
(TTS)
DMR in TXK
(Promoter, CpG shore)
DMR in UHRF1
(Exon 2 of 17, CpG Shore)
N
B
C
BACH2
SH2D1A
TXK
UHRF1
CGI
chr12:49,246,900-49,247,600
chrX:95,760,700-95,761,400
chr13:43,976,000-43,977,000
chr20:54,889,000-54,889,600
primers
B
BACH2 (12 CpGs)
SH2D1A (7 CpGs)
TXK (7 CpGs)
UHRF1 (22 CpGs)
methylated
unmethylated
C
Methylation (%)
ns
D
log2(TPM+1)

Figure 2.8. Targeted CpG methylation and expression analysis in representative hypermethylated genes related to immune cell activation. A) Methylation peaks in four interesting gene regions are shown. Pink dumbbells also express the loci where primers have been designed. The DMR in the BACH2 gene is located in the second intron of 6 introns, the DMR in the SH2D1A gene is located in TTS, DMR in TXK is located CpG shore promoter, and the DMR in UHRF1 is located in the second exon of 17 exons overlapped with CpG shore. B) The methylation validation for 12 CpGs in BACH2 DMR, 7 CpGs in SH2D1A and TXK DMR, and 22 CpGs in UHRF1 DMR by performing targeted bisulfite sequencing using primers listed in Table S10. Methylated CGs are indicated by black circles, and unmethylated CGs are expressed by empty (white) circles. C) Violin plots show the distribution of methylated CG (%) between groups. The total percentages of methylated CG were calculated as '(The number of methylated CG / The number of total CG in the amplified region) * 100 (%)' in each CG for every sample. D) In contrast to Violin plots in (C), Box plots show the expression levels are significantly down-regulated in Benign and Carcinoma PBMCs versus Normal PBMCs. The y-axis means the $\log_2$ (TPM+1) quantified using RNA-seq.

Computational modeling of a PBMC methylome-based two-step classifier distinguished benign and malignant as well as healthy conditions.

Methylome-based classification is a potential diagnostic method that reflects the stage or subtype of tumors. Previous studies have reported the usefulness of tissue methylation-based classifiers in diagnosing CNS tumors [103], bone sarcoma [104], and renal cell carcinoma [105]. Recently, a model using DNA methylation for discriminating cancer from para-cancerous tissue has been developed [106]. To develop a liquid biopsy-based diagnosis, we attempted to establish a model for diagnosing mammary gland tumors using genome-wide methylome data we produced. These results thus far showed immune methylome dynamics between normal and tumor PBMCs. However, it was difficult to define specific DMRs or functional terms that differentiate between benign and malignant tumors by PBMC DMRs. For efficient modeling, we devised a method to classify normal and tumor in step 1 (NT classifier), then classify benign and carcinoma in step 2 (BC classifier) and named it a two-step classifier (**Figure 2.9A**). The process for modeling and performance evaluation is depicted in **Figure 2.9B**.

First, NT classifier modeling was performed using 636 common DMRs with FDR-adjusted *p*-value <0.1 and $|\log_2 FC| \geq 0.585$ in NB DMR and NC DMR (**Figure 2.9C-E**). To overcome the problem that arising from the limited number of samples, 10-fold cross-validation (10-fold CV) was applied. The classifiers were modeled

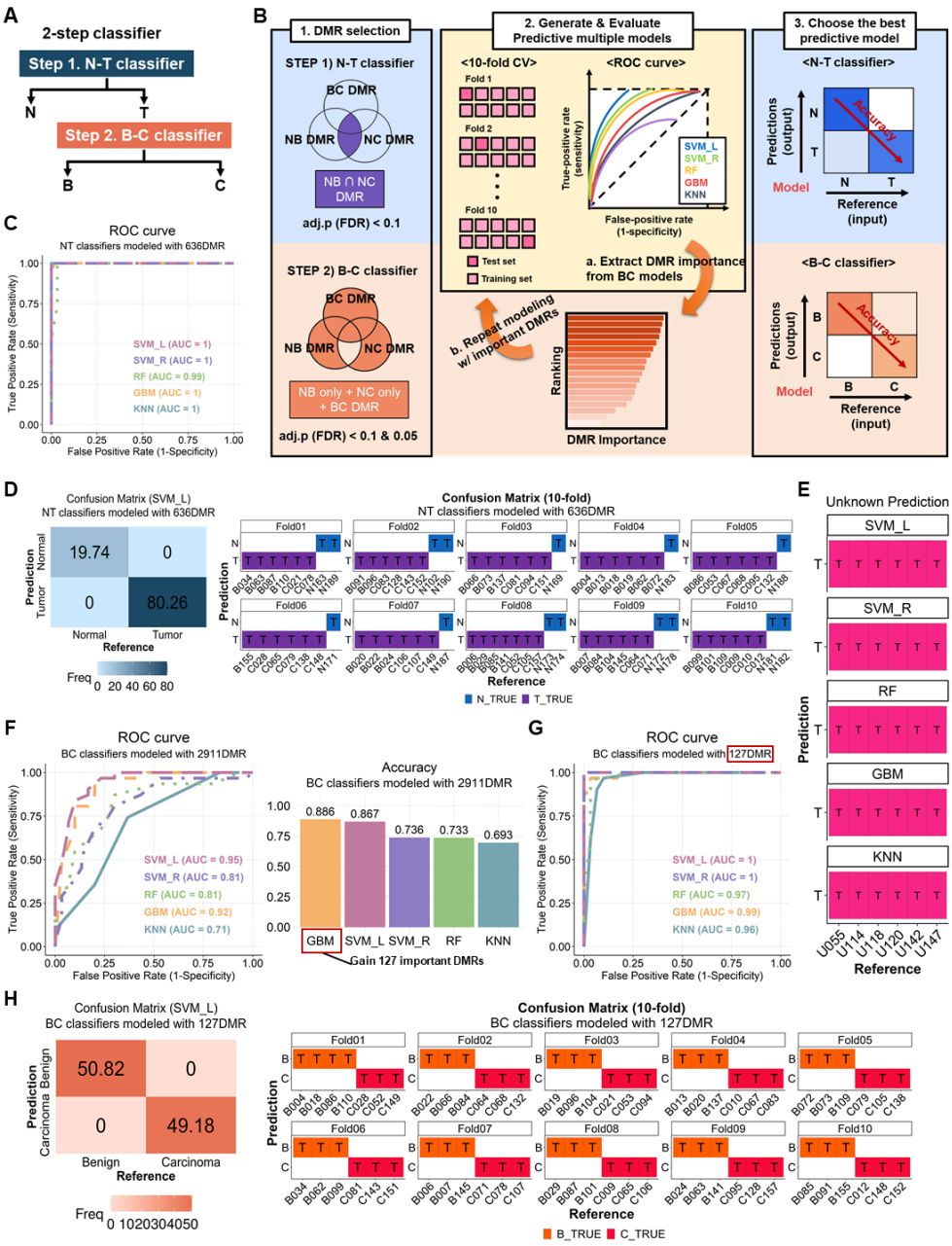

A
2-step classifier
Step 1. N-T classifier
N
T
Step 2. B-C classifier
B
C
B
1. DMR selection
STEP 1) N-T classifier
BC DMR
NB DMR
NC DMR
NB ∩ NC DMR
adj.p (FDR) < 0.1
STEP 2) B-C classifier
BC DMR
NB DMR
NC DMR
NB only + NC only + BC DMR
adj.p (FDR) < 0.1 & 0.05
2. Generate & Evaluate Predictive multiple models
<10-fold CV>
Fold 1
Fold 2
Fold 10
Test set
Training set
<ROC curve>
True-positive rate (sensitivity)
False-positive rate (1-specificity)
SVM_L
SVM_R
RF
GBM
KNN
a. Extract DMR importance from BC models
b. Repeat modeling w/ important DMRs
Ranking
DMR Importance
3. Choose the best predictive model
<N-T classifier>
Predictions (output)
Accuracy
Model
Reference (input)
<B-C classifier>
C
ROC curve
NT classifiers modeled with 636DMR
True Positive Rate (Sensitivity)
False Positive Rate (1-Specificity)
SVM_L (AUC = 1)
SVM_R (AUC = 1)
RF (AUC = 0.99)
GBM (AUC = 1)
KNN (AUC = 1)
D
Confusion Matrix (SVM_L)
NT classifiers modeled with 636DMR
Prediction
Normal
Tumor
19.74
0
0
80.26
Reference
Freq
0 20 40 60 80
Confusion Matrix (10-fold)
NT classifiers modeled with 636DMR
Fold01
Fold02
Fold03
Fold04
Fold05
Fold06
Fold07
Fold08
Fold09
Fold10
N_TRUE
T_TRUE
E
Unknown Prediction
SVM_L
SVM_R
RF
GBM
KNN
U055 U114 U118 U120 U142 U147
F
ROC curve
BC classifiers modeled with 2911DMR
SVM_L (AUC = 0.95)
SVM_R (AUC = 0.81)
RF (AUC = 0.81)
GBM (AUC = 0.92)
KNN (AUC = 0.71)
Accuracy
BC classifiers modeled with 2911DMR
0.886
0.867
0.736
0.733
0.693
GBM
SVM_L
SVM_R
RF
KNN
Gain 127 important DMRs
G
ROC curve
BC classifiers modeled with 127DMR
SVM_L (AUC = 1)
SVM_R (AUC = 1)
RF (AUC = 0.97)
GBM (AUC = 0.99)
KNN (AUC = 0.96)
H
Confusion Matrix (SVM_L)
BC classifiers modeled with 127DMR
Prediction
Benign
Carcinoma
50.82
0
0
49.18
Reference
Freq
0 10 20 30 40 50
Confusion Matrix (10-fold)
BC classifiers modeled with 127DMR
Fold01
Fold02
Fold03
Fold04
Fold05
Fold06
Fold07
Fold08
Fold09
Fold10
B_TRUE
C_TRUE

Figure 2.9. A machine learning-based diagnostic two-step classifier discriminating tumor from normal PBMCs followed by carcinoma from benign PBMCs. A) The concept of a two-step classifier for precisely distinguishing three groups (Normal, Benign, and Carcinoma). B) Schematic diagram of the diagnostic methylome-based classifier modeling. To generate the best predictive model, 10-fold cross-validation with multiple ML algorithms were employed, and then the performance of each model was evaluated. C) The ROC curves of the NT classifiers were established by SVM_L, SVM_R, RF, GBM, KNN, and logistic regression. AUC values are shown in the right-bottom area under the curves. D) Heatmap of the confusion matrix (left) for tumor detection by the SVM_L-based NT classifier, which has the best AUC value (AUC = 1) and accuracy (Accuracy = 1). The confusion matrix for 10-fold cross-validation (right) shows the prediction results for seven to nine test samples in each fold. E) Validation of the predictive performance in multiple NT classifiers. PBMC MBD-seq data from six dogs with CMT were used as the validation set. Except for the logistic classifier, which incorrectly predicted three out of six, the SVM_L, SVM_R, RF, GBM, and KNN classifiers predict tumors. F) The ROC curves (left) for the BC classifier modeled with 2911 DMRs containing 'BC_DMR' and DMRs identified 'only in NB_DMR' or 'only in NC_DMR'. BC classifiers show lower AUC values compared to NT classifiers. The bar graph (right) exhibits the highest accuracy in GBM. 127 DMRs extracted by GBM-based feature importance are used for BC classifier re-modeling. This iterative process is illustrated in the center of (B). G) The ROC curves of re-modeled BC

classifiers using 127 DMRs, which show enhanced performance compared to previous BC classifiers. H) The improved performance was confirmed via both a heatmap of the confusion matrix (left) and the 10-fold confusion matrix (right) for the final BC classifier (SVM_L) generated using 127 DMRs.

with five ML algorithms (Support Vector Machine with the linear kernel (SVM_L) or the radial kernel (SVM_R), Random Forest (RF), K-Nearest Neighbor (KNN), Gradient Boosting Machines (GBM), and Logistic Regression), and the performance of each was evaluated with the ROC curve (**Figure 2.9C**). NT classifier shows strong performance with AUC = 1 in SVM_L, SVM_R, GBM, and KNN models except for RF (AUC = 0.99) and logistic regression (AUC = 0.7). In both the representative SVM_L confusion matrix and the 10-fold validation result, it is confirmed that benign and carcinoma are classified as T (Tumor) and normal as N (Normal) (**Figure 2.9D**). The accuracy of each model is shown in **Figure 2.10A**. The high accuracy and AUC values of NT classifiers indicate that the PBMC methylome profile in tumors is completely different from that of normal samples. To evaluate the predictive ability of the NT classifiers, PBMC MBD-seq data from 6 dogs with mammary gland tumors that were not used for methylome profiling due to uncertain diagnosis were validated in the five NT classifier models (**Figure 2.9E**, the information of 6 unknown donors is listed in **Table 2.5**). All of the five NT classifiers exactly diagnosed total six PBMC samples derived from unknown MGT dogs as T (Tumor).

Next, a BC classifier was developed using significant DMRs with FDR-adjusted *p*-value <0.1 and $|\log_2 FC| \geq 0.585$ only in NB_DMR and NC_DMR and additional BC_DMR (NB only + NC only + BC DMR = total of 4,122 DMRs). Since the original BC_DMRs with FDR-adjusted *p*-value <0.1 failed to cluster benign and carcinoma (**Figure 2.1G**), the same modeling process was performed using 2,911

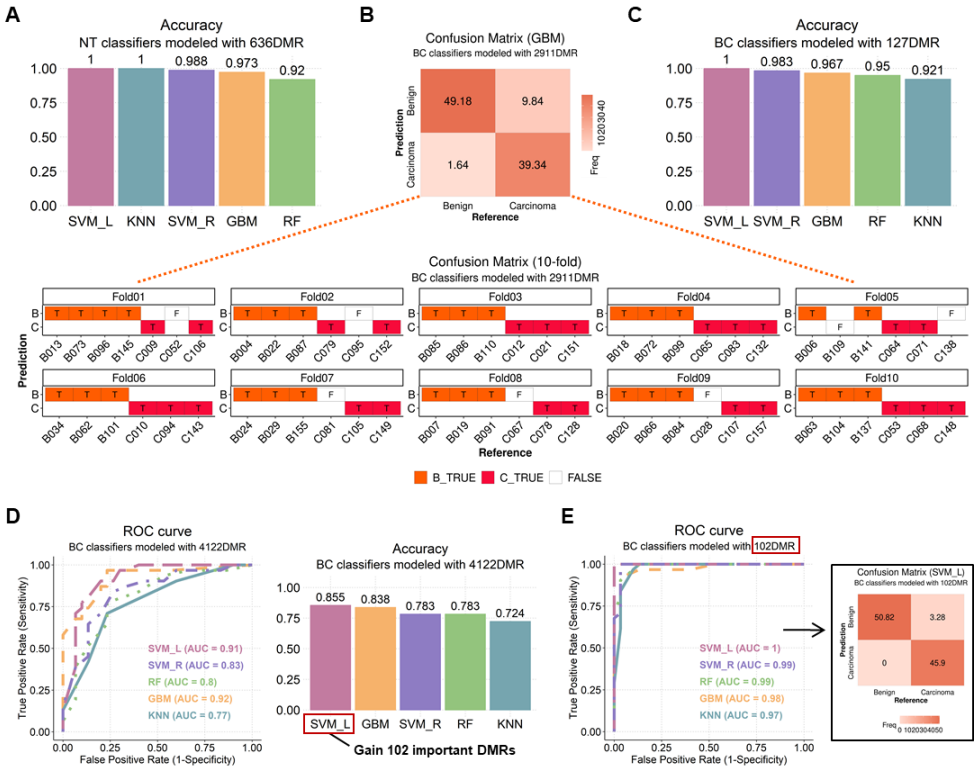


Figure 2.10. Evaluating the accuracy and predictive performance of the two-step classifier. A) Classifying accuracy of NT classifiers generated by five ML algorithms. B) Heatmap of the confusion matrix (left) for discriminating Carcinoma from Benign in the GBM-based BC classifier. The confusion matrix for 10-fold cross-validation (right) shows the prediction results for six test samples in each fold. C) Classifying accuracy of BC classifiers generated by five ML algorithms. D-E) BC classifier modeling using 4,122 DMRs (FDR <0.1) is performed in parallel with **Figure 2.9F-H**. The ROC analysis shows the performance of BC classifiers.

Table 2.5. The information of unknown dog PBMC donors (used for validation sets of NT classifier)

Donor ID	Type	Subtype	Sex	Age (years)	Breeds	Histological features
U055	T	*unknown*	FS	6	Maltese	(not inspected)
U114	T	*unknown*	F	11	Maltese	Complex mammary tumor
U118	T	*unknown*	FS	15	Yorkshire Terrier	Mammary tumor
U120	T	*unknown*	F	8	Pomeranian	Lobular hyperplasia
U142	T	*unknown*	FS	13	Maltese	Mammary tumor
U147	T	*unknown*	F	10	Shih-tzu	Mammary tumor

DMRs with FDR-adjusted *p*-value <0.05 (**Figure 2.9F-H, Figure 2.10D-E**). The BC classifier trained with the 2,911 DMRs showed the highest performance when using SVM_L (AUC = 0.95), followed by GBM (AUC = 0.92). However, the accuracy of SVM_L and GBM was 0.867 and 0.886, respectively, lower than that of the NT classifier (**Figure 2.9F**). The accuracy was about 0.85, which was inferior to that of the NT classifier (**Figure 2.10B**). To improve the performance of the BC classifier, the modeling process was repeated one more time with DMRs of high importance in the initially selected model to increase the discrimination between benign and carcinoma (depicted in **Figure 2.9B**). The performance of the models was measured using 127 DMRs, which showed high relative importance in GBM and the highest accuracy in the primary BC classifier (see the bar graph in **Figure 2.9F**). It shows improved accuracy and performance than the first-order classifier using 2,911 DMRs (**Figure 2.9G-H, Figure 2.10B-C**). As mentioned above, a parallel analysis was also executed with 4,122 DMRs with an FDR-adjusted *p*-value <0.1 (**Figure 2.10D-E**). The performance of the primary classifier was similar to that using 2,911 DMRs. However, the remodeled classifier using 102 DMRs of high importance in GBM showed slightly lower accuracy than the previous classifier in the confusion matrix of **Figure 2.11.** Both BC classifiers developed with important DMRs have the highest AUC values and accuracy in the SVM_L model. BC_DMR did not differentiate between benign and carcinoma (**Figure 2.1G**). We performed PCA analysis to evaluate whether the DMRs selected for the classifier modeling discriminate between benign and carcinoma (**Figure 2.11**). DMRs with higher

importance divided the two groups better, indicating that the GBM-based feature importance is relevant. We designed an optimal two-step classifier by utilizing various ML methods and comparing the performance of predictive models. Our result suggests a new diagnostic strategy using the PBMC methylome that can differentiate between normal, benign, and malignant tumors by liquid biopsy.

We performed permutation importance calculations to assess the biological significance of differentially methylated regions (DMRs) that distinguish between malignant and non-malignant tumors. **Figure 2.12A** displays the top 20 DMRs with high importance out of the 127 identified DMRs. These DMRs are distributed across intergenic regions, introns, and promoters, suggesting their potential involvement in epigenetic alterations associated with malignancy. Further investigations are required to determine the specific roles of these genes in mammary tumor malignancy. To assess the significance of permutation importance, a principal component analysis (PCA) was conducted using the top 10 DMRs **(Figure 2.12B)**, revealing improved separation of groups B and C compared to using all 127 DMRs (**Figure 2.11**).

We constructed a machine-learning-based classifier for diagnosing malignant tumors using PBMC Methylome. To ensure reliability of methylome classifiers, we also modeled the two-step classifier using transcriptome data with the same parameters (**Figure 2.13**). The NT classifier demonstrated the highest performance, with an AUC of 0.99 in the GBM model, followed by SVM_R with an AUC of 0.97,

which showed a similar performance to the methylome-based NT classifier. The initial BC classifier showed the highest predictive performance, with an AUC of 0.66 in SVM_R. To improve the diagnostic accuracy, we conducted secondary modeling of the BC classifier using features with high relative importance, similar to what was done in the methylome-based BC classifier. However, despite these efforts, the re-modeled BC classifier did not demonstrate improved performance, as indicated by an AUC of only 0.68 in SVM_L. This suggests that methylome data provides more informative and suitable data for discriminating malignant tumors using PBMCs compared to transcriptome data.

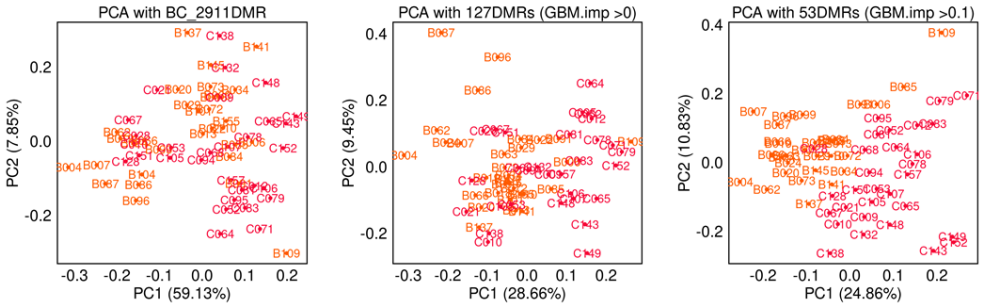


Figure 2.11. PCA analysis using DMRs involved in the BC classifiers. PCA analysis of 31 benign (orange) samples and 31 carcinoma samples (red) using 2911 DMRs (total DMRs involved in the early BC classifier), 127 DMRs (feature importance scored by GBM upper 0 used for generating the final BC classifier), and 53 DMRs (among 127 DMRs, feature importance upper 0.1). DMRs with high feature importance divide the two groups better, so the feature importance is relevant.

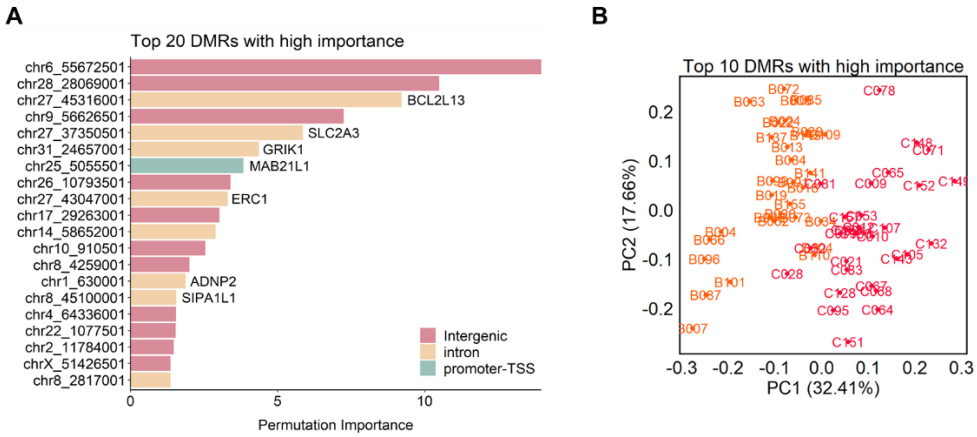
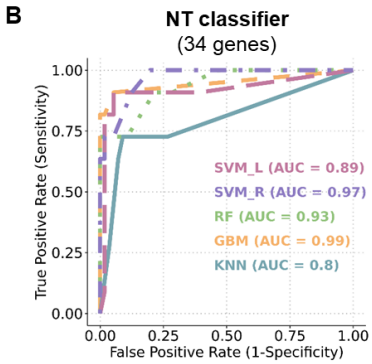
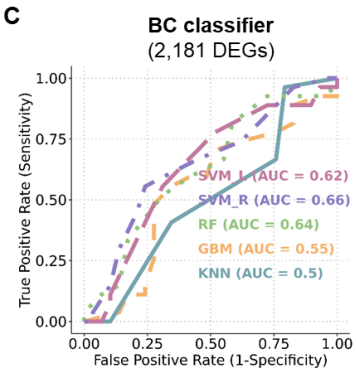


Figure 2.12. Permutation accuracy importance of DMRs used for modeling the final BC classifier. A) The top 20 DMRs with the highest importance are presented, indicating the gene symbols associated with each DMR located in intron or promoter regions. B) PCA plot conducted using the top 10 DMRs with high permutation accuracy scores to distinguish between 31 benign samples (orange) and 31 carcinoma samples (red).

A

	MBD-seq	RNA-seq
Normal	*(n=15)*	*(n=11)*
Benign	*(n=31)*	*(n=27)*
Carcinoma	*(n=30)*	*(n=29)*
Total	**76**	**67**





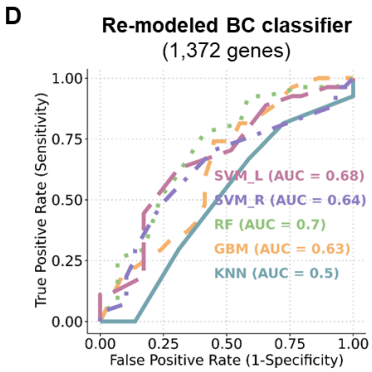


Figure 2.13. The predictive performance of transcriptome-based two-step classifier. A) The size of the dataset used for modeling classifiers. B-C) The ROC curves of the NT and BC classifiers are shown, established using SVM_L, SVM_R, RF, GBM, and KNN. The right-bottom area under the curves represents the AUC values. The NT classifier was established using 34 genes differentially expressed in benign and carcinoma versus normal PBMCs, while the BC classifier was modeled with 2,181 DEGs differentially expressed in benign versus normal PBMCs as well as carcinoma versus normal PBMCs. C) The ROC curves of the re-modeled BC classifiers using the 1,372 genes did not show improved performance compared to previous BC classifiers, unlike the methylome-based BC classifier.

Discussion

This study provides a better understanding of genome-wide epigenomic alteration, presenting a new platform for diagnosing malignant tumors from both normal and benign tumors based on liquid biopsy and DNA methylation sequencing. In several studies, blood-based DNA methylation has been profiled to develop a robust diagnostic marker for cancer. The blood-based methylation studies are broadly divided into investigating global DNA methylation [107] and gene-specific targeted DNA methylation [96]. In addition, according to the source of DNA, these studies mainly targeted circulating tumor cells (CTCs) and cell-free DNA in serum or plasma [108]. In the meantime, methylation of repetitive elements was generally investigated as surrogates for genome-wide DNA methylation measurement [109].

There have been consistent attempts to diagnose breast cancer (BC) patients using peripheral blood. BC is the most common malignant tumor in women worldwide. The prognosis of BC mainly depends on early detection; to this day, it primarily relies on mammography. CA15-3 or CA27.29 [110], approved by the FDA as blood-based protein biomarkers for BC, are recommended only for monitoring disease recurrence and therapeutic efficacy rather than diagnosis. Recently, several studies have reported genome-wide blood DNA hypomethylation in BC patients [111]. Hypermethylation of the BRCA1 gene in the blood cells and the RASSF1A gene in cfDNA has been reported in BC patients [97]. On the contrary, some studies have also reported an association between low methylation of immune cells and increased BC

risk. Thus, the evidence still needs to be more conclusive. It suggests that reliable epigenomic information based on PBMC for diagnosing BC and predicting therapeutic efficacy are needed to be studied in detail and cross-species approaches. Therefore, we performed genome-wide methylome analysis in the canine PBMC with CMT as an alternative approach for BC.

Recently, many studies have revealed that methylation, not only in the promoters but also in gene body regions such as exon, intron, and TTS regulates transcription [112]. For this reason, methylation profiling on a genome-wide scale has been steadily attempted to confirm the distribution of DMR at various locations targeting only specific genes. Since the CpG region is also an area in which epigenetic dynamics are actively occurring due to the recovery of methyltransferase and histone modifiers, it is also imperative to understand the DMR distribution from CpG islands and their surroundings (shore and shelf regions). Although CpG islands account for only 4 to 5% of the genome, approximately 70% of promoters are associated with CpG islands affecting directly annotated gene regulation [113]. Recently, the ±2 kb region on both sides of CpG islands (called 'CpG shore') has been reported to be associated with cell type specificity and highly correlated with gene expression [114]. Therefore, these methylation changes in various regions of the blood cell genome in cancer patient dogs can affect gene expressions in cancer immunity. In this study, we observed the increased methylation of CpG shore in TXK and UHRF1 strongly anti-correlated with gene expression. Although hypermethylation of CpG islands was prominent in PBMCs with carcinoma, DMRs in the CpG shore region showed

a significant inverse correlation with gene expression. However, since PBMC methylome has more variables depending on the cell type and composition, this study has limitations in elucidating the epigenetic regulation dependent on the CpG region.

PBMC has been used in various blood target studies conducted in clinical use. However, a recent study raised the question of whether PBMC transcriptome can reflect the actual state of the blood [115]. It is because PBMC contains a wide range of cells that may vary in number from patient to patient rather than a homogeneous cell population. Fortunately, projects such as the ENCODE Project and Roadmap Epigenomics have shown widespread commonality in these different cell types of transcription, but there are still distinct differences among cell types. It means that a significant difference may not be detected in PBMC if different cell types are oppositely methylated comparing two groups of DMRs. For instance, if DMRs have high methylation in T-cells but low methylation in other cells, those differences may be offset and undetected. To overcome this limitation, trials to understand PBMC data in single-cell levels via computational deconvolution or perform single-cell epigenomics are required; however, studies on PBMC methylation in single-cell resolution have not been widely conducted yet.

T-cells are vital immune mediators, differentiating into multiple subtypes in response to cancer. For this reason, T-cells have been regarded as valuable immunotherapeutic targets, and studies on tumor-infiltrating lymphocytes (TILs),

immune checkpoints, chimeric antigen receptor-engineered T cells (CAR-T), and TCR-engineered T cells (TCR-T) have been reported [116]. T-cells are programmed to attack tumors by recognizing tumor-derived antigens and secreting anti-tumorigenic cytokines [117]. Functional gene annotation analysis confirmed the aberrant methylation of genes associated with abnormal T cell differentiation as well as decreased CD8+ T cell number in cancer PBMCs. This suggests that DNA methylation is an essential key to improving the effectiveness of cancer immunotherapy in ameliorating the systemic disorder of T cells in tumors.

Hypomethylated promoters with the upregulated gene expressions of PD-1, CTLA4, and TIM3 are reported in primary breast cancer tissues [95], and CTLA4 and TIGIT promoters in colorectal cancer tissues [118]. Unlike these epigenetic characteristics shown in tumor tissues, it has been reported that methylation and expression patterns of immune checkpoints are different in peripheral blood immune cells [96]. This indicates that genome-wide scale studies on the methylome of circulating immune cells are essential to depict T-cell dysfunction and abnormal differentiation. Our PBMC methylome profiling of canine mammary tumors showed that genes involved in the differentiation and proliferation of T-cells, B-cells, and NK cells are abnormally hypermethylated. We observed increased methylation and downregulation of four representative genes (BACH2, SH2D1A, TXK, and UHRF1). BACH2 and SH2D1A are closely related to the proliferation and activation of T cells and B cells [119,120]. TXK is involved in the significant kinase signaling pathway regulating TCR signaling along with Tec family kinases ltk and Rlk [121]. The

evidence that UHRF1 is directly related to immune cell activity is insufficient. A study described that tumor-derived exosomal circulating UHRF1 promotes NK cell exhaustion in hepatocellular carcinoma [122]. Since UHRF1 is known to interact with methyltransferase to regulate the expression of other genes, it is required to study further whether methylation and expression of UHRF1 in cancer immunity are related to T-cell dysfunction.

Overall, our study highlights the unexpected epigenetic regulatory layer in silencing the activation of select circulating immune cells via hypermethylation which further associates tumor malignant states.

This hints at the possibility that the mechanism of immune exhaustion in the circulation differs from that in local TMEs. This is probably because circulating immune cells are less educated by tumors. Immune exhaustion in the peripheral blood can be explained through the expression of cell type-specific genes or kinetic pathways involved in cell activation rather than immune checkpoints. Although these assumptions require experimental validations, we exploited these genome-wide PBMC methylome profiles to develop a classification framework for biomarker discovery.

This chapter was published as:

The landscape of PBMC methylome in canine mammary tumors reveals the epigenetic regulation of immune marker genes and its potential application in predicting tumor malignancy

A-Reum Nam[1,2,3], Min Heo[3,4], Kang-Hoon Lee[1,2], Ji-Yoon Kim[1,2,3], Sung-Ho Won[3,5] and Je-Yoel Cho1[1,2,3]* (2023)

[1]Department of Biochemistry, College of Veterinary Medicine, Seoul National University, Seoul 08826, Republic of Korea. [2]BK21 Plus and Research Institute for Veterinary Science, Seoul National University, Seoul 08826, Republic of Korea. [3]Comparative Medicine Disease Research Center, Seoul National University, Seoul 08826, Republic of Korea. [4]Interdisciplinary Program of Bioinformatics, College of Natural Sciences, Seoul National University, Seoul 08826, Republic of Korea. [5]Department of Public Health Sciences, Graduate School of Public Health, Seoul National University, Seoul 08826, Republic of Korea.

BMC Genomics 24, 403 (2023). https://doi.org/10.1186/s12864-023-09471-6

General conclusion

Chapter 1

In Chapter 1, we comprehensively profiled CMT methylation and inspected its correlation with the HBC methylome. We successfully separated CMT-DMRs and subtype-DMRs, and showed their biological relevance by GO and pathway enrichment analysis. We also suggested that changes in intron-methylation play an important role in CMT by altering TF binding affinity. The importance of the intron-methylation was further confirmed in the HBC data by anti-correlation of selected gene expression with intronic hypermethylated PAX5 and hypomethylated PAX6 motifs. This study allows us to better understand both HBC and CMT at the epigenomic level, yielding new insight into cross-species mechanisms of cancer initiation and progression by DNA methylation alteration and also into the development of cancer biomarkers.

Chapter 2

In this study, we first performed the genome-wide methylome profiling in PBMCs of canine mammary gland tumors using MBD-seq. By comparing the PBMC methylomes in normal, benign, and malignant tumors, we found that benign and cancer PBMCs had distinct methylome profiles from those of normal PBMCs. We identified four hypermethylated genes (BACH2, SH2D1A, TXK, and UHRF1) involved in T-, B-, and NK cell activity and inversely correlated with gene expression by RNA-seq. Furthermore, we developed the PBMC methylome-based diagnostic classifier that distinguishes between normal and tumor and benign and malignant tumors through ML technology. This study provides an understanding of comprehensive epigenetic regulation of circulating immune cells in response to the tumor environment. We also present a new paradigm for diagnosing benign and malignant tumors based on liquid biopsy PBMC DNA methylation. Furthermore, these results provide valuable information on immune cell DNA methylation for immunotherapy, aiding in therapeutic decision-making and predicting therapeutic efficacy.

References

1 Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nature Reviews Genetics* **14**, 204-220 (2013).

2 Greenberg, M. V. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nature reviews Molecular cell biology* **20**, 590-607 (2019).

3 Stirzaker, C., Taberlay, P. C., Statham, A. L. & Clark, S. J. Mining cancer methylomes: prospects and challenges. *Trends in Genetics* **30**, 75-84 (2014).

4 Razin, A. CpG methylation, chromatin structure and gene silencing—a three-way connection. *The EMBO journal* **17**, 4905-4908 (1998).

5 Fazzari, M. J. & Greally, J. M. Epigenomics: beyond CpG islands. *Nature Reviews Genetics* **5**, 446-455 (2004).

6 Yin, Y. *et al.* Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, doi:10.1126/science.aaj2239 (2017).

7 Zakhari, S. Alcohol metabolism and epigenetics changes. *Alcohol research: current reviews* **35**, 6 (2013).

8 Cotman, C. W. & Head, E. The canine (dog) model of human aging and disease: dietary, environmental and immunotherapy approaches. *Journal of Alzheimer's Disease* **15**, 685-707 (2008).

9 LeBlanc, A. K. & Mazcko, C. N. Improving human cancer therapy through the evaluation of pet dogs. *Nature Reviews Cancer* **20**, 727-742 (2020).

10 Mestrinho, L. A. & Santos, R. R. Translational oncotargets for immunotherapy: From pet dogs to humans. *Advanced Drug Delivery Reviews* **172**, 296-313 (2021).

11 Han, L. & Zhao, Z. Contrast features of CpG islands in the promoter and other regions in the dog genome. *Genomics* **94**, 117-124 (2009).

12 Han, L., Su, B., Li, W.-H. & Zhao, Z. CpG island density and its correlations with genomic features in mammalian genomes. *Genome biology* **9**, 1-12 (2008).

13 Wai-Shin, Y., Hsu, F.-M. & Pao-Yang, C. Profiling genome-wide DNA methylation. *Epigenetics & Chromatin* (2016).

14 Torre, L. A., Islami, F., Siegel, R. L., Ward, E. M. & Jemal, A. Global Cancer in Women: Burden and Trends. *Cancer Epidemiol Biomarkers Prev* **26**, 444-457, doi:10.1158/1055-9965.EPI-16-0858 (2017).

15 Weiss, A. *et al.* Validation Study of the American Joint Committee on Cancer Eighth Edition Prognostic Stage Compared With the Anatomic

Stage in Breast Cancer. *JAMA Oncol* **4**, 203-209, doi:10.1001/jamaoncol.2017.4298 (2018).
16 Johnson, K. C. Risk factors for breast cancer. Smoking may be important. *BMJ* **322**, 365 (2001).
17 Mahdavi, M. *et al.* Hereditary breast cancer; Genetic penetrance and current status with BRCA. *J Cell Physiol* **234**, 5741-5750, doi:10.1002/jcp.27464 (2019).
18 Saleem, M. *et al.* The BRCA1 and BRCA2 Genes in Early-Onset Breast Cancer Patients. *Adv Exp Med Biol*, doi:10.1007/5584_2018_147 (2018).
19 Rajendran, B. K. & Deng, C. X. Characterization of potential driver mutations involved in human breast cancer by computational approaches. *Oncotarget* **8**, 50252-50272, doi:10.18632/oncotarget.17225 (2017).
20 Korkola, J. & Gray, J. W. Breast cancer genomes--form and function. *Curr Opin Genet Dev* **20**, 4-14, doi:10.1016/j.gde.2009.11.005 (2010).
21 Cancer Genome Atlas Research, N. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-1068, doi:10.1038/nature07385 (2008).
22 Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47**, D941-D947, doi:10.1093/nar/gky1015 (2019).
23 Cava, C., Bertoli, G. & Castiglioni, I. Integrating genetics and epigenetics in breast cancer: biological insights, experimental, computational methods and therapeutic potential. *BMC Syst Biol* **9**, 62, doi:10.1186/s12918-015-0211-x (2015).
24 Pfeifer, G. P. Defining Driver DNA Methylation Changes in Human Cancer. *Int J Mol Sci* **19**, doi:10.3390/ijms19041166 (2018).
25 Herceg, Z. & Hainaut, P. Genetic and epigenetic alterations as biomarkers for cancer detection, diagnosis and prognosis. *Mol Oncol* **1**, 26-41, doi:10.1016/j.molonc.2007.01.004 (2007).
26 Carmona, F. J. *et al.* A comprehensive DNA methylation profile of epithelial-to-mesenchymal transition. *Cancer Res* **74**, 5608-5619, doi:10.1158/0008-5472.CAN-13-3659 (2014).
27 Sproul, D. & Meehan, R. R. Genomic insights into cancer-associated aberrant CpG island hypermethylation. *Brief Funct Genomics* **12**, 174-190, doi:10.1093/bfgp/els063 (2013).
28 Han, M., Jia, L., Lv, W., Wang, L. & Cui, W. Epigenetic Enzyme Mutations: Role in Tumorigenesis and Molecular Inhibitors. *Front Oncol* **9**, 194, doi:10.3389/fonc.2019.00194 (2019).
29 Torano, E. G., Petrus, S., Fernandez, A. F. & Fraga, M. F. Global DNA hypomethylation in cancer: review of validated methods and clinical significance. *Clin Chem Lab Med* **50**, 1733-1742, doi:10.1515/cclm-2011-0902 (2012).

30 Ehrlich, M. DNA hypomethylation in cancer cells. *Epigenomics* **1**, 239-259, doi:10.2217/epi.09.33 (2009).
31 Wang, L. H., Wu, C. F., Rajasekaran, N. & Shin, Y. K. Loss of Tumor Suppressor Gene Function in Human Cancer: An Overview. *Cell Physiol Biochem* **51**, 2647-2693, doi:10.1159/000495956 (2018).
32 Kaminska, K. *et al.* Prognostic and Predictive Epigenetic Biomarkers in Oncology. *Mol Diagn Ther* **23**, 83-95, doi:10.1007/s40291-018-0371-7 (2019).
33 Locke, W. J. *et al.* DNA methylation cancer biomarkers: Translation to the clinic. *Frontiers in Genetics* **10** (2019).
34 Abdelmegeed, S. M. & Mohammed, S. Canine mammary tumors as a model for human disease. *Oncol Lett* **15**, 8195-8205, doi:10.3892/ol.2018.8411 (2018).
35 Fragomeni, S. M., Sciallis, A. & Jeruss, J. S. Molecular Subtypes and Local-Regional Control of Breast Cancer. *Surg Oncol Clin N Am* **27**, 95-120, doi:10.1016/j.soc.2017.08.005 (2018).
36 Lee, K. H., Park, H. M., Son, K. H., Shin, T. J. & Cho, J. Y. Transcriptome Signatures of Canine Mammary Gland Tumors and Its Comparison to Human Breast Cancers. *Cancers (Basel)* **10**, doi:10.3390/cancers10090317 (2018).
37 Fish, E. J. *et al.* Malignant canine mammary epithelial cells shed exosomes containing differentially expressed microRNA that regulate oncogenic networks. *BMC Cancer* **18**, 832, doi:10.1186/s12885-018-4750-6 (2018).
38 Kim, K. K. *et al.* Whole-exome and whole-transcriptome sequencing of canine mammary gland tumors. *Sci Data* **6**, 147, doi:10.1038/s41597-019-0149-8 (2019).
39 Andrews, S. (Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom, 2010).
40 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, 10-12 (2011).
41 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357 (2012).
42 Lienhard, M., Grimm, C., Morkel, M., Herwig, R. & Chavez, L. MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. *Bioinformatics* **30**, 284-286 (2013).
43 Piazza, R. *et al.* OncoScore: a novel, Internet-based tool to assess the oncogenic potential of genes. *Sci Rep* **7**, 46290, doi:10.1038/srep46290 (2017).
44 Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research* **44**, W90-W97 (2016).

45 Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128, doi:10.1186/1471-2105-14-128 (2013).
46 Huang, D. W. *et al.* DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic acids research* **35**, W169-W175 (2007).
47 Kinsella, R. J. *et al.* Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database* **2011** (2011).
48 Li, L.-C. & Dahiya, R. MethPrimer: designing primers for methylation PCRs. *Bioinformatics* **18**, 1427-1431 (2002).
49 Díez-Villanueva, A., Mallona, I. & Peinado, M. A. Wanderer, an interactive viewer to explore DNA methylation and gene expression data in human cancer. *Epigenetics & chromatin* **8**, 22 (2015).
50 Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26, doi:10.1038/nbt.1754 (2011).
51 Lokk, K. *et al.* DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome biology* **15**, 3248 (2014).
52 Unoki, M. & Nakamura, Y. Methylation at CpG islands in intron 1 of EGR2 confers enhancer-like activity. *FEBS letters* **554**, 67-72 (2003).
53 Zhang, X. *et al.* Methylation of a single intronic CpG mediates expression silencing of the PMP24 gene in prostate cancer. *The Prostate* **70**, 765-776 (2010).
54 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010).
55 Benzina, S. *et al.* Pax-5 is a potent regulator of E-cadherin and breast cancer malignant processes. *Oncotarget* **8**, 12052 (2017).
56 Leblanc, N., Harquail, J., Crapoulet, N., Ouellette, R. J. & Robichaud, G. A. Pax-5 inhibits breast cancer proliferation through MiR-215 up-regulation. *Anticancer research* **38**, 5013-5026 (2018).
57 Zong, X. *et al.* Possible role of Pax-6 in promoting breast cancer cell proliferation and tumorigenesis. *BMB reports* **44**, 595-600 (2011).
58 Eccles, M. R. & Li, C. G. PAX genes in cancer; friends or foes? *Frontiers in genetics* **3**, 6 (2012).
59 Nagy, Á., Lánczky, A., Menyhárt, O. & Győrffy, B. Validation of miRNA prognostic power in hepatocellular carcinoma using expression data of independent datasets. *Scientific reports* **8**, 9227 (2018).
60 Györffy, B. *et al.* An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast cancer research and treatment* **123**, 725-731 (2010).

61 Ball, M. P. *et al.* Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat Biotechnol* **27**, 361-368, doi:10.1038/nbt.1533 (2009).
62 Gallegos, J. E. & Rose, A. B. Intron DNA Sequences Can Be More Important Than the Proximal Promoter in Determining the Site of Transcript Initiation. *Plant Cell* **29**, 843-853, doi:10.1105/tpc.17.00020 (2017).
63 Hoivik, E. A. *et al.* DNA methylation of intronic enhancers directs tissue-specific expression of steroidogenic factor 1/adrenal 4 binding protein (SF-1/Ad4BP). *Endocrinology* **152**, 2100-2112 (2011).
64 Blattler, A. *et al.* Global loss of DNA methylation uncovers intronic enhancers in genes showing expression changes. *Genome biology* **15**, 469 (2014).
65 Jeziorska, D. M. *et al.* DNA methylation of intragenic CpG islands depends on their transcriptional activity during differentiation and disease. *Proc Natl Acad Sci U S A* **114**, E7526-E7535, doi:10.1073/pnas.1703087114 (2017).
66 Kim, D. *et al.* Population-dependent Intron Retention and DNA Methylation in Breast Cancer. *Mol Cancer Res* **16**, 461-469, doi:10.1158/1541-7786.MCR-17-0227 (2018).
67 Keshet, I., Yisraeli, J. & Cedar, H. Effect of regional DNA methylation on gene expression. *Proceedings of the National Academy of Sciences* **82**, 2560-2564 (1985).
68 Magdinier, F. *et al.* Regional methylation of the 5′ end CpG island of BRCA1 is associated with reduced gene expression in human somatic cells. *The FASEB Journal* **14**, 1585-1594 (2000).
69 Strachan, T. & Read, A. P. PAX genes. *Current opinion in genetics & development* **4**, 427-438 (1994).
70 Czerny, T. & Busslinger, M. DNA-binding and transactivation properties of Pax-6: three amino acids in the paired domain are responsible for the different sequence recognition of Pax-6 and BSAP (Pax-5). *Molecular and cellular biology* **15**, 2858-2871 (1995).
71 Lang, D., Powell, S. K., Plummer, R. S., Young, K. P. & Ruggeri, B. A. PAX genes: roles in development, pathophysiology, and cancer. *Biochemical pharmacology* **73**, 1-14 (2007).
72 Oki, S. *et al.* ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO reports* **19** (2018).
73 Gajewski, T. F., Schreiber, H. & Fu, Y.-X. Innate and adaptive immune cells in the tumor microenvironment. *Nature immunology* **14**, 1014-1022 (2013).
74 Titov, A. *et al.* Adoptive immunotherapy beyond CAR T-cells. *Cancers* **13**, 743 (2021).

75 Mosallaei, M. *et al.* PBMCs: A new source of diagnostic and prognostic biomarkers. *Archives of Physiology and Biochemistry* **128**, 1081-1087 (2022).
76 Hogg, S. J., Beavis, P. A., Dawson, M. A. & Johnstone, R. W. Targeting the epigenetic regulation of antitumour immunity. *Nature reviews Drug discovery* **19**, 776-800 (2020).
77 Villanueva, L., Álvarez-Errico, D. & Esteller, M. The contribution of epigenetics to cancer immunotherapy. *Trends in immunology* **41**, 676-691 (2020).
78 Ramchandani, S., Bhattacharya, S. K., Cervoni, N. & Szyf, M. DNA methylation is a reversible biological signal. *Proceedings of the National Academy of Sciences* **96**, 6107-6112 (1999).
79 de Vos, L. *et al.* CTLA4, PD-1, PD-L1, PD-L2, TIM-3, TIGIT, and LAG3 DNA Methylation Is Associated With BAP1-Aberrancy, Transcriptional Activity, and Overall Survival in Uveal Melanoma. *Journal of Immunotherapy* **45**, 324-334 (2022).
80 Langevin, S. M. *et al.* Peripheral blood DNA methylation profiles are indicative of head and neck squamous cell carcinoma: an epigenome-wide association study. *Epigenetics* **7**, 291-299 (2012).
81 Zhang, Y. *et al.* The signature of liver cancer in immune cells DNA methylation. *Clinical epigenetics* **10**, 1-17 (2018).
82 Marsit, C. J. *et al.* DNA methylation array analysis identifies profiles of blood-derived DNA methylation associated with bladder cancer. *Journal of Clinical Oncology* **29**, 1133 (2011).
83 Li, L. *et al.* DNA methylation signatures and coagulation factors in the peripheral blood leucocytes of epithelial ovarian cancer. *Carcinogenesis* **38**, 797-805 (2017).
84 Carson, W. F., Cavassani, K. A., Dou, Y. & Kunkel, S. L. Epigenetic regulation of immune cell functions during post-septic immunosuppression. *Epigenetics* **6**, 273-283 (2011).
85 Park, J. S. *et al.* Canine cancer immunotherapy studies: linking mouse and human. *Journal for immunotherapy of cancer* **4**, 1-11 (2016).
86 Nam, A. *et al.* Alternative methylation of intron motifs is associated with cancer-related gene expression in both canine mammary tumor and human breast cancer. *Clinical epigenetics* **12**, 1-15 (2020).
87 Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics* **14**, 178-192 (2013).
88 Franzén, O., Gan, L.-M. & Björkegren, J. L. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database* **2019** (2019).

89 Schabort, J. J. *et al.* Ank2 hypermethylation in canine mammary tumors and human breast cancer. *International journal of molecular sciences* **21**, 8697 (2020).
90 Tamura, K., Stecher, G. & Kumar, S. MEGA11: molecular evolutionary genetics analysis version 11. *Molecular biology and evolution* **38**, 3022-3027 (2021).
91 Kuhn, M. *et al.* Package 'caret'. *The R Journal* **223**, 7 (2020).
92 Greenwell, B., Boehmke, B., Cunningham, J., Developers, G. & Greenwell, M. B. Package 'gbm'. *R package version* **2** (2019).
93 Robin, X. *et al.* Package 'pROC'. *Package 'pROC'.* (2021).
94 Rex: Excel-based statistical analysis software. . doi:URL http://rexsoft.org/. (2018).
95 Sasidharan Nair, V. *et al.* DNA methylation and repressive H3K9 and H3K27 trimethylation in the promoter regions of PD-1, CTLA-4, TIM-3, LAG-3, TIGIT, and PD-L1 genes in human primary breast cancer. *Clin Epigenetics* **10**, 78, doi:10.1186/s13148-018-0512-1 (2018).
96 Elashi, A. A., Sasidharan Nair, V., Taha, R. Z., Shaath, H. & Elkord, E. DNA methylation of immune checkpoints in the peripheral blood of breast and colorectal cancer patients. *Oncoimmunology* **8**, e1542918 (2019).
97 Cao, X. *et al.* Evaluation of Promoter Methylation of RASSF1A and ATM in Peripheral Blood of Breast Cancer Patients and Healthy Control Individuals. *Int J Mol Sci* **19**, doi:10.3390/ijms19030900 (2018).
98 Iwamoto, T., Yamamoto, N., Taguchi, T., Tamaki, Y. & Noguchi, S. BRCA1 promoter methylation in peripheral blood cells is associated with increased risk of breast cancer with BRCA1 promoter methylation. *Breast Cancer Research and Treatment* **129**, 69-77, doi:10.1007/s10549-010-1188-1 (2011).
99 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature genetics* **25**, 25-29 (2000).
100 Smith, C. L. & Eppig, J. T. The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* **1**, 390-399 (2009).
101 Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences* **101**, 6062-6067 (2004).
102 Fürst, R. W., Kliem, H., Meyer, H. H. & Ulbrich, S. E. A differentially methylated single CpG-site is correlated with estrogen receptor alpha transcription. *The Journal of steroid biochemistry and molecular biology* **130**, 96-104 (2012).
103 Karimi, S. *et al.* The central nervous system tumor methylation classifier changes neuro-oncology practice for challenging brain tumor diagnoses and directly impacts patient care. *Clinical epigenetics* **11**, 1-10 (2019).

104 Wu, S. P. *et al.* DNA methylation–based classifier for accurate molecular diagnosis of bone sarcomas. *JCO precision oncology* **1**, 1-11 (2017).
105 Chen, W. *et al.* DNA methylation-based classification and identification of renal cell carcinoma prognosis-subgroups. *Cancer cell international* **19**, 1-14 (2019).
106 Ma, B. *et al.* Diagnostic classification of cancers using DNA methylation of paracancerous tissues. *Scientific Reports* **12**, 1-14 (2022).
107 Parashar, S. *et al.* DNA methylation signatures of breast cancer in peripheral T-cells. *BMC cancer* **18**, 1-9 (2018).
108 Cristall, K. *et al.* A DNA methylation-based liquid biopsy for triple-negative breast cancer. *NPJ Precision Oncology* **5**, 1-13 (2021).
109 Zheng, Y. *et al.* Prediction of genome-wide DNA methylation in repetitive elements. *Nucleic acids research* **45**, 8697-8711 (2017).
110 Hou, M.-F. *et al.* Evaluation of serum CA27. 29, CA15-3 and CEA in patients with breast cancer. *The Kaohsiung journal of medical sciences* **15**, 520-528 (1999).
111 Severi, G. *et al.* Epigenome-wide methylation in DNA from peripheral blood as a marker of risk for breast cancer. *Breast cancer research and treatment* **148**, 665-673 (2014).
112 Yang, X. *et al.* Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell* **26**, 577-590, doi:10.1016/j.ccr.2014.07.028 (2014).
113 Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A* **103**, 1412-1417, doi:10.1073/pnas.0510310103 (2006).
114 Irizarry, R. A. *et al.* The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet* **41**, 178-186, doi:10.1038/ng.298 (2009).
115 Day, K. *et al.* Transcriptomic changes in peripheral blood mononuclear cells with weight loss: systematic literature review and primary data synthesis. *Genes Nutr* **16**, 12, doi:10.1186/s12263-021-00692-6 (2021).
116 Zhang, Z. *et al.* T cell dysfunction and exhaustion in cancer. *Frontiers in cell and developmental biology* **8**, 17 (2020).
117 Kishton, R. J., Sukumar, M. & Restifo, N. P. Metabolic Regulation of T Cell Longevity and Function in Tumor Immunotherapy. *Cell Metab* **26**, 94-109, doi:10.1016/j.cmet.2017.06.016 (2017).
118 Sasidharan Nair, V., Toor, S. M., Taha, R. Z., Shaath, H. & Elkord, E. DNA methylation and repressive histones in the promoters of PD-1, CTLA-4, TIM-3, LAG-3, TIGIT, PD-L1, and galectin-9 genes in human colorectal cancer. *Clin Epigenetics* **10**, 104, doi:10.1186/s13148-018-0539-3 (2018).

119 Roychoudhuri, R. *et al.* BACH2 regulates CD8(+) T cell differentiation by controlling access of AP-1 factors to enhancers. *Nat Immunol* **17**, 851-860, doi:10.1038/ni.3441 (2016).
120 Morra, M. *et al.* Defective B cell responses in the absence of SH2D1A. *Proc Natl Acad Sci U S A* **102**, 4819-4823, doi:10.1073/pnas.0408681102 (2005).
121 Mihara, S. & Suzuki, N. Role of Txk, a member of the Tec family of tyrosine kinases, in immune-inflammatory diseases. *Int Rev Immunol* **26**, 333-348, doi:10.1080/08830180701690835 (2007).
122 Zhang, P. F. *et al.* Cancer cell-derived exosomal circUHRF1 induces natural killer cell exhaustion and may cause resistance to anti-PD1 therapy in hepatocellular carcinoma. *Mol Cancer* **19**, 110, doi:10.1186/s12943-020-01222-5 (2020).

국문초록

개 유선종양 조직 및 면역세포의 메틸롬 분석을 통한 후성유전학적 암 조절 기전 규명 및 악성종양 예측모델 개발

남 아 름
서울대학교 대학원
수의과대학 수의생명과학 전공
(지도교수: 조 제 열)

개 유성 종양은 사람 유방암과 병리학, 분자생물학적 유사성으로 인해 유방암을 연구하는 좋은 동물 모델로 알려져 있다. 또한, 사람과 개는 환경 요인에 의한 암 발병에 있어 후성유전학적 조절 기전이 유사하기 때문에 개 유선암과 사람 유방암에서의 후성유전체 연구는 중요하다. 하지만 현재까지의 연구들에서는 특정 유전자의 프로모터 메틸화에 집중이 되어있고, 유전체 전반에 거친 메틸롬 연구는 거의 진행된 바 없다. 유전체 전반에 거친 CpG 메틸화의 조절 이상은 암의 진행을 유발하며 암세포의 특정 상태를 나타내는 생체 표지자 역할을 한다고 알려져 있다. 따라서, 암 상태에서 후성유전체의

유연한 변화가 암세포나 면역세포에 어떠한 영향을 미치는 지를 연구하는 것은 매우 중요한 임상적 정보가 될 것이라고 생각한다.

본 학위논문에서는 개 유선 암의 조직과 면역세포에서 유전체 전반에 거친 광범위한 메틸롬 및 전사체 분석을 통해 진단과 치료를 위한 생체 표지자, 더 나아가 잠재적 치료 표적을 찾아내는 연구를 진행하였다. 또한, 개 유선 암에서 보이는 후성유전학적 조절 메커니즘이 사람 유방암과의 유사성을 보이는지를 비교하는 종간 분석을 수행하였다. 이 연구를 통해 본 저자는 비교 의학 관점에서 개와 인간의 후성유전체에 의한 유전자 발현 조절에 대한 이해를 높이는 동시에, 사람과 개 모두에서 암의 진단과 치료에 적용될 수 있는 새로운 전략을 제시하고자 하였다.

제 1 장에서는 개 유선 종양 및 인근 정상 조직에서의 유전체 전체 메틸롬 프로파일에 초점을 맞추어 연구를 진행하였으며, 특히 유전자의 인트론 지역이 후성유전적 조절의 잠재적 표적이 될 수 있다는 것을 증명하였다. 11 쌍의 개 유선 암 조직과 인근 정상 조직의 메틸롬을 분석한 결과, 수많은 종양 억제자와 종양 유전자의 과메틸화가 확인되었다. 특히, 인트론 부위에 과메틸화가 일어난 유전자들이 암의 항상성과 활성을 조절하는 주요 유전자 군집에 속해있다는 것을 발견하였다. 흥미롭게도, 정상 대비 암에서 과메틸화를 나타내는 PAX5 모티브 (종양 억제성)와 저메틸화를 나타내는 PAX6 모티브 (종양 유발성)가 인트론 영역에서 빈번하게 관찰되었다.

추가적으로 수행된 상관성 분석에서, 종양 억제자로 알려진 CDH5 와 LRIG1 유전자의 인트론 영역에서 PAX5 모티브의 과메틸화와 해당 유전자의 발현 간에 역 상관 관계를 발견하였으며, 반대로 종양 촉진자로 알려진 CDH2 와 ADAM19 유전자는 인트론 영역에서 저메틸화된 PAX6 모티브를 가짐과 동시에 발현이 높아지는 것을 확인할 수 있었다. 이러한 결과는 메틸화 CpG 결합 도메인 시퀀싱 (MBD-seq)뿐만 아니라 추가적인 임상 시료에서 모두 검증되었다. 더 나아가, 비교의학 연구에서 인간 유방 침윤성 암에 대한 TCGA 데이터베이스를 이용하여 이러한 유전자 인트론 영역의 과메틸화와 유전자 발현의 감소를 확인할 수 있었다. 해당 인트론 지역의 메틸화의 변화는 인간 유방암에서 유전자 발현도 변화되도록 유도했다. 이러한 연구 결과는 개 유선암과 인간 유방암에서 후성유전체 조절의 종간 보존성에 대한 증거를 제공하며, 다양한 질환에서 유전자 조절을 이해하는 데 있어 인트론 메틸화의 역할이 중요하다는 것을 시사한다.

한편, 면역세포의 암에 대한 반응은 암의 예후와 항암 치료의 효과를 결정하는 중요한 역할을 한다. 최근 많은 연구들이 면역 치료의 주요 대상인 면역 관문 (Immune checkpoint)이 후성 유전적 조절을 받는다는 증거들을 제시하고 있다. 또한, 암 진행과정에서 면역세포의 탈진 (Exhaustion), 면역 회피 (Escape)도 후성 유전체의 변화를 동반한다고 알려져 있으며, 이는 면역세포 치료나 면역 관문 억제제 (Immune checkpoint inhibitor; ICI) 치료의 예후를 결정하는 중요한 단서가 된다.

본 학위논문의 제 2 장에서는 유방 종양 환자로부터 채취한 말초 혈액 단핵구 세포 (PBMC)의 DNA 메틸롬 프로파일 차이를 조사하는 데 초점을 맞추었다. 메틸화 CpG 결합 도메인 시퀀싱 (MBD-seq)을 수행하여, 유선 종양을 가진 개와 정상 개에서 유래한 총 76 개의 PBMC 에서 전장 유전체 메틸롬을 분석하였다. 유전자 기능 군집 분석 (Gene ontology analysis; GO analysis)을 통해, T-세포 및 B-세포의 성장과 분화에 관여하는 유전자들이 종양 PBMC 에서 고도로 메틸화되어 있는 것을 확인하였다. 또한, 면역세포 증식을 조절하는 대표적인 면역 표지 유전자들 (BACH2, SH2D1A, TXK, UHRF1)에서 높아진 메틸화와 역 상관 관계를 갖는 유전자 발현을 확인할 수 있었다. 악성 종양과 양성 종양 간에 PBMC 메틸롬의 현저한 차이는 없었지만, 본 연구의 메틸롬 데이터집합을 활용하여 악성 종양을 예측하는 기계학습기반 분류기를 모델링하였다. 본 연구는 유전체 전반에 거친 순환 면역 세포의 메틸화 프로파일을 통해 암에서의 말초혈액 면역 세포의 후성 유전적 조절에 대한 통찰력 있는 정보를 제공함과 동시에, 유전체 전반에 거친 메틸롬 정보를 이용한 양성 종양과 악성 종양을 식별하는 새로운 진단 전략을 제시한다.

요약하면, 본 학위 논문에서는 개 유방 종양에서 유전체 전반에 거친 후성유전적 변화에 대한 포괄적인 정보를 제공하며, 개 유선 종양의 메틸화에 의한 암 조절에 있어 인간과 개 사이에서 흥미로운 유사성을 보여준다. 또한, 면역세포 메틸롬 데이터를 활용한 악성 종양 예측은 인간과 개의 다양한 암

유형에서 악성 종양을 진단할 수 있는 잠재적인 확장성을 제시한다. 진단 모델의 임상 적용을 위해 추가적인 검증 연구가 필요하지만, 이 연구는 개와 사람의 암 치료와 진단을 위한 중요한 기반이 될 것이라 기대한다.

주요어: 메틸롬, 전사체, 개 유선 종양, 인간 유방암, 비교 의학, 말초 혈액 단핵구 세포, 기계 학습, 생체 표지자

학번: 2016-31835