Identifying Cancer Subtypes with Aberrant DNA Methylation Regulation in Specific CpG Islands: A Study on Colorectal and Thyroid Cancer

Yeongun Lee

Department of Medical Science

The Graduate School, Yonsei University

Identifying Cancer Subtypes with Aberrant DNA Methylation Regulation in Specific CpG Islands: A Study on Colorectal and Thyroid Cancer

Directed by Professor Lark Kyun Kim

The Doctoral Dissertation
submitted to the Department of Medical Science,
the Graduate School of Yonsei University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Medical Science

Yeongun Lee

December 2023

This certifies that the Doctoral Dissertation of
Yeongun Lee is approved.

Thesis Supervisor : Lark Kyun Kim

Thesis Committee Member#1 : Hyoung Pyo Kim

Thesis Committee Member#2 : Young-Joon Kim

Thesis Committee Member#3: Tae Il Kim

Thesis Committee Member#4: Hyunki Kim

The Graduate School
Yonsei University

December 2023

ACKNOWLEDGEMENTS

In the past seven years, I have experienced significant growth and learning in the fields of epigenetics and cancer research. This period has allowed me to gain a deep understanding of essential scientific principles, become proficient in research methodologies. While gathering robust evidence to support my hypotheses was challenging at times, the thrill of validating my ideas continually drove my enthusiasm.

I would like to express my sincere gratitude to my advisor, Professor Lark Kyun Kim, for his continuous guidance throughout my research journey. I extend my heartfelt thanks to the members of my doctoral dissertation committee: Prof. Young-Joon Kim (Department of Biochemistry, Yonsei University, and CEO of Lepidyne Co., Ltd.), Prof. Tae Il Kim, Prof. Hyunki Kim, and Prof. Hyong Pyo Kim (all esteemed colleagues from Yonsei University College of Medicine). Their invaluable feedback and advice have greatly contributed to my thesis. I'm also grateful to my lab colleagues: Dr. So Hee Dho, Dr. Ji Young Kim, Dr. Su Min Kim, Jiyeon Lee, Minjeong Cho, Wonjin Woo, Hyojin Park, Jina Lee, and Eunji Lee. I'm especially thankful to Professor Jungmin Choi and Dr. Jin-Young Lee for their support and help when I was trying to understand the complexities of bioinformatics. I would also like to extend my gratitude to Dr. Jiwon Woo for all the assistance provided.

Lastly, I want to thank my family. They've been there for me throughout my academic journey. Their love and belief in me have been a great source of strength in my research.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Identifying cancer subtypes with aberrant DNA methylation regulation in specific CpG islands: a study on colorectal and thyroid cancer

Yeongun Lee

Department of Medical Science
The Graduate School, Yonsei University

(Directed by Professor Lark Kyun Kim)

Despite numerous observations regarding the relationship between DNA methylation changes and cancer progression, only a few genes have been verified as diagnostic biomarkers of cancer. To more practically detect methylation changes, I performed targeted bisulfite sequencing. Through co-analysis of RNA-seq, I identified cohort-specific DNA methylation markers. I validated that these genes have oncogenic features in CRC and that their expression levels are increased in correlation with the hypermethylation of intragenic regions. The reliable depth of the targeted bisulfite sequencing data enabled me to design highly optimized quantitative methylation-specific PCR primer sets that can successfully detect subtle changes in the methylation levels of candidate regions. Furthermore, these methylation levels can divide CRC patients into two groups denoting good and poor prognoses. My discovery of intragenic CpG island in the *PDX1*, *EN2*, and *MSX1* as DNA methylation markers of CRC suggests their promising performance as prognostic markers and their clinical application in CRC patients. In parallel, I identified heterogeneous cancer subgroups within papillary thyroid cancer using the

differentially methylated regions from targeted bisulfite sequencing of own cohort and TCGA cohort. Multiomics data (RNA-seq and ATAC-seq) from TCGA THCA project and GSE162515 were utilized to examine the molecular characteristics of these subgroups and to catalog the candidate biomarkers of PTC with worse prognosis. In this study, I present a streamlined workflow for screening clinically significant differentially methylated regions.

Key words : CpG island, DNA methylation, colorectal cancer, thyroid cancer, targeted bisulfite sequencing

Identifying cancer subtypes with aberrant DNA methylation regulation in specific CpG islands: a study on colorectal and thyroid cancer

Yeongun Lee

Department of Medical science
The Graduate School, Yonsei University

(Directed by Professor Lark Kyun Kim)

I. INTRODUCTION

Colorectal cancer (CRC) is the third most common cancer worldwide, accounting for the second-highest mortality in 2020[1]. CRC is widely known to occur due to the accumulation of genetic and epigenetic alterations. Several molecular pathways involved in the onset and development of CRC have been identified, including the adenoma–carcinoma pathway (also called the chromosomal instability sequence), the serrated neoplasia pathway, and microsatellite instability (MSI)[2,3]. The adenoma–carcinoma pathway accounts for 70–90% of CRC cases and is generally initiated by APC mutations, followed by KRAS activation or loss of TP53 function. Conversely, the serrated neoplasia pathway develops via KRAS and BRAF mutations, and epigenetic dysregulation is uniquely distinguished by the CpG island methylator phenotype (CIMP). MSI typically occurs with Lynch syndrome, mainly due to mismatch repair (MMR) gene inactivation[4-7].

Early detection of CRC is highly critical because adjuvant chemotherapy is no longer efficient and survival rates are significantly decreased for patients with CRC diagnosed at late cancer stages (stage III or IV)[8,9]. With the clinical need for early CRC diagnosis, many diagnostic and prognostic markers based on genomic alterations have been comprehensively studied. Unfortunately, few markers are used in marker development to predict the probability of metastasis or recurrence despite their unmet clinical needs.

Among the epigenetic modifications in mammals, DNA methylation plays a key role in regulating gene expression. This epigenetic regulation affects tumor suppressor gene and oncogene expression, which may lead to cancer progression. This mode of action is slightly

different among cancer types, and DNA methylation markers have been extensively established in CRC. Because of the hypomethylation and activation of repetitive sequences, such as long interspersed nuclear element-1 and Alu repeats, genomic instability is thought to occur and could boost CRC initiation[10-12]. Conversely, researchers also found a panel of genomic regions and genes aberrantly hypermethylated at the promoter regions in some CRCs, which was later identified as a type of CRC called CIMP[13]. In general, gene expression is decreased when DNA hypermethylation occurs in the promoter of a gene; thus, hypermethylated genes of the CIMP are thought to function as tumor suppressors.

Despite numerous observations regarding the relationship between DNA methylation changes and cancer progression, only a few genes, such as SEPT9 (Epi proColon), NDRG4, and BMP3 (Cologuard), have been verified as diagnostic CRC biomarkers and have been approved for commercialization via diagnostic kits[14-16]. While the surprising lack of translation into commercially viable DNA methylation-based biomarkers can be explained by methodological and experimental hurdles[17], the cornerstone of developing DNA methylation-based biomarkers is the selection of ideal genomic locations, that is, CpG islands (CGIs) and specific CpG sites[18]. For example, in several investigations, DNA methylation in the promoter region of GSTP1 has been identified as a promising diagnostic marker for hepatocellular carcinoma but with conflicting variation in terms of its specificity. It was later discovered that this variability resulted from differences in the CpG sites of the 5′ region of the GSTP1 promoter used for measuring DNA methylation levels[19]. In other words, this suggests that detection sensitivity and clinical relevance may vary depending on how the CpG sites within the same CpG island are selected.

To discover clinical biomarkers based on next-generation sequencing technology, Illumina Infinium 450K or 850K array-based detection methods have been used for massive data generation by The Cancer Genome Atlas (TCGA)[20]. This method enables me to screen and observe the methylation levels of various genes in cancer cells. Whole-genome bisulfite sequencing has emerged as a powerful method that determines DNA methylation levels on a genome-wide scale but is limited by its high cost and the time required to obtain a statistically sufficient sample size. Targeted sequencing technology has emerged as a tool for the high-throughput sequencing of genomic regions of interest. To

increase the specificity of the quantification of DNA methylation, targeted sequencing has been applied to bisulfite sequencing[21]. In detail, targeted bisulfite sequencing utilizes probes designed to bind and capture target regions for PCR-based enrichment. These capturing and enrichment steps allow me to obtain a reliable depth of DNA methylation data at the CpG site level. This method has the advantage of selecting the largest difference in DNA methylation levels and the most clinically relevant CpG sites among CpG islands or other genomic regions. However, a more straightforward methylation method, methylation-specific polymerase chain reaction (MS-PCR, MSP), has been developed and used to validate the methylation status[22]. This method offers a time- and cost-effective way of observing methylation in target regions, while designing primers and optimizing PCR conditions are relatively laborious[23,24].

This study presents my streamlined workflow for screening clinically significant differentially methylated regions and proposes primer sequences for qMSP employed as a time- and cost- effective DNA methylation detection method for clinical applications. I preliminarily selected tumor-specific methylated regions from the Infinium 450k microarray data downloaded from TCGA. I then generated hybrid capture-based targeted bisulfite sequencing data from a South Korean CRC patient cohort at Seoul National University Hospital (SNUH). I identified cohort specific DNA methylation markers in the CpG islands of *PDX1*, *EN2*, and *MSX1* and validated tumor-specific hypermethylation levels of these three genes via optimized qMSP methods with highly sensitive primer sets. I also assessed their prognostic prediction performance and found that subgroups based on the methylation status of the identified biomarkers displayed significantly different recurrence and survival rates in CRC patients. My discovery of methylation markers in the *PDX1*, *EN2*, and *MSX1* genes suggests their potential as prognostic markers and their clinical application in CRC patients.

According to the GLOBOCAN 2020 database, thyroid cancer is ranked as the ninth most common cancer worldwide[25,26]. Among various types of thyroid cancers, papillary thyroid cancer (PTC) stands out as a common variant, believed to arise from follicular cells and characterized by unique nuclear attributes[27]. While PTC typically presents a favorable prognosis, a subset progresses to an aggressive form, underscoring the need for precise prognostic indicators. Prognostic factors for PTC, such as older age at diagnosis, male

gender, tumor size (>40mm), extrathyroidal growth, and central/lateral neck lymph node metastasis, play a pivotal role in the management and therapeutic decision-making for PTC[28,29]. However, these clinicopathological indicators do not fully account for the variable aggressiveness in PTC cases, leading to a gap in personalized therapy approaches[30]. However, the translation of these biomarkers into practical clinical applications for diagnosis and management of PTC remains limited.

The exploration of the DNA methylome of thyroid cancer has not been as extensive as in other cancer types. Furthermore, previous pan-cancer study focusing on DNA methylation patterns, particularly in promoter regions, have revealed that PTC is characterized by relatively low frequencies of both hypomethylation and hypermethylation events which can frustrate the researchers[31,32]. Additionally, DNA methylation-based research on thyroid cancer has predominantly relied on microarray platforms, constraining the screening genomic regions to predefined CpG sites, thereby potentially overlooking crucial methylation events in other genomic regions[31,33-35]. Furthermore, while numerous studies have aimed to identify biomarkers differentiating PTC from benign, normal, or other thyroid histology, there is a scarcity of research delving into genome-wide biomarkers that categorize the subtypes of PTC[36-38].

In this context, the selection of appropriate target regions for DNA methylation analysis becomes crucial. The accuracy in determining these regions directly impacts the sensitivity of both prognostic and diagnostic outcomes[17]. For that reason, identification of the CpG methylation levels in the target regions become important for drawing precise conclusions. Here, target enrichment bisulfite sequencing offers a viable solution. This method, known for its relative cost-effectiveness, allows for the detailed observation of targeted regions at a high read depth[39-41].

This study aims to unravel the subtypes of PTC by finding the differentially methylated CpG island through target enrichment bisulfite sequencing with own cohort composed of total 55 papillary thyroid cancer and their paired normal tissues. I selected potential target regions for methylation analysis based on public data (TCGA and project 107738)[34,37] following a methodology similar to a previous study[42], leading to the discovery of 329 differentially methylated regions (DMRs). These DMRs enabled me to classify own cohort and TCGA samples into two distinct PTC subgroups. I then conducted a

comprehensive molecular characterization of these subgroups, integrating high-throughput techniques such as RNA sequencing and ATAC sequencing from public datasets (TCGA and GSE162515)[34,43]. I cataloged the 7 candidate genes by integrating and assessing the differential DNA methylation, RNA expression and chromatin accessibility between PTC1 and PTC2. Finally, I introduced a sophisticated quantitative methylation-specific PCR (qMSP) system capable of accurately assessing the DNA methylation levels of candidate genes. This system utilizes carefully designed primers that specifically target regions exhibiting significant methylation differences between PTC1 and PTC2 subtypes. The precision in primer design was made possible due to the detailed resolution provided by targeted bisulfite sequencing, which successfully identified the single CpG dinucleotide methylation levels in PTC samples.

II. MATERIALS AND METHODS

1. Analysis of the public DNA methylation data for design the panel of target regions

Graphical abstract for the workflow of panel design is presented in Figure 1.

A. Target selection for TBS in colorectal cancer

For targeted bisulfite sequencing, candidate genomic DNA regions were identified using the Infinium HumanMethylation450 BeadChip data from TCGA, encompassing five primary gastrointestinal malignancies: colon adenocarcinoma (COAD), rectal adenocarcinoma (READ), liver hepatocellular carcinoma (LIHC), stomach adenocarcinoma (STAD), and pancreatic adenocarcinoma (PAAD). This data was procured from the Genomic Data Commons (GDC) Data Portal (https://portal.gdc.cancer.gov/). By utilizing the human genome reference (hg19), the beta values of individual CpG sites were averaged, designating the methylation status of their corresponding CpG island. Methylation values of non-tumorous samples were aggregated, and difference in methylation between cancerous and non-tumorous samples were computed. My selection criteria centered on CpG islands reflecting methylation variations of 20% or more in at least 20% of the patient cohort.

B. Target selection for TBS in thyroid cancer

In the thyroid cancer study, Infinium HumanMethylation450 BeadChip data of TCGA-THCA and Reduced-representation bisulfite sequencing data from GSE107738 was downloaded and preprocessed in a manner analogous to the method adopted for the colorectal cancer research. In order to appropriately adjust the size of the target panel to capture the potential candidate marker of THCA, I applied a lenient threshold for the selection of DMRs. I selected CpG islands that exhibited a methylation variance of at least 10% between normal and tumor tissues in over 10% of the overall patient cohort.

2. Design of the hybridizing probe pool

A. Probe pool design for colorectal cancer research

The probe pool was designed according to the manufacturer's instructions. Basic information regarding my target genome is as follows: Application—SeqCap Epi, Organism—Homo Sapiens, Genomic builds—hg19/GRCh37. This was followed by data input in an appropriate bed format into NimbleDesign Software (version 4.3; Roche Diagnostics, Rotkreuz, Switzerland). The total number of target regions was 18,834 (10,754 CpG islands), and the total length of the regions was 23,533,457 bp.

B. Probe pool design for thyroid cancer research

Since the probe design tool and production method I previously adopted for colorectal cancer research are no longer available, I alternatively utilized a kit from another company with comparable performance (myBaits; Arbor biosciences, Ann Arbor, Michigan, USA). The probe pool was designed according to the manufacturer's instructions. Basic information regarding my target genome is as follows: Application—SeqCap Epi, Organism—Homo Sapiens, Genomic builds—hg19/GRCh37. This was followed by data input in an appropriate bed format into NimbleDesign Software (version 4.3; Roche Diagnostics, Rotkreuz, Switzerland).

3. Tumor and adjacent healthy specimens

A total of 104 colorectal tumors and their adjacent healthy tissues were obtained from Seoul National University Hospital (SNUH; Seoul, Korea). The use of samples was approved by the Institutional Review Board of Seoul National University Hospital and carried out in accordance with the ethical standards and guidelines of the institution (IRB number: 1608-040-784).

4. Sample preparation for targeted bisulfite sequencing

A. Targeted bisulfite sequencing for colorectal cancer research

1ug of genomic DNA was used to prepare a single targeted bisulfite sequencing library. All genomic DNA of healthy and tumor samples were sheared using a focused ultrasonicator (M220; Covaris, Massachusetts, USA). The quality, quantity, and fragment size (major peak in 250–300 bp) of sheared genomic DNA were verified using a 2100 Bioanalyzer system (G2939BA; Agilent Technologies, California, USA) prior to library preparation. Sheared genomic DNA was then processed through end repair, A-tailing (Kapa Library Prep Kit for Illumina NGS Platform, 7137974001; Roche Diagnostics), and sequencing adaptor ligation steps (SeqCap Adapter Kit A, 7141530001; Roche Diagnostics). After clean-up with Agencourt AMPure XP beads (A63880, Beckman Coulter, California, USA), the DNA library was bisulfite-converted using the EZ DNA Methylation- Lightning Kit (D5031; Zymo Research, California, USA) and amplified via precapture polymerase chain reaction (PCR) using KAPA HiFi HotStart Uracil+ ReadyMix (NG SeqCap Epi Accessory Kit, 714 519001; Roche Diagnostics) with Pre-LM-PCR Oligo. The quality of the amplified, bisulfite converted library samples and their sizes (main peak in 250–300 bp) were verified using a Bio-Analyzer. 1ug of each amplified, bisulfite converted library was then combined in sets of SeqCap Epi universal and indexing oligos and bisulfite capture enhancer (SeqCap EZ HE-Oligo Kit A, 6777287001; Roche Diagnostics, Rotkreuz, Switzerland). Each pool was subsequently lyophilized using a DNA vacuum concentrator (Modulspin 31; Hanil Science Co, Ltd., Daejeon, South Korea). The dried components were resuspended in hybridization buffer (SeqCap Epi Hybridization and Wash Kit, 5634253001; Roche Diagnostics, Rotkreuz, Switzerland) and hybridized with the probe pool (SeqCap Epi Choice S, 7138938001; Roche Diagnostics, Rotkreuz, Switzerland) for 72 h at 47 °C in a thermocycler with a heated lid at 57 °C. Following incubation, libraries were captured (SeqCap Pure Capture Bead Kit, 6977952001; Roche Diagnostics, Rotkreuz, Switzerland) in a 47 °C water bath and purified at room

temperature. Captured bisulfite-converted libraries were amplified via postcapture PCR and then washed with AMPure XP beads. The quality and size (single peak in 250–300 bp) of the libraries were checked using a bioanalyzer, and samples that passed quality control were sequenced on a HiSeq2500 sequencer (Illumina, San Diego, California, USA) in paired-end mode.

B. Targeted bisulfite sequencing for thyroid cancer research

In this study, genomic DNA was extracted from the collected tissue samples using the QIAamp DNA Mini Kit (Qiagen, Hilden, Germany). as outlined in the manufacturer's recommendations and guideline. After DNA extraction, the DNA concentration and purity were assessed using Nanodrop (Thermo Fisher, Carlsbad, CA, US).

A total 500ng of genomic DNA was fragmented using M220 Focused-ultrasonicator (Covaris, Woburn, MA, US) with low-EDTA TE buffer. The quality, quantity, and fragment size (major peak in 250–300 bp) of sheared genomic DNA was verified using a 2100 Bioanalyzer system (Agilent Technologies, Santa Clara, CA, USA) prior to library preparation. The DNA library was bisulfite-converted using the EZ DNA Methylation-Gold™ Kit (Zymo Research, Irvine, CA, USA). Then, the library was prepared using the Accel-NGS® Methyl-Seq DNA library kit (Swift Biosciences, Ann Arbor, MI, USA) and other specified enzymes, buffers, and reagents in manufacturer’s protocol. Finally, 8 libraries were pooled and incubated with probe pool designed for targeting the regions of interest. After clean-up, the libraries were sequenced on HiSeq2500 sequencer (Illumina, San Diego, CA, US) with 2 X 100bp pair-end reads with unique dual index and generated 2Gb of sequencing data from each sample.

5. Preprocessing of next-generation sequencing data

A. Targeted bisulfite sequencing

Trim Galore (version 0.5.0) was used to remove the adaptor sequences from the targeted bisulfite sequencing data based on the human CpG island reference hg19 file. Bismark (version 0.19.1) was used to align sequencing reads with Bowtie2. The sort and index commands from SAMtools (version 1.9) were used. The number of methylated and unmethylated cytosines at each CpG site was listed using a Bismark methylation extractor from post-indexed data, and only those 10× or higher were selected for downstream analysis. Finally, the methylation values of CpG sites included in the same CpG island were calculated by averaging the methylation value based on the hg19 reference file. The following analyses were performed based on the assumption that the averaged value represents each respective CpG island. Targeted bisulfite sequencing data were screened for targets in which DNA methylation increased or decreased by >30% in tumor samples compared with healthy tissue samples in >50% of the 90 patients. In addition, hypermethylated CpG islands in tumor samples were further filtered to retrieve regions that showed <30% DNA methylation in the healthy tissue samples and 50% or greater DNA methylation in the tumor samples. Conversely, hypomethylated CpG islands, in which the average DNA methylation was <30% in tumor samples and greater than 50% in the healthy tissue samples, were selected. Finally, I selected CpG islands where the mean DNA methylation in healthy tissue samples and tumor samples differed by >30%.

B. RNA sequencing

For TCGA RNA-seq data preprocessing, read count data which had been aligned by HT-seq was downloaded. Each RNA-seq data was integrated into a matrix. To gain a normalized gene expression data (TPM value), the scaled-estimate value of RNA-seq data aligned by STAR was multiplied by 10^6.

For preprocessing of RNA sequencing data from GSE162515, I aligned the paired-end sequencing data in reference genome hg19 with HISAT2 (version 2.2.1) and converted the result SAM file to BAM format via SAMtools. The aligned read was quantified via htseq-count and the data from each sample was integrated into a matrix, similar to the TCGA RNA-seq case. For visualization of signal track, each BAM file was normalized and converted to bigwig file with deeptools bamCoverage (version 2.2).

C. Assay for Transposase-Accessible Chromatin (ATAC) using sequencing

For preprocessing of ATAC sequencing data from GSE162515, I adopted the standardized pipeline called PEPATAC. This pipeline utilizes TRIMMOMATIC for read trimming, refgenie and bowtie2 for building the hg19 genome assembly, and MACS2 for peak calling. Furthermore, I utilized IterativeOverlapPeakMerging[44] to generate a consensus peak set of GSE162515 cohort.

6. Analysis of next-generation sequencing data

A. Targeted bisulfite sequencing

To analyze the CpG site methylation levels in candidate CpG islands from healthy tissue and tumor samples, beta values of CpG sites in candidate CpG islands were extracted using the tabix command of SAMtools (version 1.9), and only the beta values of cytosines in the same strand of adjacent genes were used in the subsequent analysis to identify the optimal MSP target sites. To filter out the low-quality sequencing data, only sequencing data in which the methylation levels of CpG sites were present in more than 1/3 of the total CpG sites in each CpG island were used. Hierarchical clustering with Canberra distance was applied to the methylation level of each sample using the pheatmap package (version 1.0.12) in R software. Line graphs were also drawn with the same methylation data using ggplot2 (version 3.3.3) and ggsci (version 2.9) in R software. To display the methylation

differences of candidate CpG islands between healthy tissue and tumor samples, hierarchical clustering with Manhattan distance was conducted using pheatmap. Using IGV, the data regarding the average methylation levels of genes in healthy and tumor tissues were visualized in tandem with the CpG island and CpG site information.

B. RNA sequencing

The correlation of each sample on the level of gene expression pattern was analyzed from principal component analysis (PCA) generated by ggplot2 (version 3.3.3). For differential expression analysis, the read count data was normalized and compared with DESeq2 (version 3.12). The only genes with at least 1.5-fold expression difference and adjusted p value of less than 0.05 were selected. The total DEGs between each subgroup were integrated, divided with K-means clustering, and visualized in heatmap with ComplexHeatmap (version 2.16.0). To investigate the functional roles of each DEG cluster, I conducted pathway analysis by gprofiler2 (version 0.2.2).

C. ATAC sequencing

For differential accessibility analysis, I utilized DiffBind (version 3.10.0) with DESeq2. Subsequent analytical tools employed for visualization were consistent with those used in the RNA sequencing data analysis.

III. RESULTS

1. Identification of differentially methylated regions in CRC tissues by targeted bisulfite sequencing

To observe methylation levels in CRC and other types of cancers, I collected 450K microarray data of five cancer types (COAD, READ, LIHC, AD, and PAAD) from TCGA **(Figure 1)**. The beta value of each CpG site was averaged to represent the methylation value of their matched CpG island in accordance with the human genome reference (hg19). The selected CpG islands were further filtered using two criteria. One was that the difference in methylation values between healthy and tumor tissues should be more than 20%, and the other was that such a difference should be present in >20% of cancer patients. Therefore, I obtained 10,754 differentially methylated CpG islands. The selected CpG islands were designed to probe the pool using NimbleDesign (Roche), a software that predicts the coverage of the input sequence and optimizes the probe design according to its criteria so that the probe pool captures the target regions more efficiently.

Next, I performed bisulfite sequencing using the probe pool in CRC tissues. To do this, I obtained genomic DNA from the tissues of 104 Korean CRC patients (90 paired tumors and adjacent healthy tissues, an additional two healthy tissues, and 12 tumor tissues). Targeted bisulfite sequencing libraries were prepared according to the manufacturer's instructions (Roche) **(Figure 2)**, and sequencing was performed. Through targeted bisulfite sequencing of the 194 CRC tissues, I obtained the beta values of each CpG site, which were averaged to constitute the methylation value of their matched CpG island **(Figure 3)**. After obtaining the methylation values of CpG islands, I applied more stringent criteria to my data. First, the difference in the methylation values of CpG islands between paired healthy and tumor tissues (i.e., from the same patient) had to be >30%. Second, this difference had to be present in >50% of the patients. Third,

even if the difference in methylation values between healthy and tumor tissues was >30%, the lower value had to be <30%, enabling the easy optimization of MSP by maximizing the signal-to-noise ratio. Finally, to identify the differentially methylated regions that are not specific to some patients, after calculating the overall average of healthy and tumor tissues, the regions with a difference of more than 30% were selected **(Figure 4)**.

Thus, I ultimately identified 40 differentially methylated CpG islands consisting of 35 hypermethylated regions and 5 hypomethylated regions in tumor tissues. For instance, the genomic location of chromosome 7:27,147,589–27,148,389 is the intragenic region of HOXA3, where 67 CpG sites are located. On average, the methylation level in this region was 29% in healthy tissues and 78.7% in tumor tissues. This difference was observed in 83.3% of CRC patients (75 out of 90) **(Table 1)**.

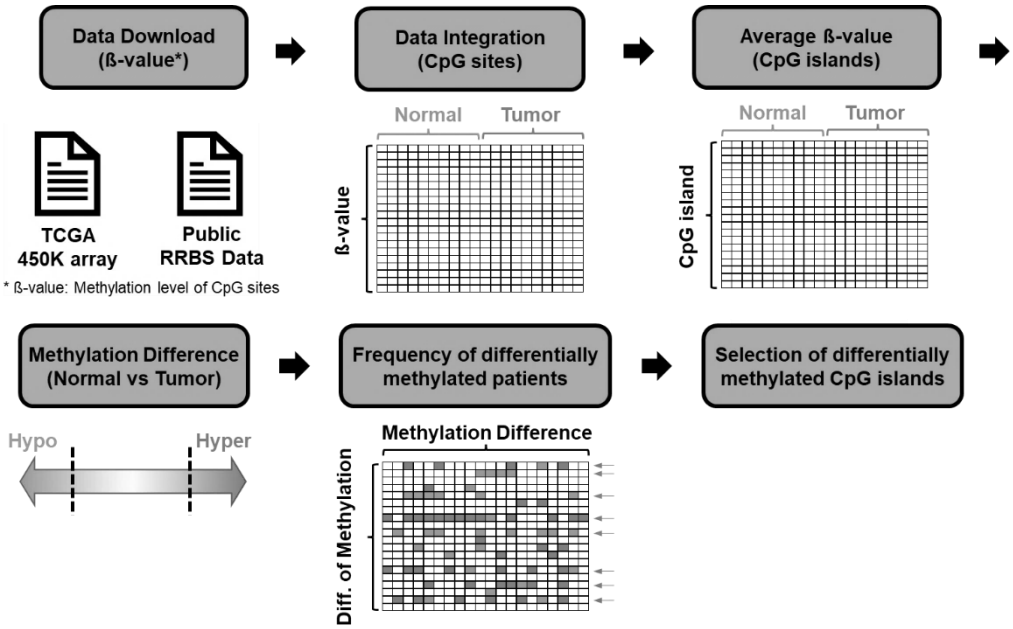

Figure 1. Preprocessing of the Infinium HumanMethylation450 BeadChip data and RRBS data for panel design of targeted bisulfite sequencing. The public data were downloaded. To estimate the methylation value of CpG islands, CpG dinucleotides on the same CpG island according to hg19 were averaged in each array datum. The methylation difference between tumor and average of normal were calculated. Based on specific criteria, I selected CpG islands where the methylation differences were observed in a significant proportion of patients.

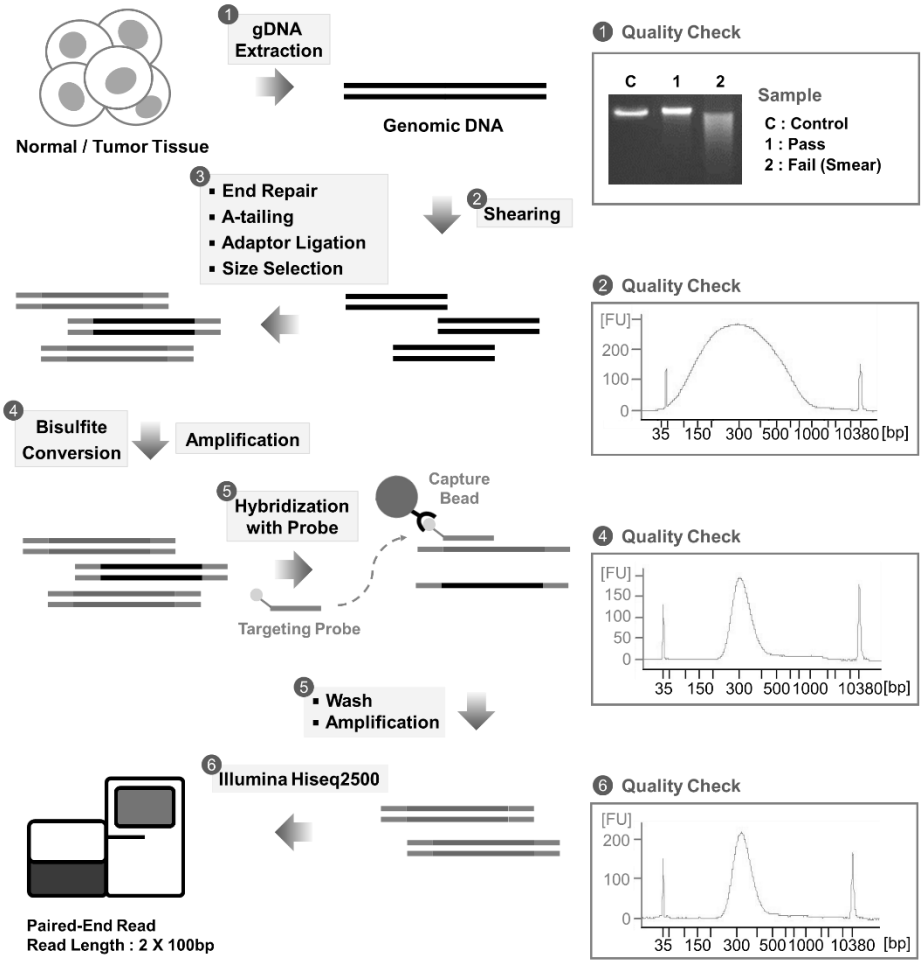

Figure 2. Preparation of the targeted DNA methylation sequencing library. Genomic DNA from healthy and tumor tissues from the colorectal and thyroid cancer cohort was extracted. Only QC-passed samples were used for the preparation of the targeted bisulfite sequencing library. Each genomic DNA was sheared to 250-300 bp, the gold standard for high-throughput sequencing. Single-stranded ends of sheared genomic DNA were repaired, followed by A-tailing, adaptor ligation, and size selection. Bisulfite conversion of genomic DNA was conducted to differentiate unmethylated cytosines from their methylated counterparts. To recover an appropriate quantity of bisulfite-converted genomic DNA, PCR amplification was performed before and after hybridization. After each amplification

step, the quality and quantity of the PCR products were confirmed using the Agilent 2100 Bioanalyzer system. The prepped samples were then used for high-throughput sequencing using HiSeq2500. Detailed library preparation procedures vary depending on the selected kit which is described comprehensively in Materials and Methods section.

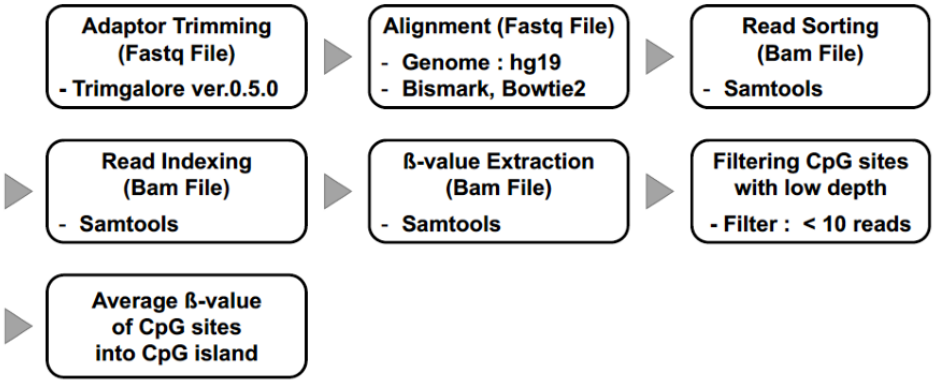

Figure 3. Preprocessing pipeline for targeted bisulfite sequencing data. Trimgalore (ver. 0.5.0) was used to trim the adaptor sequence from each targeted bisulfite sequencing data, and sequencing reads were aligned on the hg19 human genome reference using Bismark and Bowtie2. The sequencing reads were then sorted and indexed, and their methylation counts were extracted. CpG sites with a read depth below 10 were filtered out. Methylation values of CpG sites were averaged to estimate the methylation values of CpG islands.

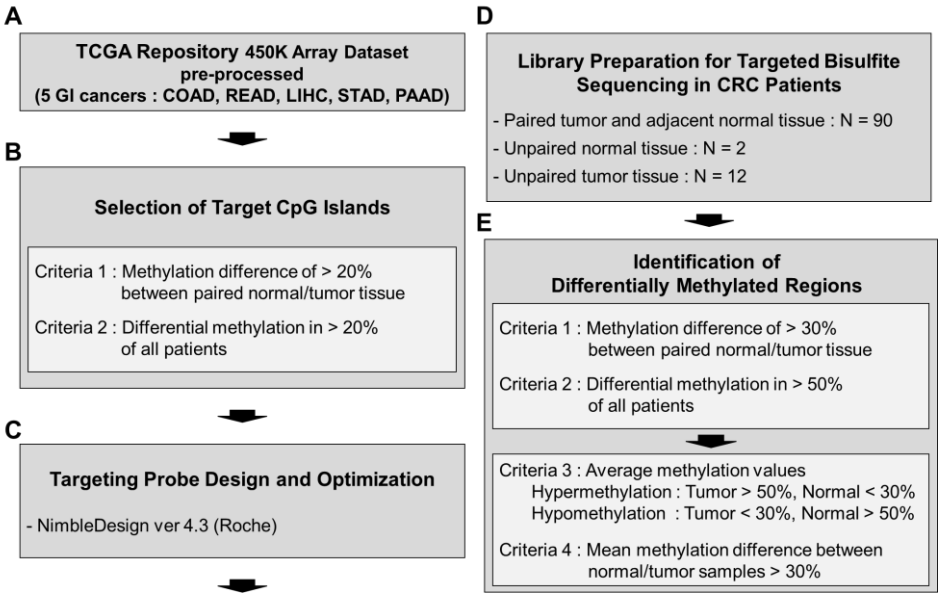

Figure 4. Overall workflow for cohort-specific DNA methylation biomarker selection in colorectal cancer. (A) Illumina Infinium 450K array data of five major gastroenterological cancers (COAD, READ, LIHC, STAD, and PAAD) downloaded from TCGA were preprocessed. **(B)** Then, 10,754 differentially methylated CpG islands (CGIs) were shortlisted from processed 450 K array data based on my criteria. **(C)** The hybridizing probe pool targeting selected CGIs was designed using NimbleDesign. **(D)** Targeted bisulfite sequencing was conducted for 104 CRC patients from the South Korean cohort, of which 90 samples were paired tumor-adjacent healthy tissue sets, while two healthy samples and ten tumor samples were unpaired. **(E)** Generated targeted bisulfite sequencing data were analyzed to select differentially methylated regions (DMRs) in tumors relative to healthy tissues, giving rise to 40 DMRs for further examination.

Table 1. Candidate CpG islands and their matched genes selected from the targeted bisulfite sequencing data of CRC study

CGI_location	CGI_info	Gene	30%_Diff	*McoM	**McaM	(McaM-McoM)
chr7:27147589–27148389	intragenic	HOXA3	0.833 (75/90)	29	78.7	49.7
chr7:27146069–27146600	intragenic	HOXA3	0.822 (74/90)	26	74	48
chr19:49669275–49669552	intragenic	TRPM4	0.811 (73/90)	24.2	73.7	49.5
chr2:54086776–54087266	promoter	GPR75-ASB3	0.8 (72/90)	23.9	74.3	50.3
chr1:200010625–200010832	intragenic	NR5A2	0.789 (71/90)	9.1	57.7	48.7
chr13:28498226–28499046	intragenic	PDX1	0.722 (65/90)	9.1	55	45.9
chr5:140857864–140858065	intragenic	PCDHGA2	0.722 (65/90)	17.3	62.8	45.5
chr7:27182613–27185562	promoter	HOXA-AS3	0.711 (64/90)	21.4	62.6	41.2
chr19:48918115–48918340	intragenic	GRIN2D	0.699 (58/83)	10.7	53.1	46.2
chr5:140864527–140864748	promoter	PCDHGA2	0.689 (62/90)	9.1	52.3	43.1
chr5:134363092–134365146	intragenic	PITX1	0.678 (61/90)	21.5	59.8	38.3
chr7:158936507–158938492	promoter	VIPR2	0.656 (59/90)	12.4	50.1	37.7
chr6:62995855–62996228	promoter	KHDRBS2	0.633 (57/90)	11.7	51.3	39.6
chr6:10398573–10398812	intragenic	TFAP2A	0.633 (57/90)	16.1	53	36.9
chr7:27143181–27143479	intergenic	-	0.633 (57/90)	26	62.6	36.7
chr7:24323558–24325080	promoter	NPY	0.633 (57/90)	16.5	52.7	36.2
chr8:97171805–97172022	promoter	GDF6	0.633 (57/90)	19.8	53.5	33.7
chr13:53313127–53314045	promoter	CNMD	0.622 (56/90)	15.6	50.9	35.3
chrX:142721410–142722958	promoter	SLITRK4	0.607 (54/89)	19.2	54.8	35.5
chr7:155255098–155255311	intragenic	EN2	0.6 (54/90)	17	52.2	35.2
chr13:102568425–102569495	promoter	FGF14	0.6 (54/90)	15.6	50.6	35
chrX:66766037–66766279	intragenic	AR	0.589 (53/90)	20.3	55.8	35.5
chr9:37002489–37002957	promoter	PAX5	0.589 (53/90)	22.1	56.3	34.1
chrX:101906001–101907017	promoter	ARMCX5-GPRASP2	0.578 (52/90)	21.6	58.2	36.6
chr4:111549879–111550203	intragenic	PITX2	0.578 (52/90)	22.9	53.7	30.8
chr4:4864456–4864834	intragenic	MSX1	0.573 (51/89)	29.7	64.3	35.3
chr8:72753874–72754755	promoter	MSC	0.567 (51/90)	26.7	58.7	32
chr19:46915311–46915802	intragenic	CCDC8	0.556 (50/90)	17.7	52.1	34.5
chr8:130995921–130996149	intragenic	FAM49B	0.544 (49/90)	20.9	53.1	32.1
chr2:98962873–98964187	promoter	CNGA3	0.544 (49/90)	19.6	51.7	32.1
chr2:5836068–5837643	intragenic	SOX11	0.544 (49/90)	20.8	51.7	30.9
chr11:65359292–65360328	intragenic	EHBP1L1	0.533 (48/90)	26.6	58	31.4
chr6:108495654–108495986	intragenic	NR2E1	0.533 (48/90)	21.5	52	30.5
chr1:120905971–120906396	promoter	HIST2H2BA (H2BP1)	0.533 (48/90)		28.8	59.1
chr13:70681732–70682219	promoter	KLHL1	0.5 (45/90)	25.1	55.5	30.4
chr16:87441387–87441671	intragenic	ZCCHC14	0.789 (71/90)	77.98	28.81	−49.17
chr7:5342299–5342599	intragenic	SLC29A4	0.778 (70/90)	73.15	26.4	−46.75
chr20:33762403–33762774	intragenic	PROCR	0.667 (60/90)	68.94	29.9	−39.04
chr1:235805318–235805771	intragenic	GNG4	0.567 (51/90)	62.69	29.03	−33.66
chr2:233925091–233925318	promoter	INPP5D	0.578 (52/90)	52.94	20.31	−32.63

***McoM: the mean of control (healthy) methylation**

****McaM the mean of case (cancer) methylation**

2. Selection of candidate genes for developing CRC biomarkers

The methylation location plays an important role in the correlation between methylation states and gene expression[18,45-47]. However, while it is well accepted that hypermethylation in the promoter region inhibits gene expression[48], the effect of methylation of the intragenic regions on gene expression is still controversial[49-55].

When I looked at the locations of my 40 differentially methylated CpG islands in terms of the promoter, intragenic, and intergenic regions, I observed that among the 35 hypermethylated regions in the tumor, 16 CpG islands were in the promoter region, 18 were in the intragenic region, and 1 was in the intergenic region. Among the five hypomethylated regions, one was in the promoter region, and four were in the intragenic region **(Figure 5 and Table 1)**.

After identifying the 40 differentially methylated CpG islands in CRC tissues, I next wanted to develop a system to detect methylation states in these regions in association with cancer status. To do this, I examined the regions whose methylation changes have a direct correlation with the expression changes of the related genes. I speculated that it would be much easier to detect the changes if both methylation and gene expression are increased in tumor tissues compared with healthy tissues because it is easy to determine what exists from what does not, but it is not easy to quantify its importance. Therefore, I was interested in the hypermethylated regions, particularly in intragenic regions, because it is difficult to connect the intergenic region to gene expression, and hypermethylation in the promoter is well accepted to be related to decreased gene expression. To examine gene expression, I took advantage of the TCGA RNA-seq dataset of colon adenocarcinoma **(Figure 5)**. Among the 18 hypermethylated intragenic regions, two regions were contained in the HOXA3 gene, so I sought to check the expression of 17 genes. According to the count data analyzed by DESeq2, the expression of only seven genes (*PDX1, GRIN2D, PITX1, TFAP2A, EN2, MSX1, and NR2E1*) was increased by more than two times in tumors **(Figure 5B)**. To ascertain the level of upregulation of

these seven genes, I also checked the expression of other candidate genes along with that of the seven genes in terms of the TPM value and then excluded NR2E1 due to lack of statistical significance **(Figure 5C)**. To further confirm the relationship between methylation changes and gene expression using Pearson and Spearman correlations, I used the Infinium HumanMethylation 450 BeadChip data and RNA sequencing data obtained from the same samples from TCGA-COAD. I found that the methylation level of the promoter CpG islands was inversely correlated with the expression of matched genes in tumor samples, regardless of whether it was significant **(Figure 6)**. In contrast, the methylation of some intragenic CpG islands had a positive correlation with matched gene expression **(Figure 7)**. That is, *PDX1*, *EN2*, and *MSX1* had higher expression levels in tumors than in normal tissues, and methylation and expression levels were positively correlated **(Figure 5B, 5C, and 7-8)**.

Next, I examined the relationship between the expression of the six genes obtained and the survival rate of CRC patients. The greater the role of abnormally expressed genes in tumor tissues, the lower the survival rate is. According to UALCAN analysis[56], high expression of *PDX1*, *EN2*, and *MSX1* was negatively correlated with patient survival **(Figure 5D)**. Therefore, I decided to focus on examining these three genes.

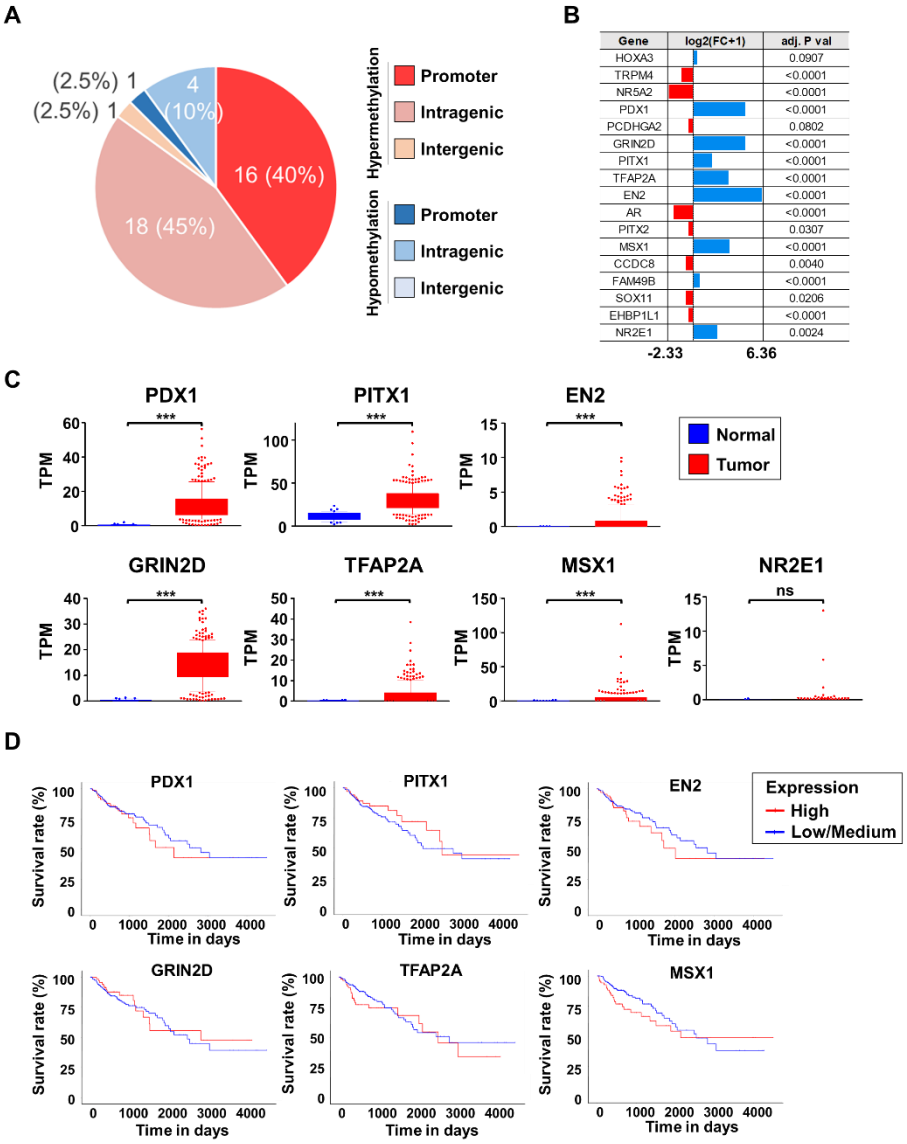

Figure 5. Streamlining of candidate DNA methylation biomarker genes based on differential gene expression and correlation with CRC patient survival outcomes. (A) Genomic location analysis of differentially methylated CGIs in targeted bisulfite sequencing data indicates that most hypermethylated regions are evenly distributed between the promoter and intragenic regions, while a larger proportion of hypomethylated

regions are in intragenic regions. My focus was on hypermethylated intragenic regions. **(B)** The expression data (read counts) downloaded from TCGA were examined to identify upregulated genes in tumor samples relative to healthy tissue samples. Downloaded RNA-seq data were processed with DESeq2 in R. **(C)** Gene expression representation of seven upregulated candidate genes in terms of TPM. Their differential expression status was further verified, and genes with nonsignificant differences were omitted from downstream analysis. Expression data between normal and tumor tissues were downloaded from TCGA, and TPM values were derived by multiplying the scaled estimate value of RNA-seq data by 106. Significance levels are presented as ns: nonsignificant, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. **(D)** Kaplan–Meier survival plots (generated by the UALCAN database) of the six upregulated genes indicated the difference between patients with high expression of the shortlisted genes (top 25%) and patients with low or medium expression (bottom 75%). Gene expression and clinical data were based on TCGA-COAD.

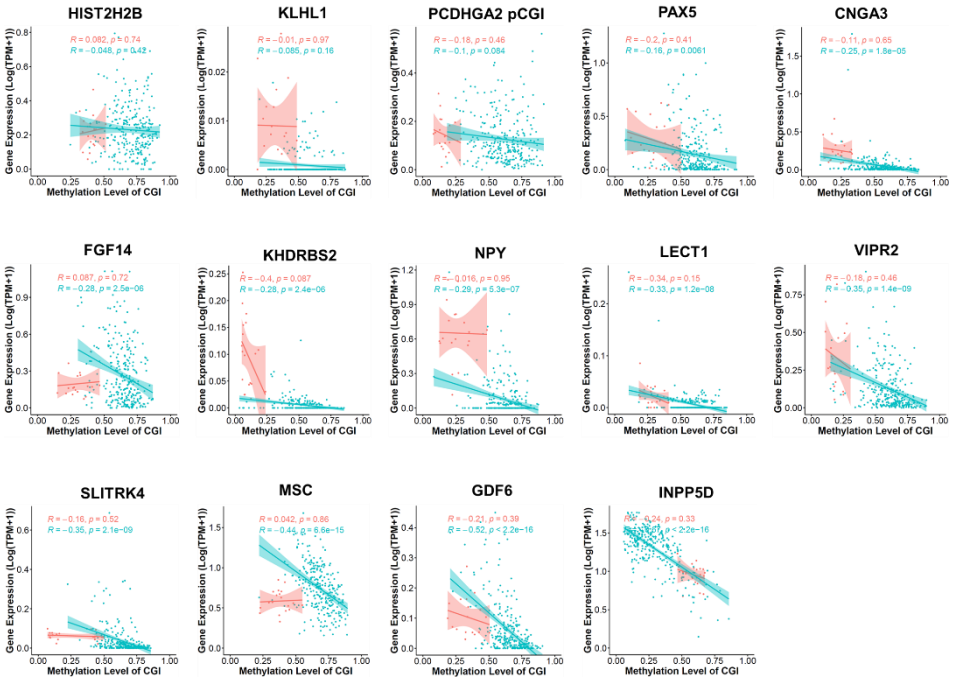

Figure 6. Pearson correlation between promoter CGI methylation and matched gene expression. To find the correlation between CGI methylation level and gene expression of candidate promoter CGIs, I called the Infinium humanmethylation450 data and RNA sequencing from TCGA-COAD (Normal = 19, Tumor = 279). RNA-Seq TPM data were log-transformed to reduce and correct the difference between Pearson and Spearman correlation coefficient [Log(TPM+1)]. Red indicates normal and blue indicates tumor. The correlation coefficient of tumor samples is organized in a table.

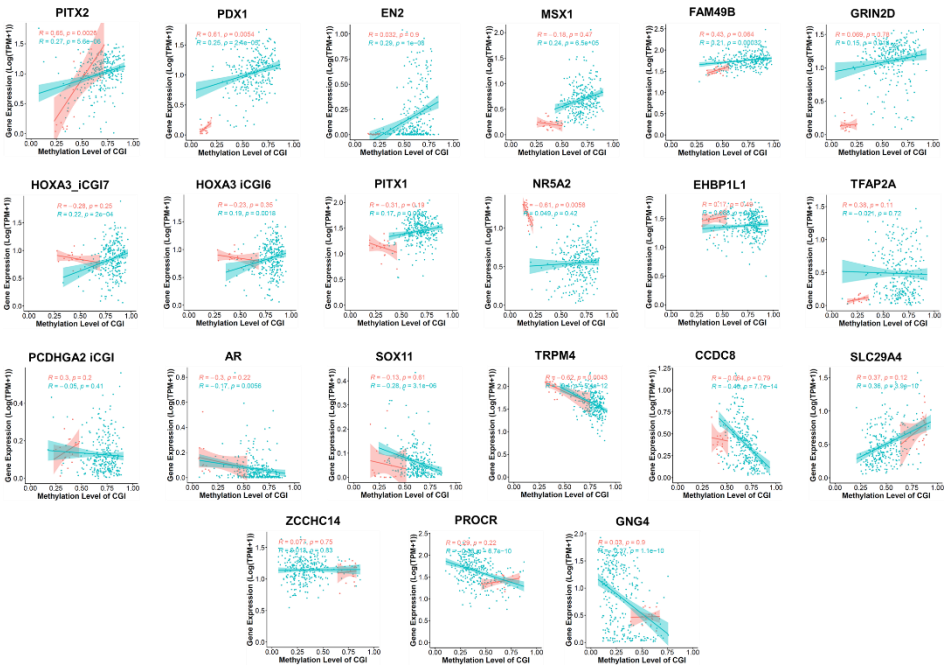

Figure 7. Pearson correlation between intragenic CGI methylation and matched gene expression. To find the correlation between CGI methylation level and gene expression of candidate intragenic CGIs, I called the Infinium humanmethylation450 data and RNA sequencing from TCGA-COAD (Normal = 19, Tumor = 279). RNA-Seq TPM data were log-transformed to reduce and correct the difference between Pearson and Spearman correlation coefficient [Log(TPM+1)]. Red indicates normal and blue indicates tumor. The correlation coefficient of tumor samples is organized in a table.

3. Overexpression of *PDX1*, *EN2*, or *MSX1* promotes cell proliferation and invasion in human colon cancer cells

Pancreatic and duodenal homeobox 1 (*PDX1*) is a critical transcription factor for pancreatic development and beta-cell maturation[57]. *PDX1* is overexpressed in pancreatic cancer cells, but its role is different at each cancer stage[58-60]. Although *PDX1* has already been reported as a potential cancer marker in CRC, it is based on the observation of *PDX1* expression in cancer cells, and its role has not been studied in detail. Homeobox protein engrailed-2 (*EN2*) is a homeobox-containing transcription factor regulating many developmental stages[61]. Very recently, *EN2* was reported to play an oncogenic role in tumor progression via CCL20 in CRC[62]. Msh homeobox 1 (*MSX1*) is also a homeobox-containing transcription factor. *MSX1* has been suggested as an mRNA biomarker for CRC, but this suggestion was based on observations, and to my knowledge, its role has never been demonstrated at the cellular level in CRC[63].

As previously mentioned, I wanted to develop a system that identifies the methylation changes of related genes that play a role in CRC. Although a literature search suggested a role for each gene in CRC, I wanted to be more confident. Thus, I transiently transfected each gene into the HCT116 colon cancer cell line and then checked the status of the cells. Proliferation was determined using CCK-8, a colorimetric reagent that indicates cell viability. Overexpression of *PDX1*, *EN2*, and *MSX1* increased cell proliferation **(Figure 8A)**. In addition, when I performed the Transwell assay, I observed that *PDX1*, *EN2*, and *MSX1* significantly promoted HCT116 cell migration **(Figure 8B)**.

Overall, I concluded that since the overexpression of *PDX1*, *EN2*, and *MSX1* is directly related to the proliferation and migration of CRC cells, if the methylation changes in the intragenic regions of these genes are correlated with changes in gene expression, the detection of methylation changes in my marker regions would be able to predict cellular conditions.

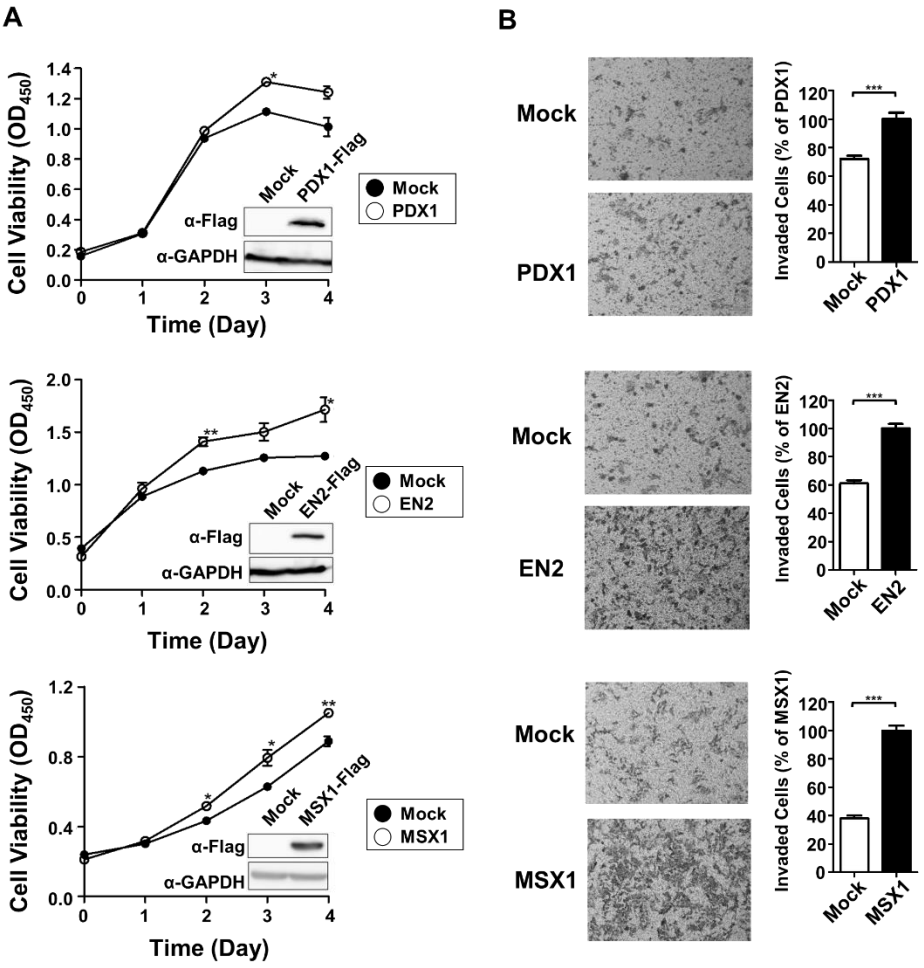

Figure 8. Selected candidate DNA methylation biomarker genes drive oncogenic properties by promoting cell proliferation and migration in vitro. (A) The cell proliferation test with CCK-8 reagent indicated that overexpression of *PDX1*, *EN2*, and *MSX1* promotes proliferation of the HCT116 colon cancer cell line. The overexpression of each gene was verified through FLAG-tag capture. **(B)** Transwell invasion assays with HCT116 cells overexpressing *PDX1*, *EN2*, and *MSX1* were conducted, and invading cells were stained with crystal violet. Overexpression of *PDX1*, *EN2*, and *MSX1* was found to accelerate migration and confer invasive properties.

4. Design of MSP primers for the optimal detection of methylation changes

To detect the methylation changes in my marker regions, I decided to set up a qMSP for each region, but factors had to be considered first. Since MSP is a PCR-based experiment, the choice of primer region is very important. If each of the forward and reverse primers has as many CpG sites as possible, the ideal methylation difference between healthy and tumor tissue is large. However, because it would be preferred to perform PCR of methylated primers with unmethylated primers in the same machine, too many CpG sites may cause a Tm difference between methylated and unmethylated primers. Last, I attempted to make the amplicon length 100–160 bp because longer products may not be efficiently amplified. Overall, after many trials and errors, I decided that the forward and reverse primers had at least six CpG sites in total, the Tm of each primer was 55–60 °C, and the amplicon length was 100–160 bp.

To design MSP primers specifically for the intragenic CpG island of *PDX1* (chr13:28,498,226-28,499,046), I examined the methylation changes of 80 individual CpG sites in that region. Although most CpG sites had large differences in methylation changes between tumor and healthy tissues, in an effort to identify the region that satisfies my criteria, I designed MSP primers according to the heatmap and the line graph of the methylation level for each CpG site in the candidate CpG islands **(Figure 9 and Figure10)**. Since I was interested in the methylation level of the same strand of the target CpG island, I mainly focused on the methylation level of CpG sites on the sense strand. The forward primer for *PDX1* has four CpG sites, and the reverse primer has three CpG sites. The beta value of these seven CpG sites was approximately 10% in normal tissues but 70% in tumor tissues on average. The amplicon size was 126 bp and 123 bp, and the Tm was 55–57 °C **(Figure 9A and Figure 10A)**. For *EN2* and *MSX1*, MSP primers were designed through similar efforts. In brief, the forward primer and the reverse primer for EN2 had three CpG sites. The beta value of the six CpG sites was approximately 10% in healthy tissues but 70% in tumor tissues on average. The amplicon sizes were 127 bp and 112 bp, and the Tm was 57–58 °C **(Figure 9B and Figure 10B)**. The forward primer and the reverse primer for MSX1 had three CpG sites. The beta value of the six CpG

sites was approximately 10% in healthy tissues but 70% in tumor tissues on average. The amplicon sizes were 151 bp and 144 bp, and the Tm was 55–57 °C **(Figure 9C and Figure 10C)**.

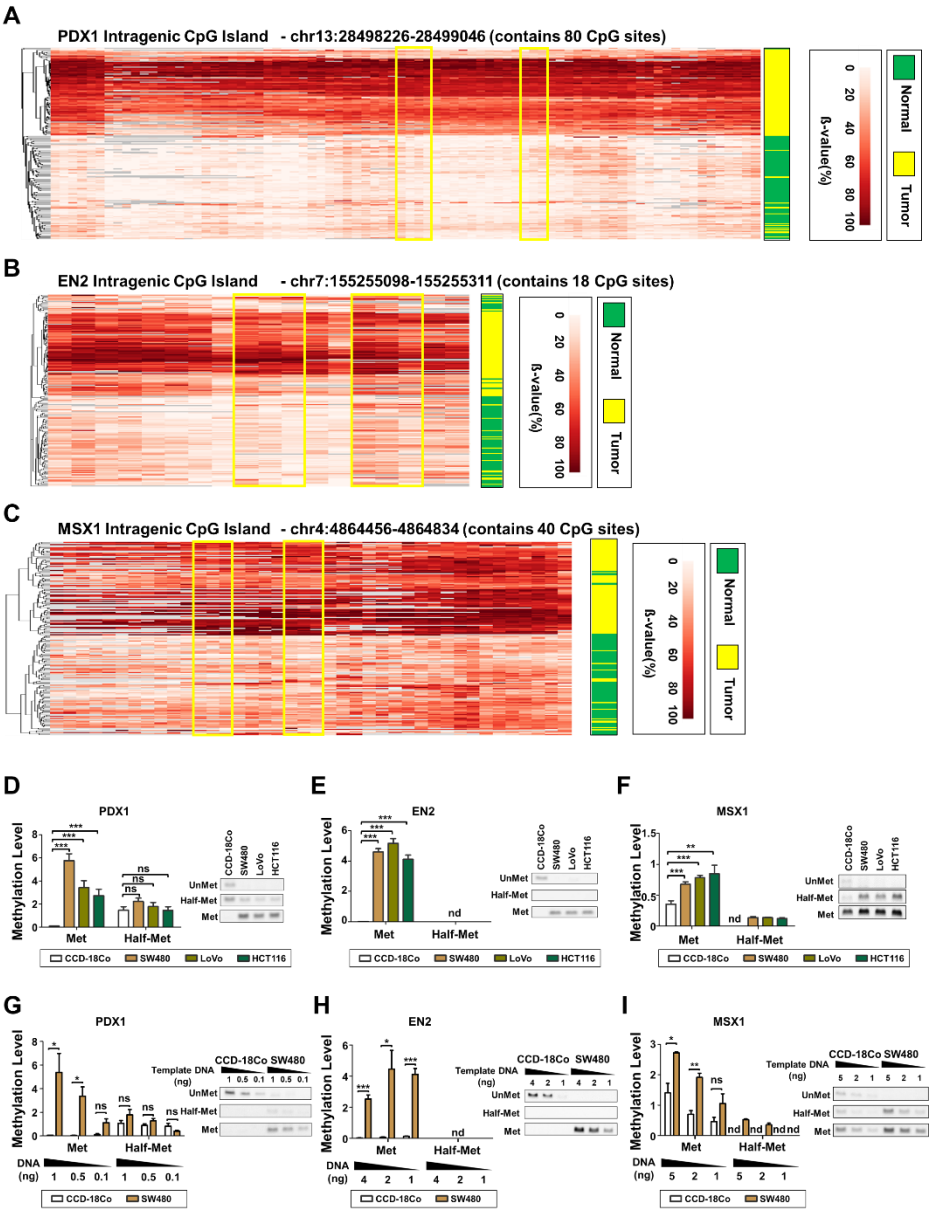

Figure 9. Optimized benchmark for primer-binding site selection and primer design in methylation-specific PCR (MSP). (A)–(C) MSP-targeting genomic regions in the intragenic CpG islands of **(A)** *PDX1*, **(B)** *EN2*, and **(C)** *MSX1* are boxed in yellow. Hierarchical clustering of healthy tissue and tumor samples of targeted bisulfite sequencing

data confirmed the hypermethylation of each target region in the tumor relative to healthy tissues. Each column corresponds to the cytosine of CpG sites within the respective intragenic CpG islands of *PDX1*, *EN2*, and *MSX1*. Low-quality sequencing data were then filtered out. **(D)–(F)** The efficacy of methylation detection and quantification of manually designed MSP primers were validated in vitro, in which three colon cancer cell lines (SW480, LoVo, HCT116) and one healthy colon cell line (CCD-18Co) were used. Agarose gel electrophoresis of quantitative MSP (qMSP) products also confirmed the methylation level detection efficacy of the designed primers for *PDX1*, *EN2*, and *MSX1*. **(G)–(I)** qMSP with varying CCD-18Co and SW480 template DNA quantities was conducted to verify DNA quantity dependent signal changes of **(G)** *PDX1*, **(H)** *EN2*, and **(I)** *MSX1* methylation. Met: MSP primer that binds to genomic DNA where all the target CpG sites are methylated. Half-Met: the MSP primer that binds with genomic DNA where some of the target CpG sites are methylated. Unmet: MSP primer that binds with genomic DNA where all the target CpG sites are not methylated. nd: not determined. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

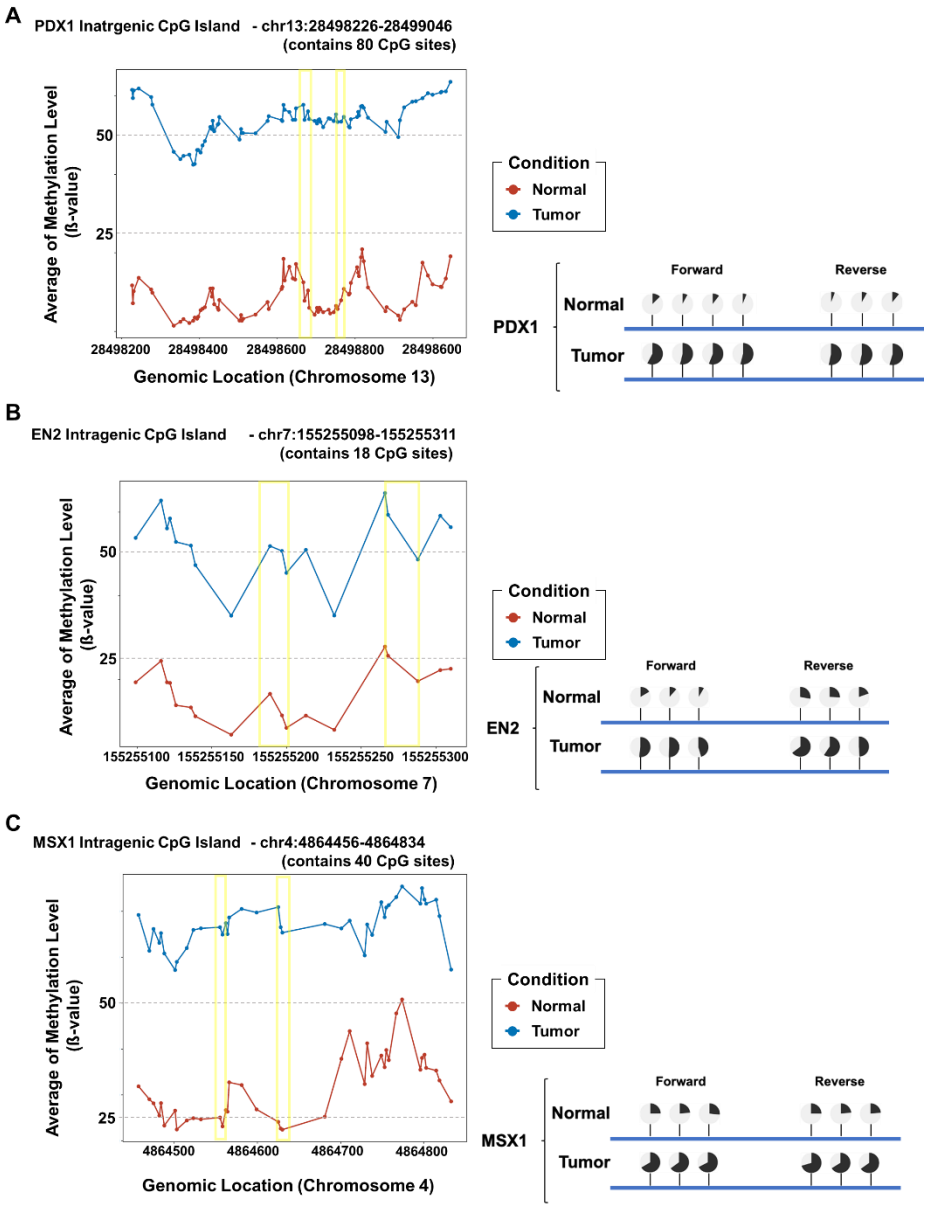

Figure 10. MSP targeting genomic regions in the intragenic CpG island of *PDX1*, *EN2*, and *MSX1*. (A)-(C) Line graph indicating the average DNA methylation level of the CpG sites in the candidate CpG island and their targeted MSP primer binding sites. Targeted bisulfite sequencing data were used in the plotting process. (Left) The red line represents

the average methylation level of healthy samples, while the blue line corresponds to tumor samples. Each dot in the line graph denotes the CpG sites included in the CpG island. The yellow boxes indicate the MSP forward and reverse primer binding sites. (Right) DNA methylation status of CpG sites in healthy and tumor colon tissues where custom-made MSP primers anneal. Each dot represents the CpG site, and the dark portion of each dot represents the average methylation level.

5. MSP primers efficiently detect the methylation states of the region of interest

Next, I wanted to confirm whether my MSP primers properly detected methylation levels. Since my MSP primers had a total of six or seven CpG sites, I not only made a primer set that retained cytosine (methylation primers) or changed all cytosine to thymine (unmethylated primers) but also created a primer set that changed only half of the cytosine to thymine (half-methylation primers). Using these primers, I performed qPCR with bisulfite treated genomic DNA from the CCD-18Co normal colon cell line and the SW480, LoVo, and HCT116 colon cancer cell lines.

In each CpG island, the methylation primer gave a PCR product in SW480, LoVo, and HCT116 cells but not in CCD-18Co cells. Unmethylated primers, on the contrary, were detected in CCD- 18Co cells but not in SW480, LoVo, and HCT116 cells. The half methylation primer failed to show clear differences among CCD-18Co, SW480, LoVo, and HCT116 cells **(Figure 9D-F)**. I quantitatively calculated the methylation level by dividing the methylation primer value or the half-methylation primer value by the unmethylated primer value. SW480, LoVo, and HCT116 cells showed significantly higher methylation levels than CCD-18Co cells when I used methylation primers but not when I used half-methylation primers **(Figure 9D–F)**. I next examined how sensitively the methylation primers could distinguish cancer cells from healthy cells in terms of the amount of template DNA. I observed the differential methylation levels of CCD-18Co and SW480 cells via qMSP and found that even 0.5 ng of template DNA, in the case of *PDX1*, was sufficient to observe the difference **(Figure 9G–I)**.

From these results, I confirmed that my MSP primers could distinguish cancer cells from normal cells very efficiently. Interestingly, although half-methylation primers also have four CpG sites where methylation levels between healthy and cancer cells are different, they could not produce clear differences when I executed MSP, suggesting that only MSP primers have more than enough CpG sites to provide substantially different results.

6. The developed MSP primers could detect dynamic changes in methylation states

I next examined whether my MSP primers could distinguish the dynamic changes in methylation levels out of concern that the data from cell lines might not sufficiently reflect physiological methylation changes due to fixed methylation values. To induce methylation changes, I used the CRISPR/dCas9-TET1 system (hereafter the dCas9-TET system), which enables me to decrease methylation levels in a location-specific manner **(Figure 11A)**[64].

After introducing the dCas9-TET system into the *PDX1* genomic region, I detected a significant reduction in methylation levels using my methylation primers, which contain seven CpG sites. However, I could not detect this difference using half methylation primers **(Figure 11B)**. I noted that *PDX1* expression was significantly decreased according to the reduction in methylation level in the intragenic region, suggesting that the methylation changes are directly related to gene expression changes **(Figure 11C)**. I obtained similar results with *EN2* and *MSX1*. I successfully detected a reduction in the methylation levels in the intragenic regions of *EN2* and *MSX1* using my methylation primers, consistent with the reduction in gene expression **(Figure 11D–G)**. Thus, I concluded that my methylation primers are sensitive enough to detect methylation changes that precede gene expression changes.

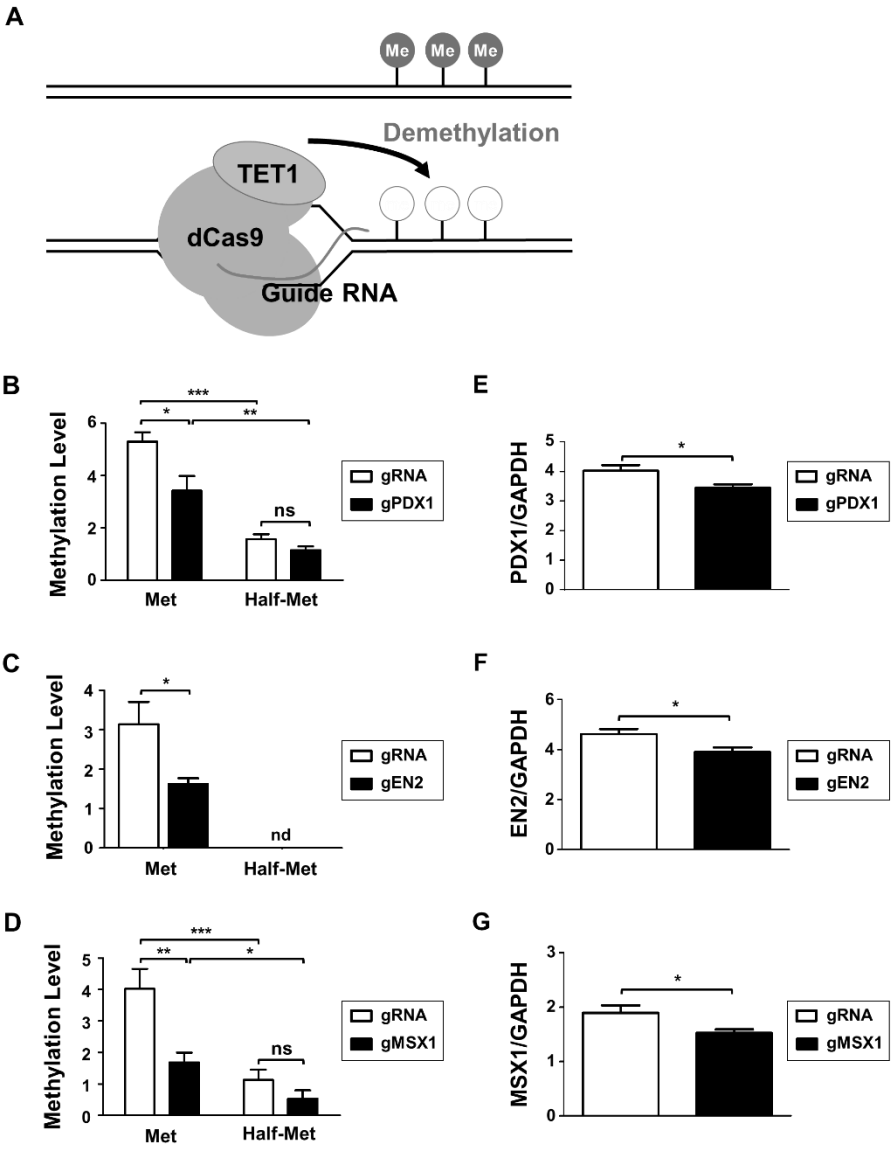

Figure 11. Customized MSP primers detect methylation changes in SW480 candidate biomarkers modulated by the CRISPR/dCas9-gRNA system. (A) A representation of my designed CRISPR/dCas9-gRNA system whereby specific gRNAs recruit the dCas9 protein and the catalytic domain of TET1 to demethylate the targeted genomic locus. **(B), (D), (F)** qMSP with SW480 cells transfected with dCas9-TET1CD mock or gRNA specific

to **(B)** *PDX1*, **(D)** *EN2*, and **(F)** *MSX1* indicates that the designed primers can distinguish the lack of methylation modulated by the CRISPR/ dCas9-gRNA system compared with controls. **(C), (E), (G)** qPCR with SW480 cells transfected with dCas9-TET1CD mock or gRNA of **(C)** *PDX1*, **(E)** *EN2*, and **(G)** *MSX1* shows a reduction in gene expression with decreased methylation. Genomic DNA and RNA used in qMSP and qPCR were simultaneously extracted from the cell lines.

7. The methylation levels of *PDX1*, *EN2*, and *MSX1* predict CRC metastasis

Next, I examined whether the methylation levels of the intragenic CpG regions of *PDX1*, *EN2*, and *MSX1* have clinical implications. I classified patients based on the methylation levels of these regions by conducting hierarchical clustering with the Manhattan distance. Consequently, I created two groups: the hypermethylated group (Group 1, N = 26) and the intermediate methylation and hypomethylated group (Group 2, n = 61) **(Figure 11A)**. Interestingly, these two groups showed a substantial difference in OS **(Figure 11B)** and PFS rates **(Figure 11C)**. In addition, peripheral lymphatic, vascular and perineural invasions, which are characteristic events followed by cancer metastasis, occurred more frequently in Group 1 than in Group 2. However, differences in cell differentiation, microsatellite instability, and tumor location were not observed. When I reviewed the information of patients, I realized that the majority of stage IV (after metastasis) patients were included in Group 1, whereas the majority of stage III (before metastasis) patients were included in Group 2 **(Table 2)**. These results suggest that *PDX1*, *EN2*, and *MSX1* methylation levels can predict CRC patient prognosis.

Finally, I examined whether my MSP system could distinguish between these two patient groups. I executed qMSP using bisulfite-treated genomic DNA from the tumor tissues of seven patients. Two patients in Group 1 showed higher methylation levels in the intragenic regions of *PDX1*, *EN2*, and *MSX1* than five individual patients in Group 2 **(Figure 11D)**. This result suggests that my MSP detection system can be clinically applied to predict the prognosis and metastasis of CRC patients after surgery.

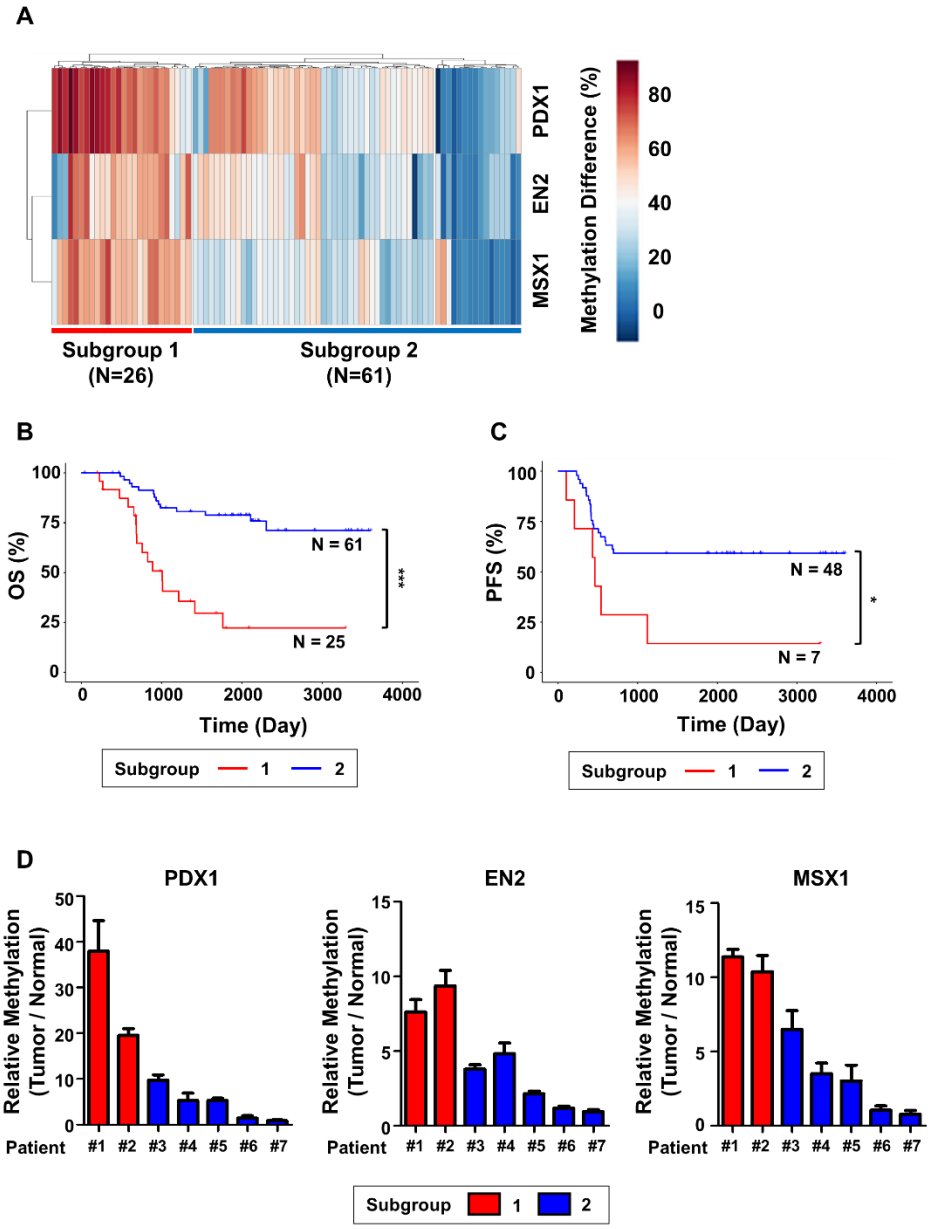

Figure 12. Prognostic potential of the 3-gene methylation signature is indicated through the classification of CRC patients. (A) Hierarchical clustering was conducted with DNA methylation data of intragenic CpG islands of *PDX1*, *EN2*, and *MSX1*, where

two distinct subgroups of CRC patients were observed. Kaplan–Meier plots for analyzing the significant differences in **(B)** overall survival and **(C)** CRC recurrence between the subgroups reveal the prognostic potential of the methylation data of the three biomarkers. The log-rank test was used to compare the significant differences between the two subgroups. One sample was excluded from the analysis of clinical data due to missing clinical data. Additionally, 31 patients were excluded from the recurrence analysis because they were diagnosed with stage IV CRC with metastatic cancers, and differentiating cancer recurrence would be challenging. **(D)** qMSP data generated with genomic DNA originating from the tumor and healthy tissues of the seven CRC patients displayed similar patterns to the cohort-specific methylation change analysis in a. The relative methylation levels of intragenic CpG islands of *PDX1*, *EN2*, and *MSX1* were calculated by dividing the methylation level of the tumor by that of healthy tissue.

Table 2. Clinical data of the subgroups classified by the methylation level of the intragenic CpG island of *PDX1, EN2,* and *MSX1*

Parameter	Subgroup 1	Subgroup 2	P
N	25	61	
Age (year)	58.2 (40-74)	63.2 (36-83)	0.0343*
Sex (male:female)	13:12	39:22	0.304, ns
Stage			
I	0% (0)	1.64% (1)	2.113E-06***
II	8% (2)	0% (0)	
III	20% (5)	78.7% (48)	
IV	72% (18)	26.2% (12)	
Invasion			
Lymphatic	56% (14)	45.9% (19)	0.0314*
Vascular	44% (11)	19.6% (8)	0.00172**
Perineural	80% (20)	50.8% (31)	0.0124*
Differentiation			
Well	0% (0)	1.7% (1)	0.706, ns
Moderate	91.7% (22)	93.1% (54)	
Poor	8.3% (2)	5.2% (3)	
Microsatellite			
Stable	91.3% (21)	93.1% (54)	0.969, ns
Instable - low	4.3% (1)	5.2% (3)	
Instable - high	4.3% (1)	5.2% (3)	
Site of Tumor			
Ascending	20% (5)	25.9% (15)	0.667, ns
Descending	4% (1)	0% (0)	
Transverse	4% (1)	1.7% (1)	
Sigmoid	40% (10)	36.2% (21)	
Rectal	16% (4)	20.7% (12)	
Rectosigmoid Junction	16% (4)	15.5% (9)	

***p<0.05**

****p<0.01**

*****p<0.001**

8. Differentially methylated regions of thyroid cancer clearly divide the cohort into two major subgroups

Following the methodology used in my colorectal cancer research, I conducted targeted bisulfite sequencing in my prospective thyroid cancer cohort to examine the DNA methylation status. By employing the same approach as in the colorectal cancer study, I identified differentially methylated regions (**Figure 3 and Figure 4**). As a result, I detected 248 hypermethylated regions and 83 hypomethylated regions within the cohort (**Figure 13A and Figure 13B**). Utilizing the DNA methylation data from these regions for comparative analysis between own cohort and the TCGA cohort, both independent cohorts evidently bifurcated into two major subgroups. The methylation patterns in own cohort's Tumor 1 (T1) and Tumor 2 (T2) subgroups closely mirrored those of PTC1 and PTC2 subgroups of TCGA-THCA, respectively (**Figure 13C**). To assess the clinical characteristics of PTC1 and PTC2, I generated a Kaplan-Meier plot to compare the survival rates of each subgroup. Notably, patients in the PTC1 group exhibited a significantly lower overall survival than those in the PTC2 group (**Figure 13D**). Furthermore, PCA analysis using the DMRs conclusively differentiated between T1 and T2 (**Figure 13E**). In conclusion, by defining differentially methylated regions from the thyroid cancer cohort, I was able to delineate two distinct subgroups with differing prognostic outcomes.

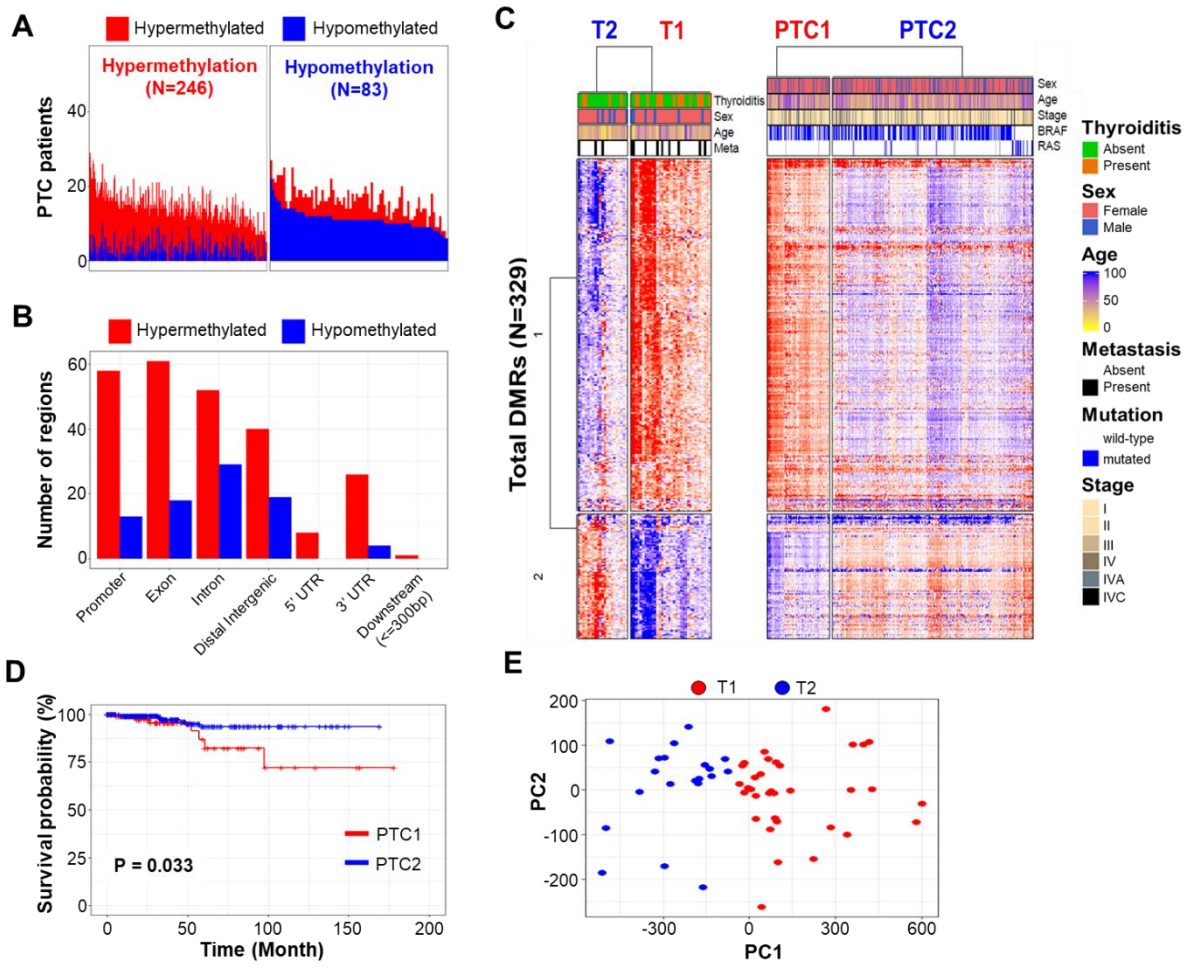

Figure 13. Analysis of differentially methylated regions in thyroid cancer cohorts. (A) The number of DMRs and the frequency of differentially methylated patients. **(B)** Regional distribution of selected hyper/hypomethylated DMRs. **(C)** Heatmap representation of DNA methylation status for DMRs in both own cohort (left) and TCGA THCA (right). **(D)** Kaplan-Meier survival curve analysis of PTC1 and PTC2 from TCGA cohort. **(E)** PCA analysis using DMRs.

9. Determination of thyroid cancer subgroups through DNA methylation data reveals clear molecular characteristics

Leveraging the differential methylation regions (DMRs) derived from targeted bisulfite sequencing data in own cohort, I undertook a transcriptomic assessment of patient subgroups within the TCGA dataset. Principal component analysis (PCA) unveiled a distinct segregation in the transcriptomic profiles of these subgroups (**Figure 14A**). To delineate the differential gene expression landscapes between these clusters, I employed DESeq2, pinpointing several differentially expressed genes (DEGs) (**Figure 14B**). A subsequent heatmap visualization of these DEGs segregated them into three primary gene clusters, each underpinned by unique expression patterns (**Figure 14C**). In my endeavor to decode the functional attributes of these clusters, a gene ontology analysis was executed (**Figure 14D**). Notably, gene cluster 1, characterized by diminished expression in both PTC1 and PTC2, was enriched with genes historically documented to exhibit suppressed expression in thyroid cancer. In contrast, gene cluster 3, which displayed augmented expression in both PTC1 and PTC2, was enriched with genes previously reported to have elevated expression in thyroid cancer[65]. Gene cluster 2 presented a unique scenario: heightened expression exclusively in PTC1 was accompanied by a preponderance of immune-related gene ontology (GO) terms, aligning with the well-established nexus between cancer progression and immune response. To identify genes among the differentially expressed genes (DEGs) that are regulated by DNA methylation, I annotated the DMRs and conducted a comparative analysis. As a result, I was able to identify 77 DEGs associated with the DMRs (**Figure 15**). In summary, it is unequivocal that the subgroups stratified on the basis of DNA methylation imprints manifest divergent transcriptomic hallmarks.

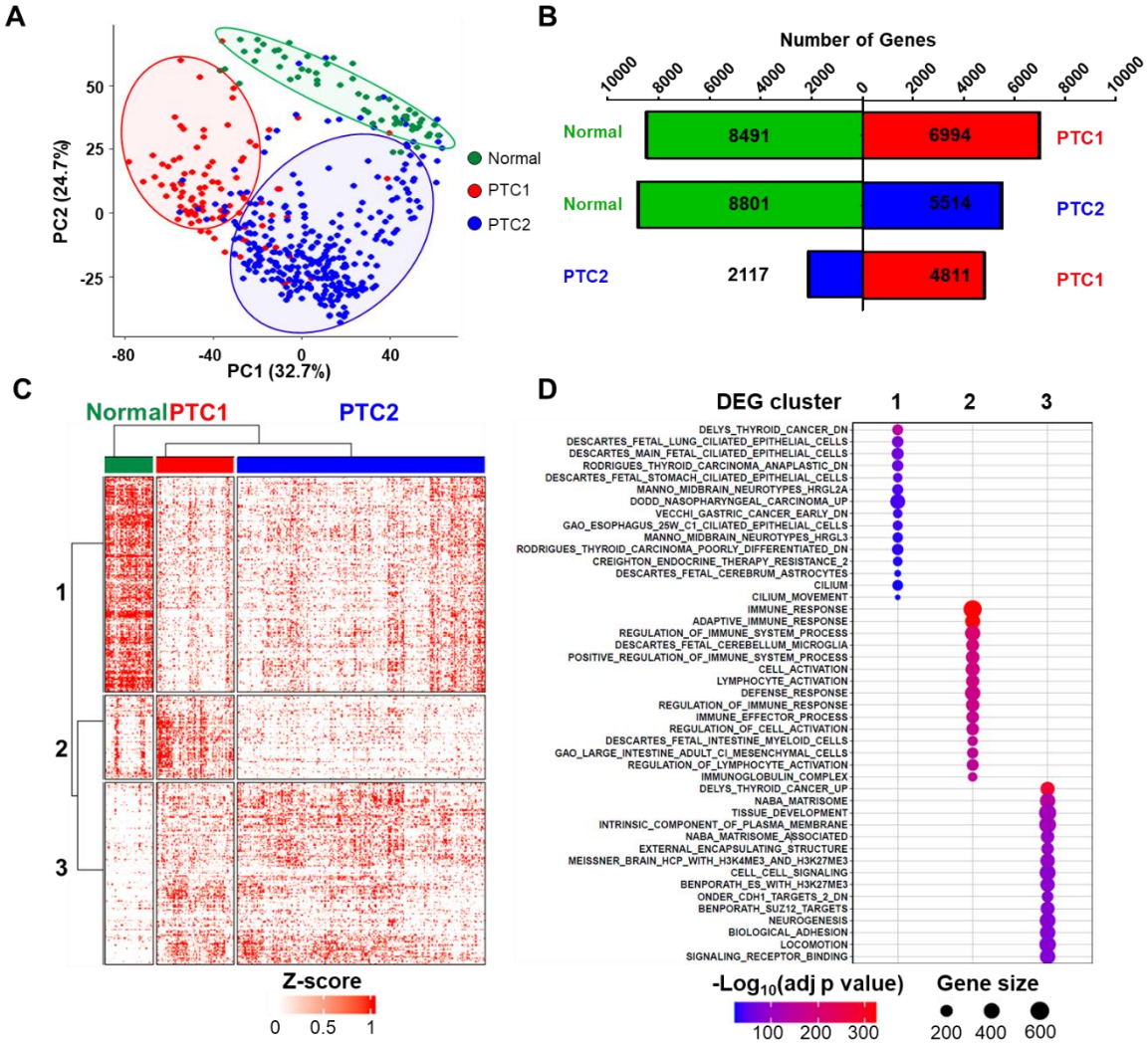

Figure 14. Analysis of transcriptomics between PTC1 and PTC2. (A) PCA analysis using RNA sequencing data of TCGA THCA. **(B)** The number of differentially expressed genes between each subgroup. **(C)** The heatmap of total differentially expressed genes **(D)** Gene ontology analysis of the DEG clusters.

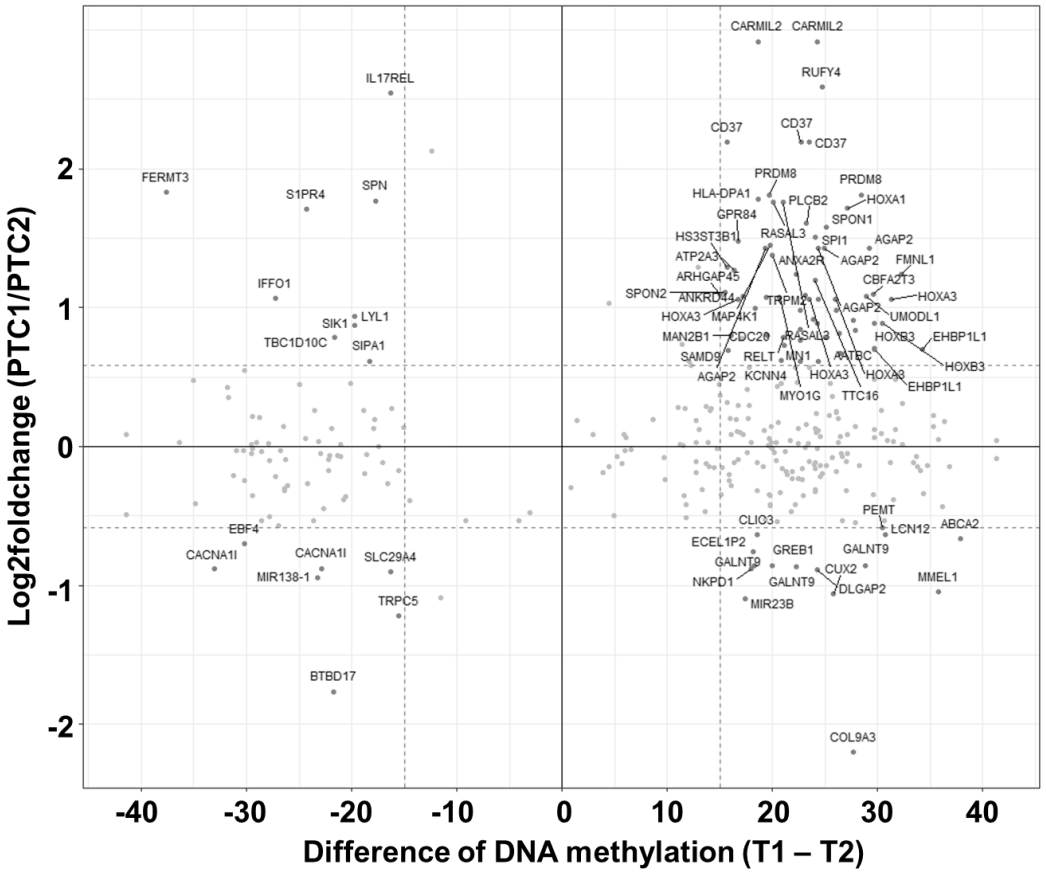

Figure 15. Overlap of differentially methylated regions of own cohort and differentially expressed genes of TCGA THCA cohort.

10. Analysis of chromatin accessibility confirmed that PTC1 is related to immune response

I sought to investigate the chromatin accessibility in thyroid cancer utilizing previously conducted RNA sequencing and ATAC sequencing data from the thyroid cancer cohort (GSE162515)[44]. Building upon my earlier findings of 77 genes that were both differentially methylated and expressed, I stratified the thyroid cancer patients in the GSE162515 cohort based on these genes. Notably, I observed a clear distinction in the thyroid cancer patients based on their gene expression patterns (Figure 16A and Figure 16B). I then identified differentially accessible regions between PTC1 and PTC2 within the GSE162515 dataset (Figure 16C). Mirroring the results from the RNA sequencing data analysis, genes associated with regions exhibiting increased accessibility in PTC1 were enriched for those related to immune response (Figure 16D). This led me to infer the occurrence of aberrant immune responses in PTC1 based on ATAC-seq data. Concurrently, I pinpointed genes from the previously identified set of 77, which also exhibited changes in chromatin accessibility (Table 3). These genes are postulated to play a pivotal role in the onset or progression of thyroid cancer.

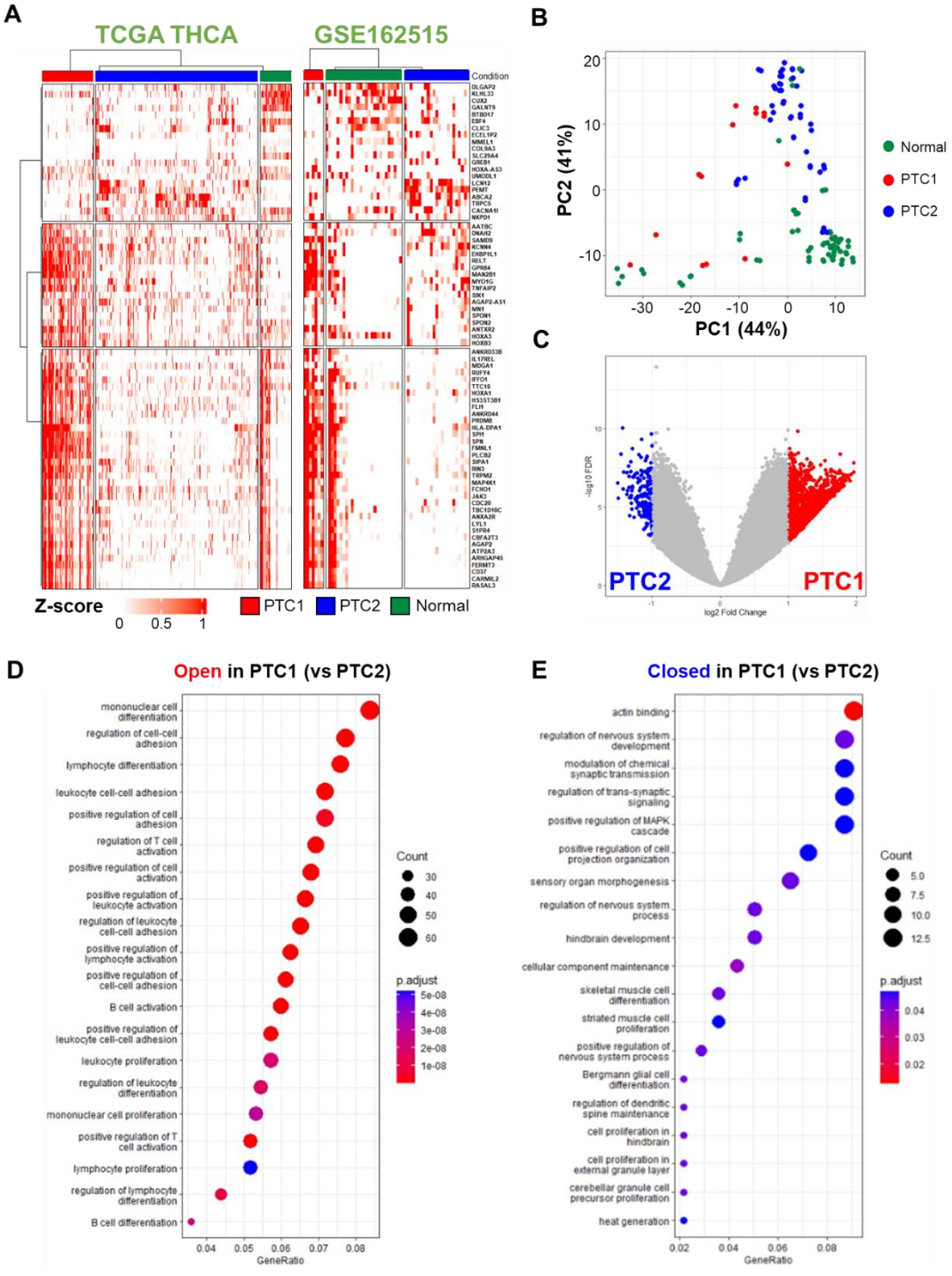

Figure 16. Analysis of ATAC sequencing of thyroid cancer cohort. (A) Comparison of thyroid cancer cohort with differentially methylated and expressed genes in TCGA THCA cohort. **(B)** PCA analysis with differentially accessible regions. **(C)** Volcano plot of differentially accessible regions between PTC1 and PTC2. Gene ontology analysis of **(D)** open regions and **(E)** closed regions in PTC1 versus PTC2.

Table 3. List of genes and genomic location where differentially methylated, accessible, and expressed between PTC1 and PTC2.

Location	Gene	*Met	**Exp	padj	***Acc	FDR
chr9:139901954-139902205	ABCA2	37.9	-0.67	2.69754E-21	-0.34	0.00523
chr11:65352231-65353134	EHBP1L1	34.3	0.70	4.40782E-20	0.44	0.01733
chr7:27163819-27164098	HOXA3	31.3	1.06	1.07583E-16	0.60	0.03001
chr17:46627787-46628444	HOXB3	30.5	0.89	1.00639E-15	0.61	0.01996
chr17:17465135-17465397	PEMT	30.5	-0.59	1.44107E-17	-0.38	0.00011
chr7:27179062-27179511	HOXA-AS3	29.7	0.71	0.034785118	0.99	0.0033
chr12:58132478-58132734	AGAP2	29.2	1.43	1.38311E-36	1.60	2E-05
chr4:81109887-81110460	PRDM8	28.4	1.81	1.81784E-47	0.48	0.03454
chr10:135038085-135038506	UTF1	27.9	0.84	0.002014953	0.43	0.00167
chr5:10649367-10650352	ANKRD33B	27.7	0.91	3.05992E-12	0.39	0.0477
chr7:27134097-27134303	HOXA1	27.1	1.71	2.18713E-42	0.65	0.03744
chr14:93153278-93154759	RIN3	26.4	0.82	2.59277E-28	0.39	0.00102
chr7:27153187-27153647	HOXA3	26.0	1.06	1.07583E-16	0.55	0.01914
chr11:14280741-14281164	SPON1	25.2	1.58	2.24875E-14	0.43	0.03614
chr17:46697413-46697701	HOXB8	25.1	0.78	0.005565291	1.28	0.0006
chr12:58130870-58132047	AGAP2	24.9	1.43	1.38311E-36	0.58	0.00136
chr12:58119909-58121551	AGAP2-AS1	24.4	0.61	6.49569E-07	0.36	0.00274
chr12:58119909-58121551	AGAP2	24.4	1.43	1.38311E-36	0.36	0.00274
chr17:7690145-7690411	DNAH2	23.9	0.91	3.87177E-10	-0.41	0.03224
chr7:27147589-27148389	HOXA3	23.5	1.06	1.07583E-16	0.48	0.02702
chr19:49841187-49841628	CD37	22.7	2.19	5.2369E-57	0.98	0.00022
chr5:43037259-43037520	ANXA2R	22.3	1.24	1.07086E-27	1.99	1.3E-06
chr2:11774310-11774521	GREB1	22.2	-0.87	4.6938E-05	0.72	0.00239
chr19:15563869-15564223	RASAL3	21.0	1.76	5.65964E-57	0.32	0.02826
chr19:44278273-44278777	KCNN4	20.8	0.62	0.002803386	0.42	0.01669
chr7:45002111-45002845	MYO1G	20.7	1.08	1.23836E-08	0.64	0.00161
chr19:15568027-15569227	RASAL3	20.1	1.76	5.65964E-57	0.87	1.7E-05
chr21:45789090-45789373	TRPM2	20.0	1.38	1.9613E-36	0.63	0.01813
chr1:43814305-43815277	CDC20	19.5	0.80	1.5117E-10	0.23	0.00354
chr16:67681975-67683924	CARMIL2	18.7	2.92	1.66885E-77	0.76	0.00029
chr12:54764065-54764510	GPR84	16.7	1.48	8.19556E-33	0.30	0.01611
chr17:3847999-3848570	ATP2A3	16.4	1.27	6.77505E-32	1.47	3.1E-07
chr17:14201726-14202052	HS3ST3B1	15.7	1.29	8.72421E-28	1.88	5.5E-06
chr19:49842654-49843628	CD37	15.7	2.19	5.2369E-57	0.64	0.00028
chr4:1205817-1206203	SPON2	15.5	1.11	7.99542E-20	0.70	0.00348
chr19:1070985-1071812	ARHGAP45	15.1	1.09	5.43616E-33	0.52	0.00751
chr7:5336513-5336894	SLC29A4	-16.3	-0.90	5.195E-06	2.48	7.5E-07
chr16:29675845-29676120	SPN	-17.7	1.77	2.52324E-44	2.61	4.7E-08
chr11:65408344-65408631	SIPA1	-18.3	0.62	3.83801E-14	2.42	2.5E-08
chr19:13207375-13207621	LYL1	-19.7	0.94	1.00055E-16	1.84	2.6E-06
chr21:44818894-44819446	SIK1	-19.7	0.88	0.00392773	3.24	1.7E-11
chr11:67176945-67177169	TBC1D10C	-21.6	0.79	9.22E-15	2.71	8.1E-09
chr17:72347924-72348322	BTBD17	-21.7	-1.76	1.11635E-05	0.70	0.01167
chr22:40057941-40058844	CACNA1I	-22.8	-0.88	2.93923E-09	1.25	0.00046
chr19:3178741-3179986	S1PR4	-24.2	1.71	1.96522E-45	2.61	1.7E-10
chr12:6664425-6665336	IFFO1	-27.2	1.07	3.00958E-37	2.66	1.2E-13
chr22:40081519-40082390	CACNA1I	-33.0	-0.88	2.93923E-09	0.85	0.00166
chr11:63974829-63975048	FERMT3	-37.6	1.83	2.41594E-54	2.47	6.4E-09

***Met: methylation difference between T1 and T2 (T1 - T2)**

****Exp: differential expression between PTC1 and PTC2 in TCGA THCA**

*****Acc: difference of chromatin accessibility between PTC1 and PTC2 in GSE165212**

IV. DISCUSSION

In this study, I present my discovery of novel CRC prognostic markers based on a comprehensive analysis of multiomics data and the validation of their functional impact in vitro. First, I used a public database for the preliminary screening of CRC-specific differentially methylated regions. In addition, I generated deep depth of targeted bisulfite sequencing data from the South Korean CRC cohort. For functional validation, I analyzed RNA-seq data and generated CRISPR/dCas-based cell lines. Finally, I established qMSP-based primer sequences and protocols for the quick and easy prediction of CRC prognosis.

I aimed to identify intragenic CGIs in which methylation changes were significantly related to gene expression and further cancer progression. By examining the differences in the methylation levels observed in tumors and adjacent healthy tissues via hybrid capture-based targeted bisulfite sequencing, I discovered significantly hypermethylated intragenic CGI regions in *PDX1*, *EN2*, and *MSX1* in the tumor samples. Therefore, I selected genomic locations targeted by MSP and designed primers to validate the hypermethylated status of the target CGIs. My primer design system for the candidate methylation biomarkers provided the strength that enabled the effective detection of methylation changes. In other words, since the targeted bisulfite sequencing data showed the methylation level of almost all CpG sites in certain genomic regions of interest, I could select the optimal MSP target sites efficiently, where the differences in methylation levels between healthy and tumor tissues were significant **(Figure 9A–C and Figure 10A-C)**. Hence, I successfully identified tumor-specific differentially methylated CGIs as prognostic markers of CRC and developed optimized qMSP methods to detect these methylation markers effectively.

Despite extensive efforts to discover CRC prognostic markers, technical drawbacks have challenged many researchers in developing systems for the clinical application of these biomarkers. One of the most important reasons is the difficulty in optimizing the qMSP. Specifically, the methylation level is difficult to quantify when discriminating between bisulfite-treated cytosine (methylated C) and uracil (unmethylated C) simultaneously. Increasing primer sensitivity while removing nonspecific bands is the key hurdle for optimizing qMSP. Based on high-coverage targeted bisulfite sequencing data, I

identified well-performing primer sets that included six or seven CpG sites in the forward and reverse primers that significantly distinguished healthy tissues from tumor tissues, although these primer sets were not tested in a multiplexing mode of action. I assume that more CpG sites can increase the annealing temperature, which could be more effective in precisely binding primers to their target sites. The qMSP technique established in this study may be used in additional and more feasible clinical applications for prognosis prediction if it is further developed and optimized as a multiplex qMSP technique. After inspecting the DNA methylation levels of the genes of interest, I then investigated the correlation between epigenetic regulation and the subsequent gene expression changes that ultimately lead to DNA methylation. However, even if there are significant epigenetic changes, one cannot conclude that these changes are correlated with gene expression. For example, I found two CpG islands of the *HOXA3* gene as the top 1 (chr7:27,147,589-27,148,389; hereafter *HOXA3*_CGI 7) and 2 (chr7:27,146,069-27,146,600; hereafter *HOXA3*_CGI 6) candidates that satisfied my criteria, but I failed to determine whether the expression of *HOXA3* was significantly changed in CRC patients **(Table 1).** I suppose that even if it is technically possible to detect the methylation changes of a particular gene of interest, it is still not a suitable epigenetic marker unless there is confidence in its expression effects. While it is well known that hypermethylation of promoter CpG islands leads to decreased gene expression, the mechanism and regulatory roles with respect to the gene expression of hypermethylated intragenic CGIs are still debated. One of the arguments supporting the idea of tumorigenesis caused by the hypermethylation of intragenic CGIs is that it leads to the hypermethylation of certain homeobox genes in their gene body[66]. This phenomenon was also confirmed in my study because *PDX1*, *EN2*, and *MSX1* are members of the homeobox family of genes. In addition to the *PDX1*, *EN2*, and *MSX1* CGIs, several CGI regions in other genes are worth examining. Many researchers have found methylated biomarkers in *BCAT1*, *NDRG4*, *SEPT9*, *BMP3*, and *IKZF1*[67-70], which correlates with my findings. Therefore, I provide evidence supporting the role of intragenic CGIs, which warrants further research.

In this study, I propose a practical method for identifying CRC prognostic markers. I utilized public databases and generated suitable high-depth targeted bisulfite sequencing data to define South Korean-specific differentially methylated regions (DMRs). I also

validated the proliferative aspect of the intragenic CGIs of *PDX1*, *EN2*, and *MSX1* in vitro, and I present optimized qMSP methods for further application in clinical fields. Based on the follow-up data of the patients in the cohort, I found a significant decrease in OS and higher recurrence rates in CRC patients with hypermethylated target CGIs. Along with surgical biopsy, adjuvant chemotherapy, and other proper care, regular tracking of prognostic factors could be helpful for patients with late-stage CRC. I also expect that my proposed methods and biomarkers could be applied to other cancers.

My study has made a significant stride in uncovering the epigenetic landscape of PTC by identifying 329 DMRs (Figure 13A). This discovery, achieved through TBS, unveils the considerable epigenetic heterogeneity within PTC (Figure 13C). Notably, the classification of PTC into two distinct subgroups based on their DMR profiles suggests the existence of divergent epigenetic mechanisms driving tumor development. This aligns with the current trajectory towards precision medicine in oncology, underscoring the necessity of personalized treatment modalities that are informed by molecular profiles rather than solely relying on traditional histopathological criteria.

My analyses involving transcriptomics and chromatin accessibility have shed light on the distinct characteristics of the PTC1 subgroup, particularly its association with immune-related pathways and a tumorigenic gene expression profile. These observations hint at a more aggressive molecular phenotype in PTC1, possibly linked to mechanisms of immune evasion or immunotolerance. Such insights are crucial for understanding the progression dynamics of PTC1 and its potential resistance to standard treatments.

In summary, my study marks a significant advancement in the understanding of PTC by identifying 329 DMRs and delineating two distinct PTC subgroups with unique epigenetic profiles. Notably, the PTC1 subgroup exhibits a more aggressive phenotype, potentially linked to immune-related pathways and tumorigenesis, as suggested by my analyses of chromatin accessibility and gene expression patterns. While my methodological approach using TBS and MSP provides a comprehensive view of PTC's epigenetic landscape, future research necessitates longitudinal studies for deeper insights into the clinical implications of these findings, particularly in advancing personalized treatment strategies in oncology.

V. CONCLUSION

In my study on CRC, I identified specific DNA methylation markers in the genes *PDX1*, *EN2*, and *MSX1* using targeted bisulfite sequencing. I discovered that these genes expression is linked with hypermethylation, highlighting their potential roles in CRC. Moreover, using a similar methodology and multiomics data, I was able to define two distinctly different subgroups within PTC. My findings suggest these genes as promising prognostic markers, offering a new direction for clinical diagnosis and treatment strategies.

REFERENCES

1. Ferlay J EM, Lam F, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, Bray F. Global Cancer Observatory: Cancer Today.: Lyon, France: International Agency for Research on Cancer.; 2020.
2. Day DW. The adenoma-carcinoma sequence. Scand J Gastroenterol Suppl 1984;104:99-107.
3. Dekker E, Tanis PJ, Vleugels JLA, Kasi PM, Wallace MB. Colorectal cancer. The Lancet 2019;394:1467-80.
4. Zecchin D, Boscaro V, Medico E, Barault L, Martini M, Arena S, et al. BRAF V600E is a determinant of sensitivity to proteasome inhibitors. Mol Cancer Ther 2013;12:2950-61.
5. Schell MJ, Yang M, Teer JK, Lo FY, Madan A, Coppola D, et al. A multigene mutation classification of 468 colorectal cancers reveals a prognostic role for APC. Nat Commun 2016;7:11743.
6. Xia LC, Van Hummelen P, Kubit M, Lee H, Bell JM, Grimes SM, et al. Whole genome analysis identifies the association of TP53 genomic deletions with lower survival in Stage III colorectal cancer. Sci Rep 2020;10:5009.
7. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nat Med 2004;10:789-99.
8. Dashwood RH. Early detection and prevention of colorectal cancer (review). Oncol Rep 1999;6:277-81.
9. Force USPST, Bibbins-Domingo K, Grossman DC, Curry SJ, Davidson KW, Epling JW, Jr., et al. Screening for Colorectal Cancer: US Preventive Services Task Force Recommendation Statement. JAMA 2016;315:2564-75.
10. Feinberg AP, Vogelstein B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature 1983;301:89-92.
11. Ehrlich M. DNA methylation in cancer: too much, but also too little. Oncogene 2002;21:5400-13.
12. Rodriguez J, Frigola J, Vendrell E, Risques RA, Fraga MF, Morales C, et al. Chromosomal instability correlates with genome-wide DNA demethylation in human primary colorectal cancers. Cancer Res 2006;66:8462-9468.
13. Toyota M, Ahuja N, Ohe-Toyota M, Herman JG, Baylin SB, Issa J-PJ. CpG island methylator phenotype in colorectal cancer. Proceedings of the National Academy of Sciences 1999;96:8681-6.
14. Tóth K, Sipos F, Kalmár A, Patai AV, Wichmann B, Stoehr R, et al. Detection of methylated SEPT9 in plasma is a reliable screening method for both left- and right-sided colon cancers. PLoS One 2012;7:e46000.
15. A stool DNA test (Cologuard) for colorectal cancer screening. Med Lett Drugs Ther 2014;56:100-1.

16. Peterse EFP, Meester RGS, de Jonge L, Omidvari A-H, Alarid-Escudero F, Knudsen AB, et al. Comparing the Cost-Effectiveness of Innovative Colorectal Cancer Screening Tests. JNCI: Journal of the National Cancer Institute 2021;113:154-61.
17. Koch A, Joosten SC, Feng Z, de Ruijter TC, Draht MX, Melotte V, et al. Analysis of DNA methylation in cancer: location revisited. Nat Rev Clin Oncol 2018;15:459-66.
18. Tse JWT, Jenkins LJ, Chionh F, Mariadason JM. Aberrant DNA Methylation in Colorectal Cancer: What Should We Target? Trends Cancer 2017;3:698-712.
19. Jain S, Chen S, Chang KC, Lin YJ, Hu CT, Boldbaatar B, et al. Impact of the location of CpG methylation within the GSTP1 gene on its specificity as a DNA marker for hepatocellular carcinoma. PLoS One 2012;7:e35789.
20. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the Infinium Methylation 450K technology. Epigenomics 2011;3:771-84.
21. Wendt J, Rosenbaum H, Richmond TA, Jeddeloh JA, Burgess DL. Targeted Bisulfite Sequencing Using the SeqCap Epi Enrichment System. In: Tost J, editor. DNA Methylation Protocols. New York, NY: Springer New York; 2018. p.383-405.
22. Herman JG, Graff JR, Myöhänen S, Nelkin BD, Baylin SB. Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. Proceedings of the National Academy of Sciences 1996;93:9821-6.
23. Hernández HG, Tse MY, Pang SC, Arboleda H, Forero DA. Optimizing methodologies for PCR-based DNA methylation analysis. BioTechniques 2013;55:181-97.
24. Kibbe WA. OligoCalc: an online oligonucleotide properties calculator. Nucleic Acids Res 2007;35:W43-6.
25. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 2021;71:209-49.
26. Chen DW, Lang BHH, McLeod DSA, Newbold K, Haymart MR. Thyroid cancer. Lancet 2023;401:1531-44.
27. Mazzoni. FLARCAT. Papillary Thyroid Carcinoma. in StatPearls. Treasure Island (FL): StatPearls 2023.
28. LiVolsi VA. Papillary thyroid carcinoma: an update. Mod Pathol 2011;24 Suppl 2:S1-9.
29. Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, et al. 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The

American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer. Thyroid 2016;26:1-133.
30. Abdullah MI, Junit SM, Ng KL, Jayapalan JJ, Karikalan B, Hashim OH. Papillary Thyroid Cancer: Genetic Alterations and Molecular Biomarker Investigations. International Journal of Medical Sciences 2019;16:450-60.
31. Zafon C, Gil J, Perez-Gonzalez B, Jorda M. DNA methylation in thyroid cancer. Endocr Relat Cancer 2019;26:R415-R39.
32. Saghafinia S, Mina M, Riggi N, Hanahan D, Ciriello G. Pan-Cancer Landscape of Aberrant DNA Methylation across Human Tumors. Cell Rep 2018;25:1066-80 e8.
33. Cha YJ, Koo JS. Next-generation sequencing in thyroid cancer. J Transl Med 2016;14:322.
34. Cancer Genome Atlas Research N. Integrated genomic characterization of papillary thyroid carcinoma. Cell 2014;159:676-90.
35. Beltrami CM, Dos Reis MB, Barros-Filho MC, Marchi FA, Kuasne H, Pinto CAL, et al. Integrated data analysis reveals potential drivers and pathways disrupted by DNA methylation in papillary thyroid carcinomas. Clin Epigenetics 2017;9:45.
36. Ye L, Zhou X, Huang F, Wang W, Qi Y, Xu H, et al. The genetic landscape of benign thyroid nodules revealed by whole exome and transcriptome sequencing. Nat Commun 2017;8:15533.
37. Yim JH, Choi AH, Li AX, Qin H, Chang S, Tong ST, et al. Identification of Tissue-Specific DNA Methylation Signatures for Thyroid Nodule Diagnostics. Clin Cancer Res 2019;25:544-51.
38. Rodriguez-Rodero S, Fernandez AF, Fernandez-Morera JL, Castro-Santos P, Bayon GF, Ferrero C, et al. DNA methylation signatures identify biologically distinct thyroid cancer subtypes. J Clin Endocrinol Metab 2013;98:2811-21.
39. Capdevila J, Awada A, Fuhrer-Sakel D, Leboulleux S, Pauwels P. Molecular diagnosis and targeted treatment of advanced follicular cell-derived thyroid cancer in the precision medicine era. Cancer Treat Rev 2022;106:102380.
40. Morselli M, Farrell C, Rubbi L, Fehling HL, Henkhaus R, Pellegrini M. Targeted bisulfite sequencing for biomarker discovery. Methods 2021;187:13-27.
41. Gulilat M, Lamb T, Teft WA, Wang J, Dron JS, Robinson JF, et al. Targeted next generation sequencing as a tool for precision medicine. BMC Med Genomics 2019;12:81.
42. Lee Y, Dho SH, Lee J, Hwang J-H, Kim M, Choi W-Y, et al. Hypermethylation of PDX1, EN2, and MSX1 predicts the prognosis of colorectal cancer. Experimental & Molecular Medicine 2022;54:156-68.
43. Sanghi A, Gruber JJ, Metwally A, Jiang L, Reynolds W, Sunwoo J, et al.

Chromatin accessibility associates with protein-RNA correlation in human cancer. Nat Commun 2021;12:5732.
44. Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, et al. The chromatin accessibility landscape of primary human cancers. Science 2018;362.
45. Klutstein M, Nejman D, Greenfield R, Cedar H. DNA Methylation in Cancer and Aging. Cancer Res 2016;76:3446-50.
46. Ng JM, Yu J. Promoter hypermethylation of tumour suppressor genes as potential biomarkers in colorectal cancer. Int J Mol Sci 2015;16:2472-96.
47. Lu J, Wilfred P, Korbie D, Trau M. Regulation of Canonical Oncogenic Signaling Pathways in Cancer via DNA Methylation. Cancers (Basel) 2020;12.
48. Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet 2008;9:465-76.
49. Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature 2010;466:253-7.
50. Lee SM, Lee J, Noh KM, Choi WY, Jeon S, Oh GT, et al. Intragenic CpG islands play important roles in bivalent chromatin assembly of developmental genes. Proc Natl Acad Sci U S A 2017;114:E1885-e94.
51. Krinner S, Heitzer AP, Diermeier SD, Obermeier I, Längst G, Wagner R. CpG domains downstream of TSSs promote high levels of gene expression. Nucleic Acids Res 2014;42:3551-64.
52. Shenker N, Flanagan JM. Intragenic DNA methylation: implications of this epigenetic mechanism for cancer research. Br J Cancer 2012;106:248-53.
53. Kinde B, Wu DY, Greenberg ME, Gabel HW. DNA methylation in the gene body influences MeCP2-mediated gene repression. Proc Natl Acad Sci U S A 2016;113:15114-9.
54. Arechederra M, Daian F, Yim A, Bazai SK, Richelme S, Dono R, et al. Hypermethylation of gene body CpG islands predicts high dosage of functional oncogenes in liver cancer. Nat Commun 2018;9:3164.
55. Greenberg MVC, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. Nat Rev Mol Cell Biol 2019;20:590-607.
56. Chandrashekar DS, Bashel B, Balasubramanya SAH, Creighton CJ, Ponce-Rodriguez I, Chakravarthi B, et al. UALCAN: A Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses. Neoplasia 2017;19:649-58.
57. Teo AK, Tsuneyoshi N, Hoon S, Tan EK, Stanton LW, Wright CV, et al. PDX1 binds and represses hepatic genes to ensure robust pancreatic commitment in differentiating human embryonic stem cells. Stem Cell Reports

2015;4:578-90.
58. Grillone K, Riillo C, Scionti F, Rocca R, Tradigo G, Guzzi PH, et al. Non-coding RNAs in cancer: platforms and strategies for investigating the genomic "dark matter". J Exp Clin Cancer Res 2020;39:117.
59. Boons G, Vandamme T, Ibrahim J, Roeyen G, Driessen A, Peeters D, et al. PDX1 DNA Methylation Distinguishes Two Subtypes of Pancreatic Neuroendocrine Neoplasms with a Different Prognosis. Cancers (Basel) 2020;12.
60. Vinogradova TV, Sverdlov ED. PDX1: A Unique Pancreatic Master Regulator Constantly Changes Its Functions during Embryonic Development and Progression of Pancreatic Cancer. Biochemistry (Mosc) 2017;82:887-93.
61. Brunet I, Weinl C, Piper M, Trembleau A, Volovitch M, Harris W, et al. The transcription factor Engrailed-2 guides retinal axons. Nature 2005;438:94-8.
62. Li Y, Liu J, Xiao Q, Tian R, Zhou Z, Gan Y, et al. EN2 as an oncogene promotes tumor progression via regulating CCL20 in colorectal cancer. Cell Death Dis 2020;11:604.
63. Sun AJ, Gao HB, Liu G, Ge HF, Ke ZP, Li S. Identification of MSX1 and DCLK1 as mRNA Biomarkers for Colorectal Cancer Detection Through DNA Methylation Information. J Cell Physiol 2017;232:1879-84.
64. Morita S, Noguchi H, Horii T, Nakabayashi K, Kimura M, Okamura K, et al. Targeted DNA demethylation in vivo using dCas9-peptide repeat and scFv-TET1 catalytic domain fusions. Nat Biotechnol 2016;34:1060-5.
65. Delys L, Detours V, Franc B, Thomas G, Bogdanova T, Tronko M, et al. Gene expression and the biological phenotype of papillary thyroid carcinomas. Oncogene 2007;26:7894-903.
66. Su J, Huang YH, Cui X, Wang X, Zhang X, Lei Y, et al. Homeobox oncogene activation by pan-cancer DNA hypermethylation. Genome Biol 2018;19:108.
67. Imperiale TF, Ransohoff DF, Itzkowitz SH, Levin TR, Lavin P, Lidgard GP, et al. Multitarget stool DNA testing for colorectal-cancer screening. N Engl J Med 2014;370:1287-97.
68. Lu H, Huang S, Zhang X, Wang D, Zhang X, Yuan X, et al. DNA methylation analysis of SFRP2, GATA4/5, NDRG4 and VIM for the detection of colorectal cancer in fecal DNA. Oncol Lett 2014;8:1751-6.
69. Liu Y, Tham CK, Ong SY, Ho KS, Lim JF, Chew MH, et al. Serum methylation levels of TAC1. SEPT9 and EYA4 as diagnostic markers for early colorectal cancers: a pilot study. Biomarkers 2013;18:399-405.
70. Pedersen SK, Symonds EL, Baker RT, Murray DH, McEvoy A, Van Doorn SC, et al. Evaluation of an assay for methylated BCAT1 and IKZF1 in plasma for detection of colorectal neoplasia. BMC Cancer 2015;15:654.

ABSTRACT(IN KOREAN)

특정 CpG 섬에서의 비정상적인 DNA 메틸화를 가진 암의 아형 식별: 대장암 및 갑상선 암 연구

<지도교수 김 락 균>

연세대학교 대학원 의과학과

이 영 운

DNA methylation의 변화와 암 진행의 관계에 대한 수많은 연구가 진행되어오고 있지만 실제 암의 진단 바이오마커로 검증된 유전자는 몇 개에 지나지 않는다. DNA methylation의 변화를 더 효과적으로 검출하기 위해 targeted bisulfite sequencing (TBS)을 수행했다. RNA-seq과의 통합 분석을 통해 *PDX1*, *EN2*, *MSX1* 내부의 CpG island가 대장암에서 DNA methylation의 유의미한 차이를 보이는 것을 확인했다. 이 유전자들이 암 유발 특성을 가지고 있으며 발현량이 DNA methylation과 양의 상관관계를 갖는 것을 확인했다. TBS의 높은 read depth 덕분에 관심 유전체 영역에서 단일 CpG 수준의 미세한 DNA methylation 차이를 감지할 수 있는 quantitative MSP primer를 제작할 수 있었다. 이 유전자들의 DNA methylation 수준을 통해 대장암 환자들을 좋은 예후와 나쁜 예후를 보이는 두 그룹으로 나눌 수 있었다. 이는 *PDX1*, *EN2*, *MSX1*의 DNA methylation 수준이 예후 예측의 바이오마커로써 유망한 성능을 보이고 대장암 환자들에 임상적으로 응용될 수 있음을 암시한다. 또한, TBS를 갑상선암에도 적용하여 차등적인 수준을 보이는 DNA methylation 바이오마커를 동정했다. 해당 바이오마커를 통해 자체 코호트와 TCGA 코호트에서 갑상선 유두암의 이질적인 그룹을 동정하였고, 다중 오믹스 데이터를 통합하여 이들의 분자적 특성을 분석하였다. 이 연구를 통해 우리는 임상적으로 중요한 유전체 영역을 찾기 위한 간소화된 작업 흐름을 제시한다.

핵심되는 말 : CpG island, DNA methylation, 대장암, 갑상선암, Targeted

bisulfite sequencing

PUBLICATION LIST

1. Lee, Y, Dho SH, Lee J, Hwang, J, et al. Hypermethylation of PDX1, EN2, and MSX1 predicts the prognosis of colorectal cancer. Experimental & Molecular Medicine. 2022;54(2):156-168