이학박사 학위논문

Discovering genome-wide methylation and fragmentomics of cell free DNA for diagnosis and prognosis of colorectal cancer

대장암 진단 및 예후 예측를 위한 혈액내 종양DNA의 genome-wide 메틸화 및 fragmentomics 마커 발굴에 관한 연구

2022년 2월

서울대학교 융합과학기술대학원
분자의학 및 바이오제약학과

강 준 규

대장암 진단 및 예후 예측를 위한 혈액내 종양DNA의 genome-wide 메틸화 및 fragmentomics 마커 발굴에 관한 연구

지도교수 김 태 유

이 논문을 이학박사 학위논문으로 제출함.

2022 년 1 월

서울대학교 융합과학기술대학원

분자의학 및 바이오제약학과

강 준 규

강 준 규 의 이학박사 학위논문을 인준함.

2022 년 1 월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

Discovering genome-wide methylation and fragmentomics of cell free DNA for diagnosis and prognosis of colorectal cancer

By Jun-Kyu Kang
(Directed by Tae-You Kim, M.D., Ph.D.)

A thesis submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Department of Molecular Medicine and Biopharmaceutical Science at World Class University Graduate School of Convergence Science and Technology, Seoul National University, Seoul, Korea

January 2022

Approved by Thesis Committee:

Professor _____Chairman

Professor _____Vice chairman

Professor _____

Professor _____

Professor _____

ABSTRACT

Discovering genome-wide methylation and fragmentomics of cell free DNA for diagnosis and prognosis of colorectal cancer

Jun-Kyu Kang

Department of Molecular Medicine
and Biopharmaceutical Sciences
World Class University Graduate School of
Convergence Science and Technology
Seoul National University

Non-genetic signatures from liquid biopsy samples are emerging as feasible markers of cancer because plasma cell-free DNA (cfDNA) is representative of the patient's systemic state. Non-genetic signatures include cfDNA methylation, topology of cfDNA, and cfDNA fragmentomics. DNA methylation has somatic tissue specific patterns, and DNA fragment size is one of the most representative characteristics of cfDNA. In particular, cfDNA from the plasma of cancer patients, which contains circulating tumor DNA (ctDNA), can be representative of the status of both the primary tumor and minimal residual disease. For this reason, the tissue of origin (TOO) could be determined from ctDNA methylation patterns. Fragment size of ctDNA could also be a useful marker for cancer patients. However, studies on the comprehensive applications of non-genetic signatures for cancer diagnosis, monitoring, and predicted prognosis are still needed to define and validate the role of non-genetic markers in clinical practice.

Here, I show 1) an accurate prediction model that was developed using a machine learning algorithm for the comprehensive analysis of multiple CpG sites. Although many DNA methylation markers have been reported, previously reported markers were based on a single marker and a western population. My prediction model includes 305 CpG sites and was built by a machine learning algorithm based on tissue samples from Korean colorectal cancer patients. The prediction model showed high performance not only in databases of pan-cancer tissue samples but also those based on plasma from cancer patients. In addition, the prognosis of colorectal cancer patients was accurately predicted with a subset of the 305 CpG sites.

Next, I showed that 2) the fragmentation ratio of specific lengths of DNA could be a valuable prognostic marker for colorectal cancer patients. Many recent studies have shown ctDNA fragment size is shorter than that of cfDNA derived from healthy tissue and have attempted to apply this to cancer diagnosis; however, the data are limited, and the only application has been for cancer diagnosis. In order to fill this gap, cfDNA fragment size was analyzed using targeted deep sequencing from paired ends. I demonstrated that ctDNA fragment length was related to variant allele frequency, and the prognosis of colorectal cancer patients could be predicted by the fragmentation ratio at a specific sampling time in longitudinal samples.

In summary, blood based non-genetic signatures are significantly associated with the status of colorectal cancer and can be used to predict patient prognosis.

Keyword : Colorectal cancer (CRC), Epigenetics, circulating tumor DNA (ctDNA), Diagnosis, Prognosis, Next generation sequencing (NGS)

Student Number : 2018-37966

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

I. Use of an optimized machine learning algorithm to discover DNA methylation markers from Korean colorectal cancer patients

II. Combined analysis of ctDNA mutation and fragment size for predicting prognosis of colorectal cancer

I. Use of an optimized machine learning algorithm to discover DNA methylation markers from Korean colorectal cancer patients

ABSTRACT

DNA methylation is a key epigenetic regulator in mammalian development. Furthermore, DNA methylation is well known to play an important role in carcinogenesis. Pattern of DNA methylation vary in somatic cancer tissues and among subjects of different races. Although numerous cancer-related DNA methylation markers have been reported, they were based on single marker studies performed in Western populations. In this study, we investigated discovering comprehensive markers and validating the potential as diagnosis and prognosis for colorectal cancer. Patients with various stage of colorectal cancer were eligible for the current study, mainly stage III. We generated genome-wide methylation data from 379 colorectal cancer (CRC) tissues and 330 available paired adjacent normal mucosa tissues from Korean patients by Illumina EPIC Human Methylation microarray targeting 860,000~ CpG sites. A machine learning algorithm was used to build an optimized prediction model and select the tumor specific markers based on through theses CpG sites. Then, the risk score was devised for prognosis using this marker set. Finally, in order to validate the rule of CpG sites, the genomic location and pathway enrichment analysis was performed by CHROMHMM and Metascape. A total of 305 methylation markers that showed statistically significant differences between normal and cancer tissues were selected. Our model could accurately identify CRC (areas under the curve for the training and validation cohorts: 0.968 and 0.984, respectively). Using our prediction model, the colorectal cancer patients were predicted as colorectal

cancer accurately in the methylation data from TCGA (COREAD; colorectal cancer tissue DNA) and GEO dataset (plasma cfDNA from colorectal cancer patients). The risk score comprising the subset of 305 methylation markers was calculated, and poor prognosis was predicted in the high-risk score group (overall survival P = 0.073, progression-free survival P = 0.0026). Gene ontology (GO) enrichment analysis showed that the 305 CpG sites were enriched in transcription regulatory regions (160/305, 52.5%) and were associated with developmental process and carcinogenesis (GO: 0032502, log10P = -4.28; C4721208, log10P = -3.80). In summary, the performance of our prediction model with these 305 CpG sites was highly accurate for CRC diagnosis, and the optimized risk score could predict the prognosis of Korean CRC patients.

Key words : Colorectal cancer (CRC), DNA methylation, Diagnosis, Prognosis, Machine learning, Risk score

Student Number : 2018-37966

INTRODUCTION

Colorectal cancer (CRC) is a leading cause of death worldwide, accounting for 9.0 age-standardized deaths per 100,000 people in 2020 (1). According to GLOBCAN 2018 data, the cumulative risk of CRC development in Korea was ranked in the top two (2). Regardless of sex, the incidence of the CRC in Korea has been increasing steadily (3). The incidence rate increases in an age-dependent manner (15–34 years, 3.6%; 35–64 years, 10.3%; ≥65 years, 13.4%). However, CRC is curable when detected early, with survival rates >89.2% at 5 years for patients diagnosed with stage I disease. In contrast, patients with regional spread (stage IIIC) have a worse prognosis, with approximately 43.2% surviving at 5 years. Therefore, early diagnosis of CRC is important. Conventional screening methods include colonoscopy or fecal occult blood testing. Although screening programs are heterogeneous worldwide, introduction of a screening program seems to be followed by reduced CRC mortality.

DNA methylation is well known to play an important role in carcinogenesis. For example, it was reported that either the cis-regulatory elements of tumor suppressor genes were hypermethylated or the cis-regulatory elements of oncogenes were hypomethylated in tumor cells compared to normal cells (4). These aberrantly methylated regions have been considered as diagnostic or prognostic markers for cancer. Luo et al. reported the use of DNA methylation markers for diagnosing, prognosing, and subtyping CRC based on machine learning, but the reproducibility of these

markers is unclear (5). DNA methylation has somatic cancer-specific patterns, which Liu et al. used to develop a prediction model for both cancer diagnosis and tissue of origin (TOO) (6). However, DNA methylation patterns are highly race dependent (7). A comprehensive analysis of DNA methylation using East Asian population data is still needed.

Machine learning and other computational resources are increasingly used to discover epigenetic markers (8, 9). Here, I describe the identification of epigenetic markers for cancer diagnosis using a large-scale Korean CRC patient dataset. A comprehensive analysis based on machine learning was carried out to identify genome-wide methylation patterns from databases, and the utility of diagnostic markers and prognostic risk score was determined with statistical methods.

EXPERIMENTAL DESIGN

1. Extraction genomic DNA from Colorectal tissue

Genomic DNA (gDNA) of 379 tumor tissue and 330 adjacent normal colon tissues were extracted by following kit. Genomic DNA was isolated from each sample using a Qiagen DNA FFPE Tissue Kit (Qiagen, Hilden, Germany) for FFPE samples and a QIAamp DNA Mini Kit (Qiagen) for fresh-frozen tissues. After isolation, the concentrations and purities of genomic DNA were measured using a spectrophotometer (ND1000; Nanodrop Technologies, Thermo Fisher Scientific, MA, USA).

2. Illumina Infinium MethylationEPIC array BeadChip (850K) and Whole transcriptome sequencing

Genome-wide methylation data was generated by Infinium Methylation EPIC array (850K array). The signals were normalized by 'SWAN' method in R package 'minfi' (10). Through this, the **β** value of ~850,000 CpG sites were calculated as representing the level of DNA methylation. Whole transcriptome sequencing was performed and the quantification by gene symbol and gene iso-form were calculated.

3. Monte-carlo simulation

Theoretically, 6,000 dGE(diploid genome equivalent; ~40ng) can be isolated from 4ml of plasma which is 40% of the 10ml whole blood (11). Detection rate 0.01 % means the platform detect 1 anomalous signal out of

12,000 copies. In order to simulating this theoretic situation, I performed monte-carlo simulation. As Diploid genomes, 15,000 is the expected copy number in 100 ng cfDNA. I created 15,150 simulated genomes with from 1 to 10,000 independent loci with from 0.01 to 10% cancer-specific DMRs in tenfold increments. Next, I performed simulation with depending on the number of epi-mutations to detect. The process was repeated 1,000 times for each combination of parameters.

4. Building the prediction model and selecting significant CpG sites

Setting cohort for machine learning: Using R package 'caret', colorectal cohort was divided into training cohort and validation cohort (8:2) keeping the ratio (0.87:1) which is the ratio of sample size between adjacent normal tissue and colorectal cancer tissue. For cross validation, the training cohort was divided into sub-training set and test set (8:2) keeping the ratio (0.87:1), additionally. **Selection markers for machine learning:** With sub-training set, differentially methylated regions (DMRs) and each p value between tumor tissue and normal tissue were calculated by student t-test. After that, Benjamini-Hochberg correction were performed for calculating FDR. Out of 850,000 probes, TOP 1,000 probes (500 DMRs for hyper-methylated in tumor & 500 DMRs for hypo-methylated in tumor) were selected by the criteria which is FDR under 0.05 and $\Delta\beta$ value (mean of β tumor tissues - mean of β normal tissues). This step was repeated 5 times for inner cross-validation. **Selection best machine learning model:** With TOP 1,000 probes, the performance was compared among 5 suitable classification

model (Linear Discriminant Analysis; LDA, Decision Tree; CART, K Nearest Neighbours; KNN, Random Forest; RF, Support Vector Machines; SVM). Machine learning was performed by inner cross validation (inner CV; n=10). Accuracy and Kappa value were calculated in every repeat. Through this, TOP 2 accurate models were selected. **Selection probe set:** With TOP 2 models, importance score for each probe was calculated. In order to select the probe set, inflection point analysis was performed based on importance score sorted in descending order. Then, the set of probes intersecting between the TOP 2 models was selected. **Building prediction model & Calculating performance of prediction:** the prediction model was built by the set of probes and TOP 1 classification model among 5 models. Area under curve (AUC) with both sub-training set and test set was calculated by R package 'pROC'. **Outer cross validation:** The methods up to this point were repeated a total of 5 times. **Building final prediction model with the final set of probes:** the final classification model which is called 3 more times out of 5 outer CV was selected. With 305 final probes, the prediction model for cancer diagnosis was built. Final validation & visualization: With validation cohort, the AUC was calculated. And heatmap with 305 probes set were visualized by R package 'pheatmap'.

5. Processing 450k methylome data of TCGA, EWAS, GEO database

Using FireBrowse database (http://firebrowse.org/), methylome data (Infinium HumanMethylation450 BeadChip array) of pan-cancer level tissues was collected (tumor and available matched normal tissue from 33 types of

cancer) (12). From EWAS database, methylome data (Infinium HumanMethylation450 BeadChip array) of 31 type of somatic tissues was collected (13). In GEO database, large cohort of normal PBMC methylome data (GSE40279; Infinium HumanMethylation450 BeadChip) and methylome data of plasma cfDNA from various disease included colorectal cancer (GSE122126; Infinium HumanMethylation450 BeadChip) were collected. Either 'NA' or probes in sex chromosome were excluded for analysis.

6. Deconvolution and clustering

With R package 'Rtsne', the deconvolution of large-scale database was performed (dims = 2, max_iter = 500, perplexity = 5). Using prcomp method in R package 'ggfortify', PCA analysis was performed.

7. Permutation test

With TCGA and EWAS data set (n=15,646), the intersected probes were collected. Then, random forest algorithm applied to predict the tissue of origin (ntree = 500). This step repeated 1,000 times and error rate were calculated for each number of sampled probes.

8. ChromHMM status and GO analysis

Using bedtools (v.2.28.0), the 305 probes were annotated by ChromHMM status (14). In addition, the coordinate of 305 probe were annotated by Homer 'annotatepeak.pl' (15) . Then, GO analysis was performed with probes which were located in either transcription starting stie

(TSS) or gene promoter (16) .

9. Risk score

Using coxph method in R package 'survival', cox proportional-Harzards model analysis was performed with each probe. The prognostic markers which were statistically fulfilled for overall survival (OS) and progression free survival (PFS) were selected (Log-rank test; p-value < 0.05). 20 and 133 probes were selected as prognostic markers for OS and PFS, each. The risk score was determined based on coefficients for each probe from cox regression analysis. The formula is as follows:

$$Risk\ score = \sum_{i}^{n} \beta_i * x_i$$

10. Survival analysis

Using R package 'survminer', the survival rate for OS and PFS were calculated. And Kaplan-Meier plot were visualized and p value were calculated by Log-rank test.

RESULTS

1. Clinicopathological Information of the COPM Dataset

All patients provided written informed consent before any study-specific procedures. The protocol of this study was reviewed and approved by the Institutional Review Board (IRB) of SNUH (IRB number: 1708-031-875) and was conducted in accordance with the Declaration of Helsinki in biomedical research involving human subjects. Clinicopathological information was collected from 367 of 379 patients (Table 1). The median age of the cohort was 62 (23–88), and the percentages of males and females were 60.9% and 35.9%, respectively. The most common disease stage was III (61.2%), followed by II (19.5%), IV (11.9%), and I (4.2%). The anatomical sites of the primary tumor were right colon (25.5%), left colon (68.9%), and other (3.2%). The pathological subtypes were adenocarcinoma (91.3%), mucinous adenocarcinoma (4.2%), and other (1.3%). Microsatellite instability status was high (7.1%), low (4.7%), and stable (83.1%). Genome-wide methylation data and processed methylation data were generated for each patient.

Categories	(n = 379)	Number of patients (%)
Age at diagnosis, median (range)		62 (23 – 88)
Sex	Male	231 (60.9 %)
	Female	136 (35.9 %)
Stage	I	16 (4.2%)
	II	74 (19.5%)
	III	232 (61.2%)
	IV	45 (11.9%)
Primary tumor site	Right colon	95 (25.15%)
	Left colon	260 (68.9%)
	Other	12 (3.2%)
Pathology	Adenocarcinoma	346 (91.3%)
	Mucinous adenocarcinoma	16 (4.2%)
	Other	5 (1.3%)
Microsatellite instability	MSS	315 (83.1%)
	MSI-L	18 (4.7%)
	MSI-H	27 (7.1%)
	Not available	7 (1.8%)

not available patients; n = 12

Table 1. Clinicopathological information of the COPM cohort.

2. Prediction Model Performance Based on Machine Learning

Through Monte Carlo simulation, screening more than eight epi-mutations could detect cancer containing 0.1% tumor content with an accuracy of 100% (Figure 1). To precisely differentiate between tumor and normal tissue, more than eight markers needed to be comprehensively analyzed (Figure 2). DNA methylome data was used to determine which machine learning algorithm was best fitted. To avoid overfitting, 10-fold inner CV and 5-fold outer CV was performed. This generated the prediction model with the KNN algorithm and 305 cancer-specific markers. Using PCA, PC1 could separate tumor and normal tissues with all probes (17.8%) or 305 markers (85.31%), but the samples were more distinguishable using calculation with the 305 probe set (Figures 3 and 4). Overall, 292 of 304 tumors and 74 out of 75 tumors were predicted as CRC in the training cohort (n = 568; 304T/264N) and validation (n = 141; 75T/66N) cohorts, respectively (Figure 5A, C), with corresponding AUCs of 0.968 and 0.984 (Figure 5B, D). In summary, prediction model performance was highly accurate and could precisely distinguish CRC from normal tissue.

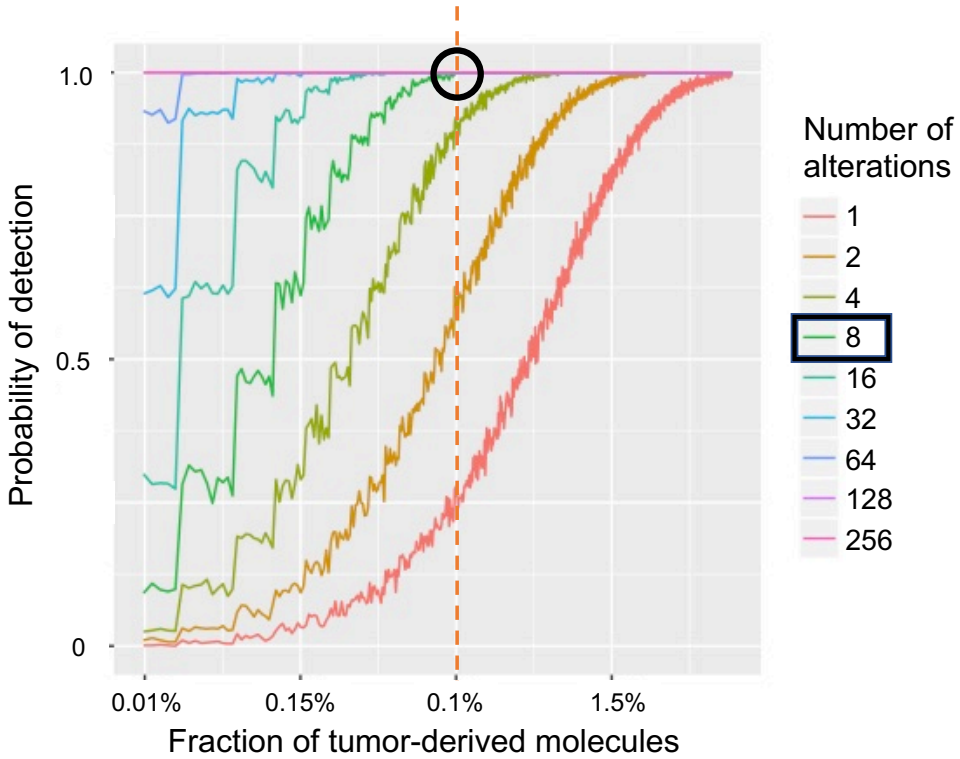

FIGURE 1. In silico simulation for setting the optimal number of DMRs.

The probabilities of detection were plotted along with fraction of tumor. The number of DMRs are annotated by color.

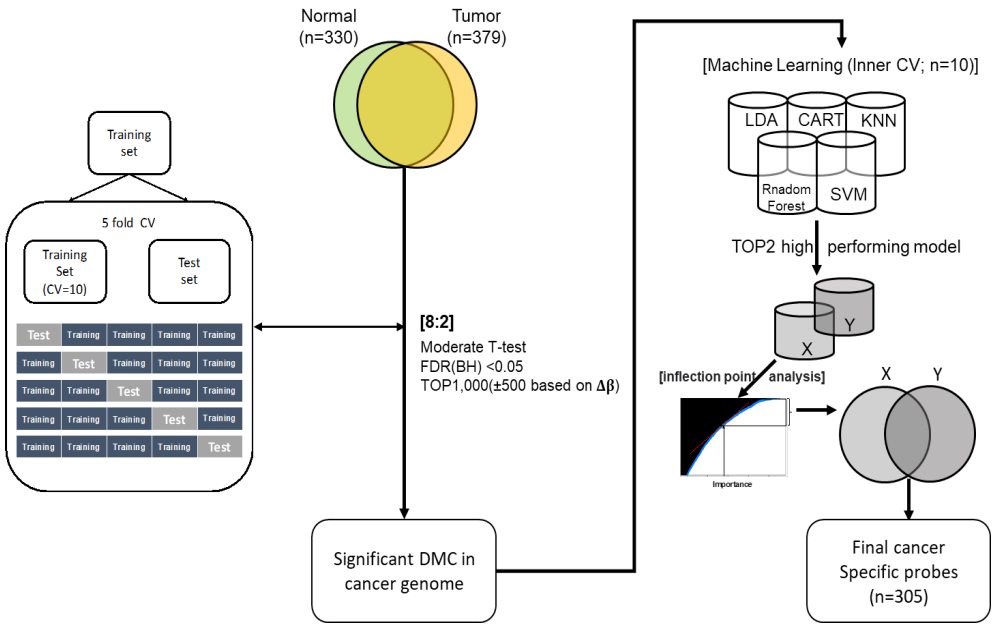

FIGURE 2. Pipeline for building the prediction model and discovering cancer-specific markers.

The outer CV for assessing model performance of the model is shown on the left box. The inner CV for building every model is shown on the right side.

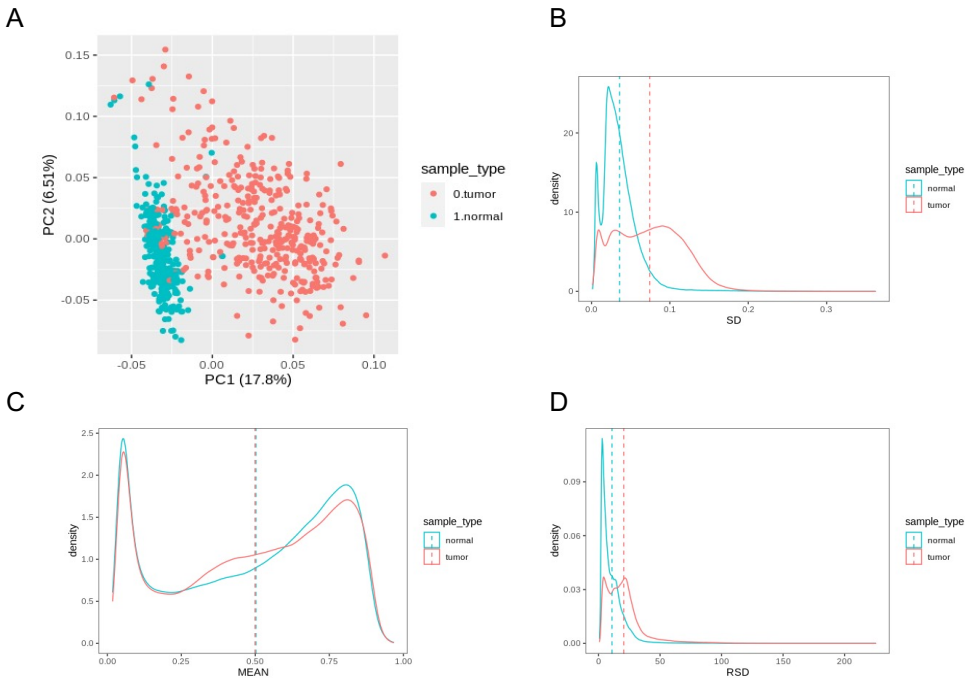

FIGURE 3. Statistical differences according to tissue type.

(A) PC 1 and 2 were calculated with the methylation level measured at ~850,000 CpG sites using the Illumina EPIC array. Standard deviation (B), mean (C), and relative standard deviation (D) were calculated as the basic statistical values. Tumor tissues (n = 379) are in red, and adjacent normal mucosa tissues (n = 330) are in green.

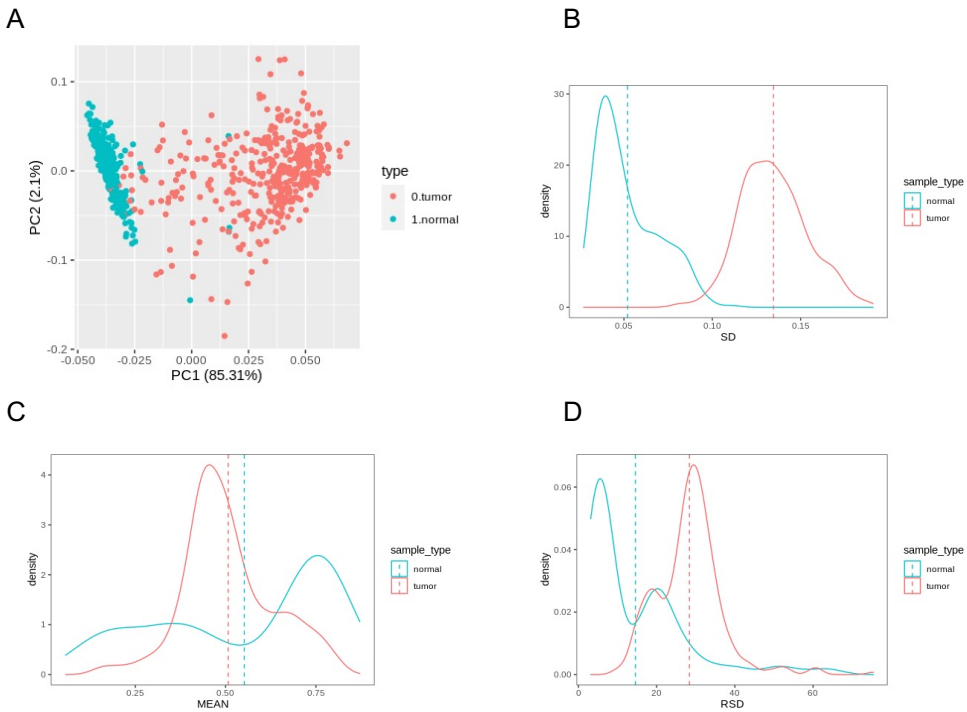

FIGURE 4. Statistical differences according to tissue type.

(A) PC1 and 2 were calculated with the methylation level measured at 305 CpG sites using Illumina EPIC arrays. Standard deviation (B), mean (C), and relative standard deviation (D) were calculated as the basic statistical values. Tumor tissues (n = 379) are in red, and adjacent normal mucosa tissues (n = 330) are in green.

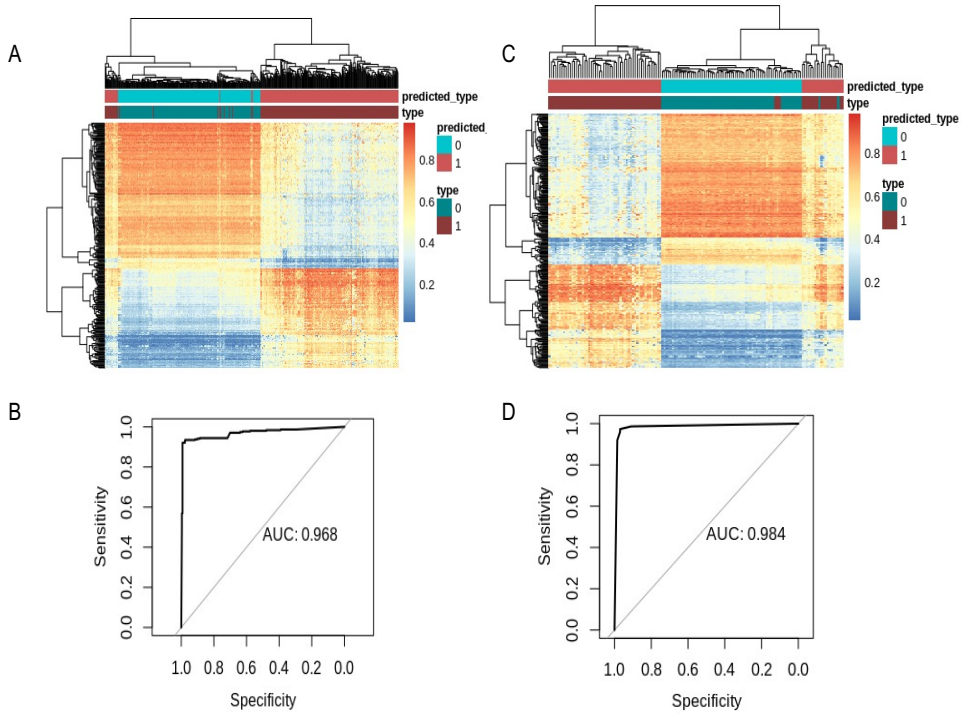

FIGURE 5. Prediction model performance using 305 DNA methylation markers for cancer diagnosis.

Unsupervised hierarchical clustering of markers differentially methylated between CRC and normal tissue DNA in the training (n=568) (A) and validation (n=141) (C) cohorts. The AUCs were calculated by ROC analysis in the training (B) and validation (D) cohorts (0, normal colon tissue; 1, colorectal cancer tissue)

3. Prediction Model Application

To validate prediction model performance, methylome data from various sources were collected and processed. First, t-distributed stochastic neighbor embedding (tSNE) analysis was performed with TCGA, EWAS, and GSE40279 datasets. Through this step, somatic tissues with tissue-specific methylation patterns and normal tissues and tumor tissues were clustered separately (Figure 6). The TOO error rate was calculated as <5% when more than 100 probes were used (Figure 7). Although TCGA and COPM were different from the platform, the intersected probes were selected, and the prediction model was re-built (Figure 8). The performance of the re-built prediction model was highly accurate (training cohort, 0.997; validation cohort, 0.976; TCGA, 1.0) (Figure 9). The re-built prediction model could accurately distinguish tumor samples from matched normal samples in almost all types of cancer except glioblastoma, kidney chromophobe, pheochromocytoma and paraganglioma, sarcoma, low-grade glioma, thyroid carcinoma, thymoma, and uveal melanoma (Figure 10A). In particular, normal PBMC methylome data (GEO40279) could accurately be predicted as not cancer. Furthermore, plasma cfDNA methylome data (GEO122126) from CRC patients could be predicted as CRC (100%; 3/3) (Figure 10B). The results showed that our prediction model could predict cancer based on plasma cfDNA, as well as gDNA.

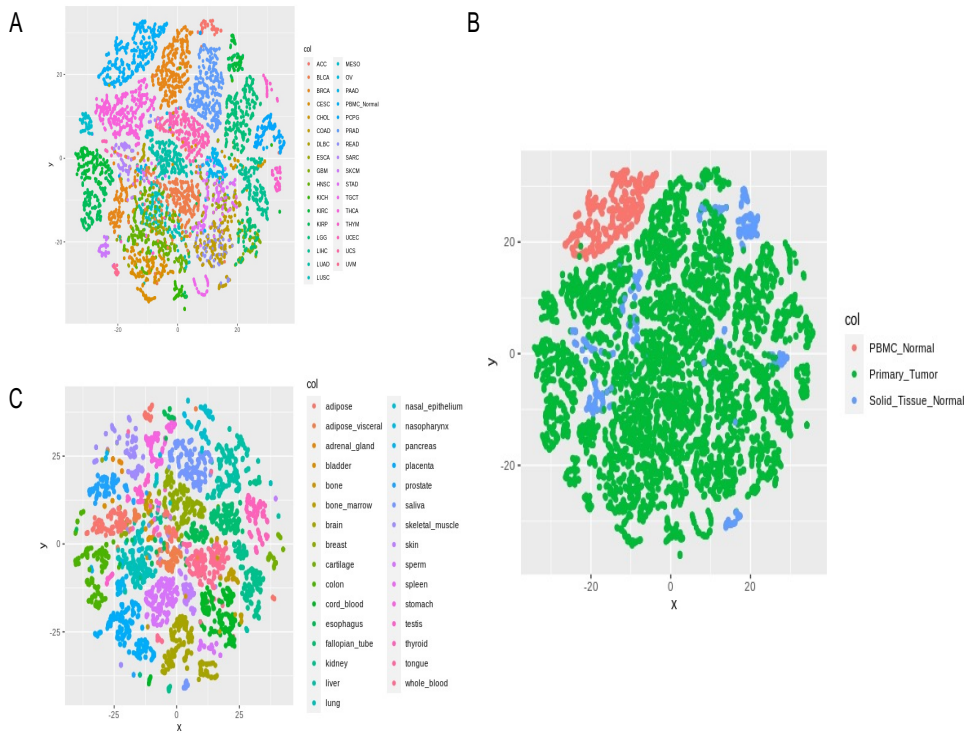

FIGURE 6. tSNE analysis with CpG methylation level.

tSNE analysis was performed using TCGA (A) and EWAS (C) datasets. Samples are annotated with tissue type-specific colors (B). (ACC, adrenocortical carcinoma; BLCA, bladder carcinoma; BRCA, basal breast invasive carcinoma; CESC, cervix squamous cell carcinoma; CHOL, Cholangiocarcinoma; COAD, colon adenocarcinoma; DLBC, Lymphoid Neoplasm Diffuse Large B-cell Lymphoma; READ, rectum carcinoma; ESCA, esophageal carcinoma; GBM, glioblastoma multiforme; HNSC, head and neck squamous cell carcinoma; KICH, kidney chromophobe carcinoma; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LGG, low grade glioma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; MESO, Mesothelioma;

OV, ovarian carcinoma; PAAD, Pancreatic adenocarcinoma; PCPG, Pheochromocytoma and Paraganglioma; PRAD, prostate adenocarcinoma; SARC, Sarcoma; STAD, stomach adenocarcinoma; SKCM, skin cutaneous melanoma; TGCT, Testicular Germ Cell Tumors; THCA, thyroid papillary carcinoma; THYM, Thymoma; UCEC, uterine corpus endometrial carcinoma; UCS, uterine carcinosarcoma; UVM, Uveal Melanoma; PBMC_Norm, PBMC from healthy individual)

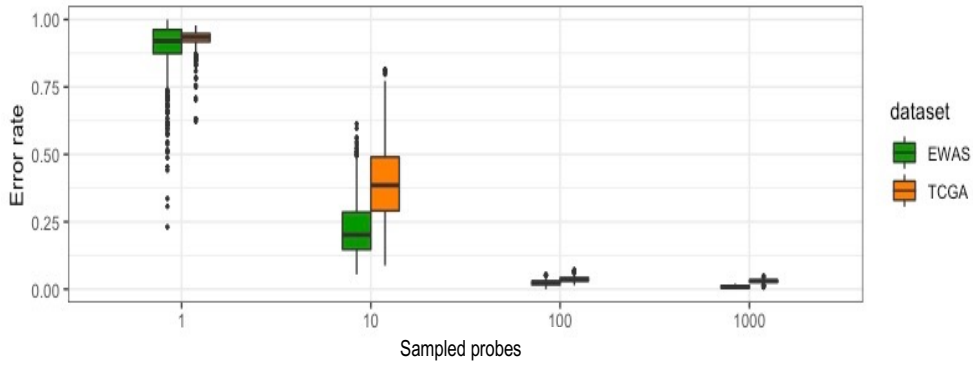

FIGURE 7. Permutation test for error rate of TOO (n = 1,000)

The error rate was calculated for sampled probes in every random forest test. The two datasets are annotated by specific colors (TCGA, n = 10,321; EWAS, n = 5,325).

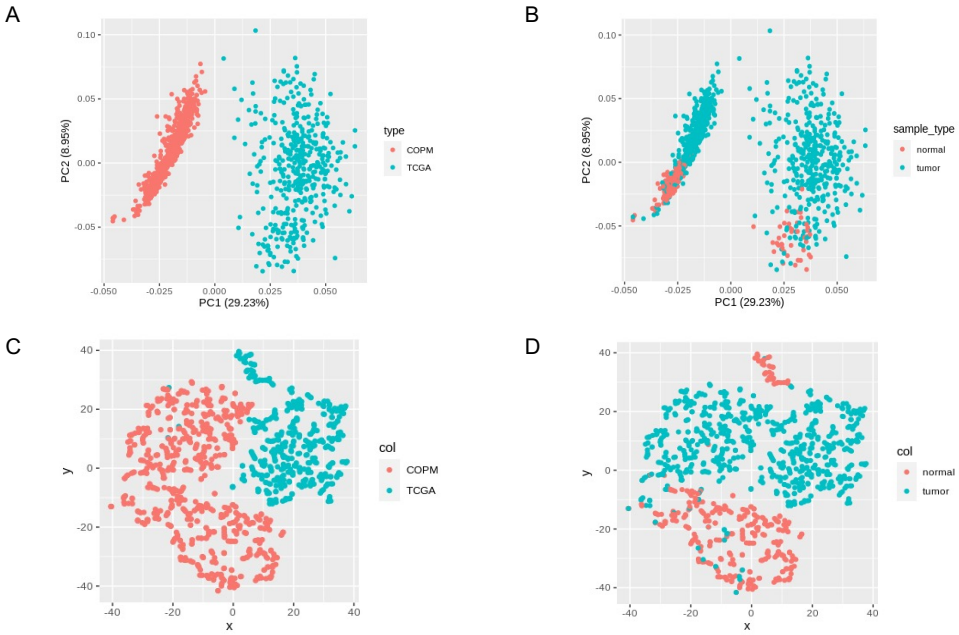

FIGURE 8. The PCA (A, C) and tSNE (B, D) analyses were performed for data and sample types. Using the intersected probes (n=76), the methylation values were used for this analysis. The data and sample types are annotated by specific colors.

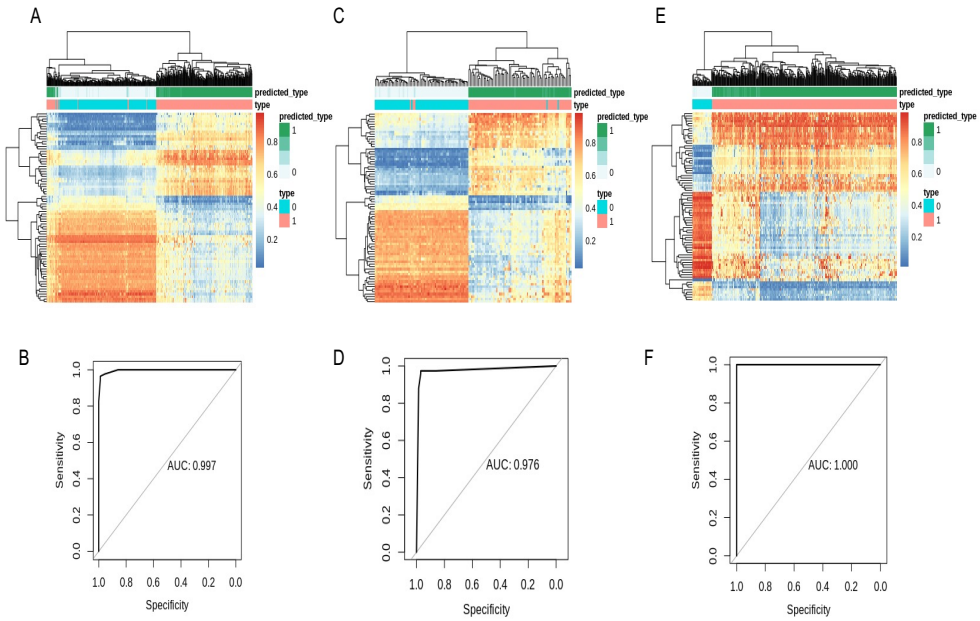

FIGURE 9. Prediction model performance using intersected 76 DNA methylation markers for cancer diagnosis.

Unsupervised hierarchical clustering of methylation markers differentially methylated between CRC and normal DNA in the training (n=568) (A), validation (n=141) (C), and TCGA COREAD (n=423) (E) cohorts. The AUCs were calculated by ROC analysis in the training (B), validation (D), and TCGA COREAD (F) cohorts. COREAD, colorectal adenocarcinoma. (0, normal colon tissue; 1, colorectal cancer tissue)

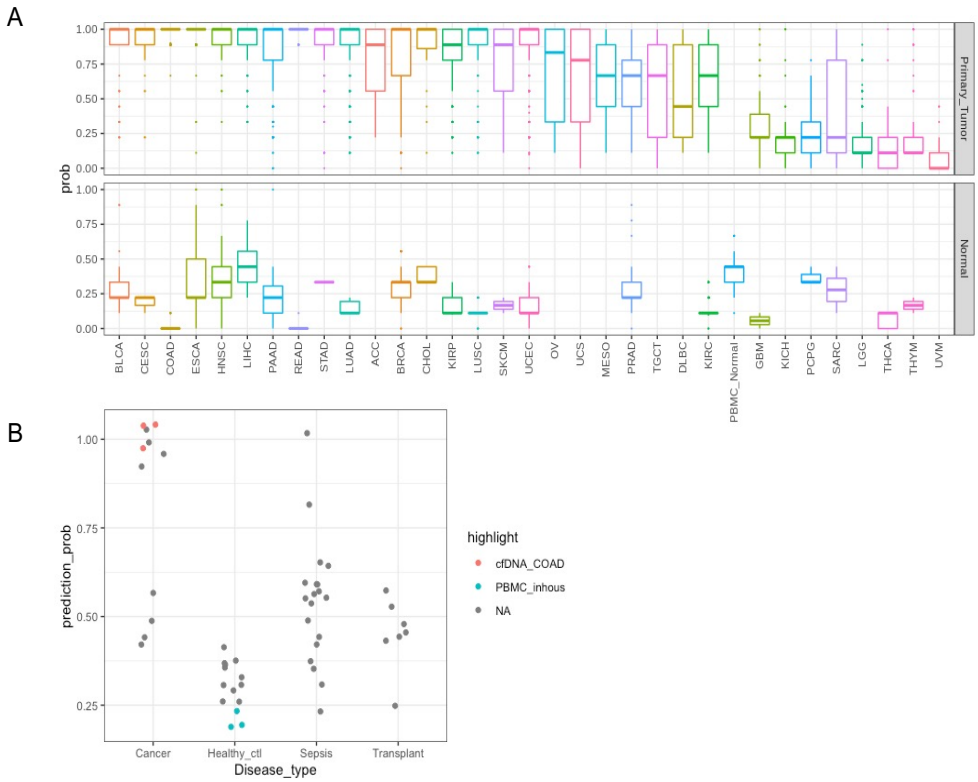

FIGURE 10. Re-constructed prediction model performance for other cancer and sample types.

TCGA database with the normal PBMC cohort (GSE40279) and plasma cfDNA database were estimated by the prediction model. (A) Primary tumor and normal tissues were predicted and plotted in TCGA dataset with GSE40279. (B) In the GSE122126 dataset, plasma cfDNA and gDNA derived from various disease types or healthy individuals were predicted and plotted.

4. Biological Rules of the Final Probe Set and Risk Score

As an important gene expression regulator, DNA methylation in cis-regulatory elements could be crucial to either carcinogenesis or disease progression (17). The 305 probes were annotated using ChromHMM and subjected to pathway analysis with Metascape (16). Overall, 160 of 305 (52.5%) probes were annotated in regions related to gene regulation (Figure 11). Pathway enrichment analysis showed that the gene set containing probes in the promoter regions were enriched in developmental processes and carcinogenesis (GO: 0032502, log10 P = -4.28; C4721208, log10 P = -3.80) (Figure 12). In addition, 9 of the 25 probes annotated in either promoter regions or coding sequence (CDS) regions were correlated with mRNA expression (Figure 13). The risk scores for OS (20 probes) and PFS (133 probes) were also determined. Poor prognosis could be predicted in high-risk score group (OS, P = 0.073; PFS, P = 0.0026) (Figure 14). These results were more statistically significant than when all probes were used (OS, P = 0.14; PFS, P = 0.015) (Figure 15). Although the risk score was not correlated with age or sex, the risk score for OS was statistically higher in the late-stage group ($P < 0.05$, t-test) (Figures 16, 17, 18). In conclusion, the probe set was associated with the cancer-related pathway, and the risk score could be a potential prognostic marker.

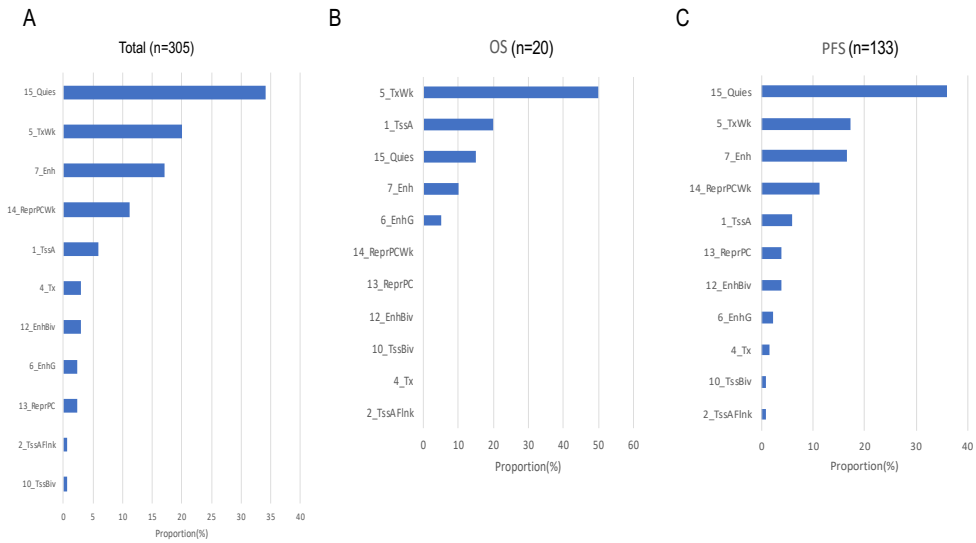

FIGURE 11. Chromatin status correlated with the probe set (ChromHMM).

The genomic regions of CpG sites were annotated using the ChromHMM database. The proportion of the total probe set (A), the probe set for OS risk score (B), and the probe set for PFS risk score (C) were calculated.

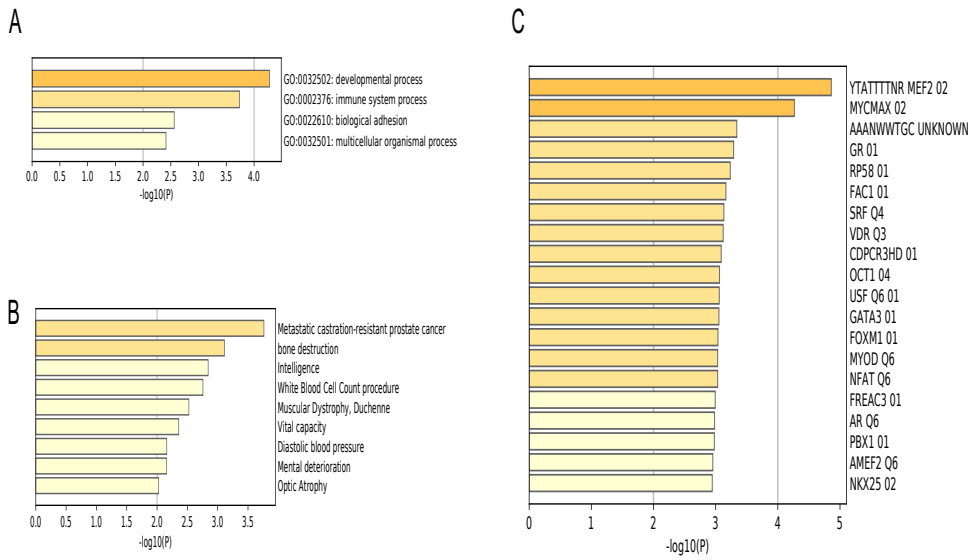

FIGURE 12. Pathway analysis using various databases through Metascape.

Pathway analysis (A), the gene-related disease (B), and the regulator of gene list (C) were analyzed and bar plotted with P values transformed by -log10

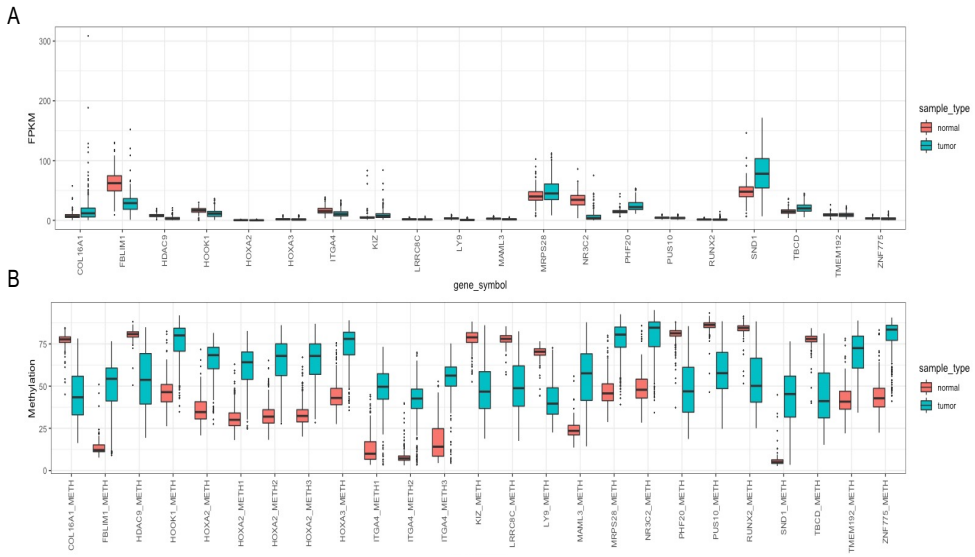

FIGURE 13. Correlation between methylation level and gene expression.

mRNA expressions were matched with the methylation levels of the gene-related regions. (A) FPKM values of gene lists related with the target probe set were plotted. (B) Beta values for CpG sites matched with gene list were plotted.

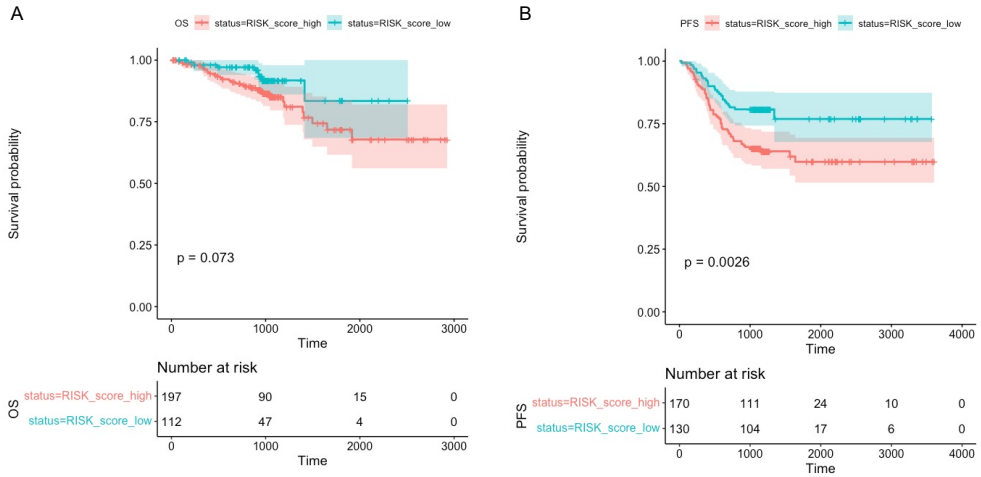

FIGURE 14. The risk score using the subset of 305 probe set as prognostic marker.

The high-risk score group for OS had poor prognosis compared to the low-risk score group for (A) OS (log-rank test, P = 0.073) and (B) PFS (log-rank test, P = 0.0026).

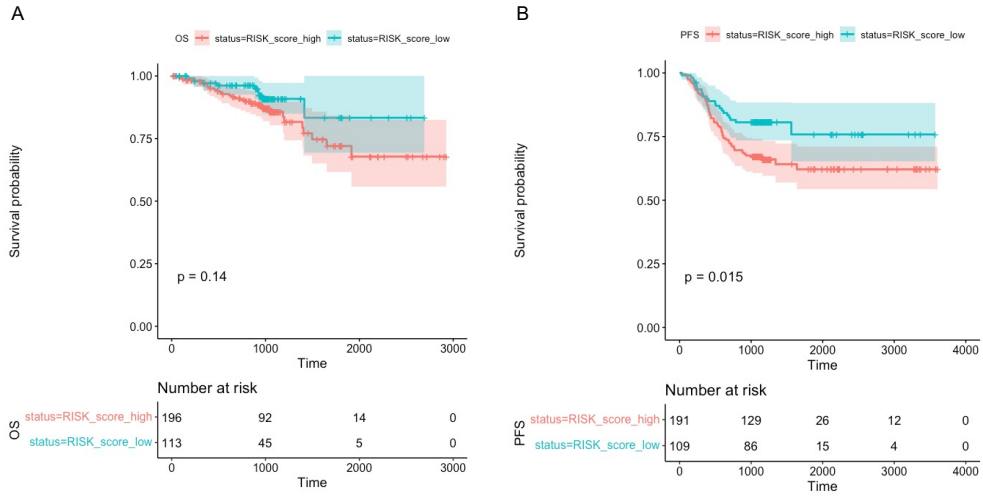

FIGURE 15. Risk score using the total of 305 probe sets as prognostic markers.

The high-risk score group for OS had poor prognosis compared to the low-risk score group for (A) OS (log-rank test, P = 0.14) and (B) PFS (log-rank test, P = 0.0026).

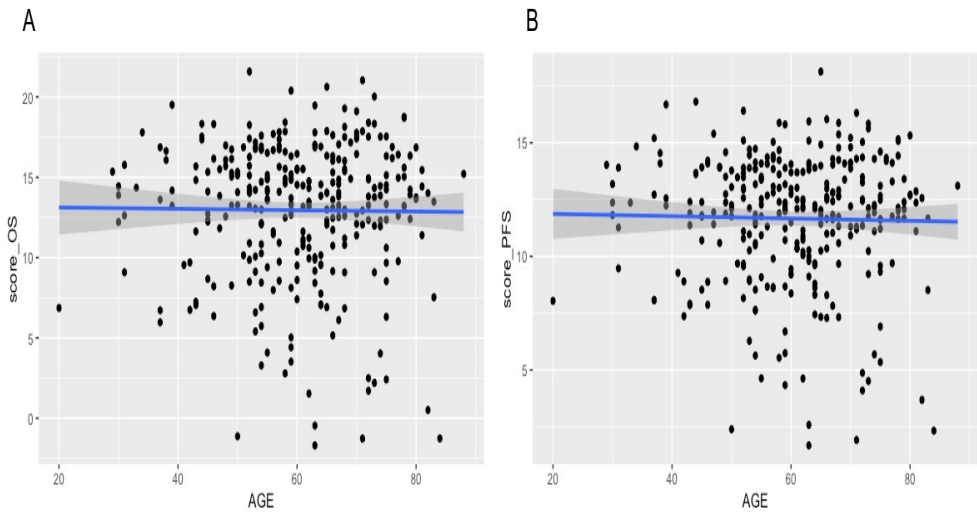

FIGURE 16. The association risk score with cancer patient age.

Risk scores for OS (A) and PFS (B) were plotted.

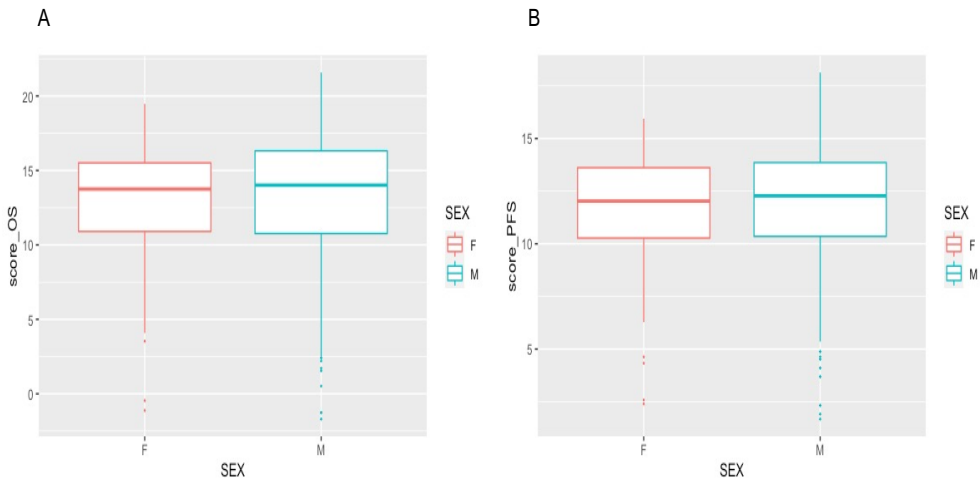

FIGURE 17. The association risk score with cancer patient sex.

Risk scores for OS (A) and PFS (B) were plotted.

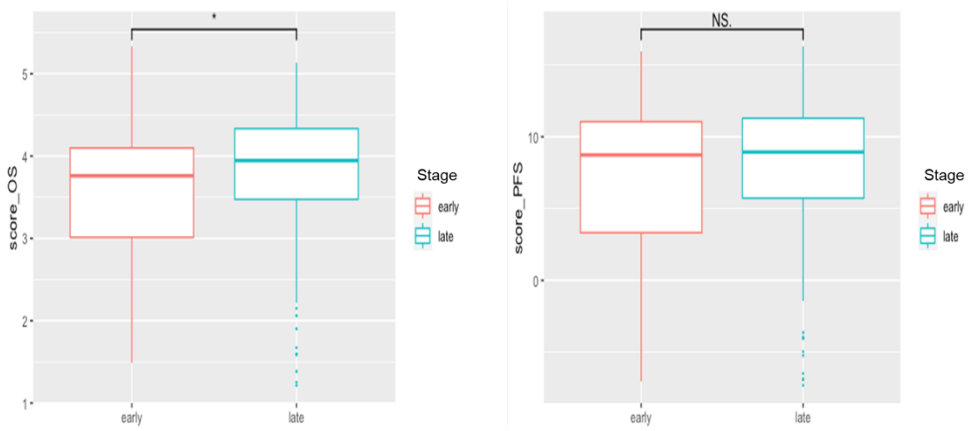

FIGURE 18. The association risk score with cancer stage.

Risk scores for OS (A) and PFS (B) were plotted. P values were calculated with Student's t-tests (OS, P = 0.05; PFS, P = N.S.). Early-stage group included the stage I and II and late-stage group included the stage III and IV.

DISCUSSION

An accurate prediction model and diagnostic markers were developed for Korean patients with CRC. To our knowledge, ours is the first description of CRC-specific methylation markers to build a prediction model using data from the largest Korean CRC cohort (COPM dataset; 330 adjacent normal mucosa tissues and 379 CRC tissues). Furthermore, risk scores based on final diagnostic marker subset could predict CRC patient prognosis.

The analysis was based on a machine learning algorithm with the goals of identifying the best model for predicting CRC and selecting an optimal probe set using genome-wide methylation data. Due to the nature of machine learning, the number of patients in the cohort is an important feature. To effectively utilize this feature, the cohort was separated prior to analysis. Comprehensive analysis is needed for the properties of DNA methylation. In a previous report, cancer tissue-specific methylation pattern was discovered by machine learning and could predict the TOO (6, 18). Although this model was built for CRC, predicting TOO would be possible by adding other types of cancer. The model accurately predicted CRC patients in TCGA and GEO datasets (n = 9,660 tissues and 59 plasma cfDNA) as having CRC. The prediction model probes were 305 CpG sites in the human genome. Therefore, the prediction model can analyze not only 850K array-based methylation values but also other platform-based methylation values. Targeted sequencing is one of the representative methods for this approach and liquid biopsy, where the amount of cell-free DNA is very low, so a platform based

on bisulfite conversion would be not applicable (19, 20). To overcome these issues, methylation levels can be detected using bisulfite-free methods based on enrichment using methylated CpG sites with specific antibodies and enzymatic C to T conversion (21) (22). The use of these methods would make the prediction model more powerful for understanding CRC.

The 305 CpG sites were associated with developmental processes and carcinogenesis (GO: 0032502, log10P = -4.28; C4721208, log10P = -3.80). DNA methylation is a critical gene regulation mechanism. In cancer, cis-regulatory elements of tumor suppressor genes and oncogenes are hypermethylated and hypomethylated, respectively, in tumor cells compared to normal cells. This underscores the importance of coordinating probes on the genome. The ChommHMM database is well-organized with regard to tissue-type chromatin status. In this study, 160 of 305 probes (52.5%) were annotated in regions related to gene regulation. In addition, 9 out of 25 probes annotated in either promoter or CDS regions correlated with mRNA expression. Methylation level could affect gene expression CpG site clusters rather than single CpG loci. Indeed, CpG islands are located in 40% of all gene promoters (23). For this reason, not all CpG site methylation was directly correlated with gene expression.

In summary, this accurate prediction model yielded risk scores of informative methylation patterns detected in CRC or a broad range of cancer types, with prediction performance approaching the goal for large-scale screening based on Korean CRC data. These results support the feasibility of employing this machine learning-based methylation analysis for early CRC

detection in the Korean population.

II. Combined analysis of ctDNA mutation and fragment size for predicting prognosis of colorectal cancer

ABSTRACT

The fragment size of cell-free DNA (cfDNA) was well characterized. It has been reported the size of cfDNA derived from the patients with cancer was shorter than the non-cancer individuals and short fragments were associated with the circulating tumor DNA (ctDNA) among cfDNA. Short fragments of ctDNA have been proposed as predictive biomarkers of disease, although their role in colorectal cancer remains unknown. We hypothesized that the fragment size of cfDNA, which include ctDNA fragments, isolated from colorectal cancer patients have the potential for prognosis after chemotherapy. Two hundred eighty plasma samples from 62 patients with colorectal cancer were collected along with plasma from 50 healthy controls. Sixty-two individuals were recruited through prospective clinical studies at Seoul National University Hospital. The chemotherapy backbone of the Cetuximab or Bevacizumab containing regimen was chosen between FOLFIRI or FOLFOX. Blood samples were obtained prior to chemotherapy and after every four cycles of chemotherapy until disease progression. ctDNA was detected by target capture panel. This panel sequencing is a tumor agnostic panel consist of 106 genes, including 10 gene fusion and MSI. Based on the panel sequencing data, the genetic alterations and the fragment size of cfDNA were calculated by our algorithm. And the fragmentation ratio was defined by the ratio of the read fragment proportion in size range P1 (100 – 155 bp) and P2 (160 – 180 bp). For survival analysis, the optimal thresholds separating the group based on the clinical response (responder vs. non-responder) were

calculated by the ROC analysis. Compared to cfDNA from healthy controls, the cfDNA fragment sizes from patients with colorectal cancer were significantly shorter (169.585 bp vs. 173.964 bp; P = 1.119e-09). Additionally, ctDNA fragments harboring mutant alleles were shorter than those harboring reference alleles of somatic mutations but not germline mutations (155.853 bp vs. 160.613 bp; P = 0.0007829 & 160.911 bp vs. 159.889bp; P = 0.992). Further, the clonality inferred from the variant allele frequency (VAF) of somatic mutation was negatively correlated with the size of the ctDNA fragment. The read fragment proportion in size range P1 was significantly associated with the clonality. We divided the samples into the following groups: baseline, first follow-up, before-last follow-up, and last follow-up (end point). We calculated the mean size of DNA fragment and the fragmentation ratio with each longitudinal sample. In the before-last follow-up group, the fragmentation ratio was found to accurately predict the prognosis of patients with colorectal cancer (average survival of 9 months; P = 0.016). The fragmentation ratio was also found to increase in a time-dependent manner (P = 0.018; ANOVA). In summary, we identified the fragmentation ratio as a prognostic marker for survival of patients with colorectal cancer.

Key words : Colorectal cancer (CRC), Fragmentomics, non-genetic marker, cell-free DNA(cfDNA), circulating tumor DNA(ctDNA), Prognosis

Student Number : 2018-37966

INTRODUCTION

In the cell nucleus, DNA is wrapped around proteins called histones. Upon cell death (apoptosis, necrosis, etc.), DNA is released into the bloodstream where it can freely circulate; such DNA is referred to as cell-free DNA (cfDNA) (24). As DNA fragmentation can result from apoptosis (25), size patterns of cfDNA fragments are highly dependent on the extent of nucleosome packaging, with shorter fragments commonly associated with transcription factor-binding sites (26) (27). The mean size of cfDNA fragments is approximately 166 bp (28). However, such fragments tend to be shorter in patients with various diseases or in pregnant patients compared to healthy controls (29). Recent studies have reported on the novel applications of cfDNA fragments.

For example, the size of circulating tumor DNA (ctDNA) tends to be shorter than that of cfDNA derived from normal tissues (30). In a study of the KRAS oncogene in patients with early pancreatic cancer, the size of a ctDNA fragment harboring a mutant allele was shorter than a ctDNA fragment harboring a reference allele using targeted deep sequencing (31). Using both targeted deep sequencing and shallow whole-genome sequencing (sWGS), copy number alterations were calculated via in silico size selection, which can enrich the tumor fraction harboring small cfDNA fragments (30, 32). On the other hand, using sWGS, cancer-specific regions harboring aberrant cfDNA fragments were identified, and a prediction model was built using a machine learning algorithm (33). The sizes of the cfDNA fragments examined were

found to be aberrant in DNase1l3-knockout mice, and their expression in liver cancer tissues was lower than that in adjacent normal liver tissues (34). However, research on ctDNA fragment size using targeted deep sequencing is still insufficient.

Herein, deep sequencing of cancer-related genes from cfDNA samples isolated from 62 patients with colorectal cancer and 50 healthy controls was performed. In this prospective cohort study, the ability of particular subgroups and sizes of cfDNA fragments to predict the prognosis of patients with cancer was assessed.

EXPERIMENTAL DESIGN

1. Information of cohort

Two hundred eighty plasma samples from 62 patients with colorectal cancer were collected along with plasma from 50 healthy controls. 62 individuals were recruited through prospective clinical studies at Seoul National University Hospital. The chemotherapy backbone of the Cetuximab or Bevacizumab containing regimen was chosen between FOLFIRI (5-Fluorouracil, Leucovorin, Irinotecan) or FOLFOX (5-Fluorouracil, Leucovorin, Oxaliplatin), at the discretion of the treating physician. Response evaluation was done in accordance to RECIST 1.1 using contrast-enhanced computed tomography (CT) obtained at baseline and repeated every four cycles or at clinician's suspicion of progressive disease. All patients provided written informed consent before any study-specific procedures. The protocol of this study was reviewed and approved by the Institutional Review Board (IRB) of SNUH (IRB number: 1805-049-944) and was conducted in accordance with the Declaration of Helsinki in biomedical research involving human subjects.

2. Blood sample collection and cell-free DNA extraction

Serial blood samples were obtained before treatment initiation (≤ 7 days before treatment) and at the time of response evaluations. Whole blood (8–10 mL) was collected into EDTA tubes during routine phlebotomy. Blood samples were centrifuged with Ficoll solution at 1500 × g for 15 min. Plasma was then separated by centrifugation at 16,000 × g for 10 min to remove cell

debris, after which 1 mL aliquots were placed in Eppendorf tubes and stored at − 80 °C before extraction. This protocol was performed within 20 min of blood collection to prevent cell-free DNA (cfDNA) degradation and release of genomic DNA from dying blood cells. cfDNA was isolated according to the manufacturer's instructions from 2 to 4 mL plasma using a cfKapture™ Kit (MagBio Genomics, USA) and quantified using a 2200 TapeStation (Agilent Technologies, Santa Clara, CA, USA). Peripheral blood mononuclear cell (PBMC) was separated following this protocol. Genomic DNA was isolated from PBMC using a QIAamp DNA Mini Kit (Qiagen).

3. Targeted deep sequencing and bioinformatics analysis

Briefly, ~ 20 ng of cfDNA and 100 ng of leukocyte DNA per patient were used for sequencing library preparation. A DNA NGS library was constructed using a IMBDx NGS DNA Library Prep Kit. Solution-based target enrichment was performed at IMBDx, Inc. (Seoul, South Korea), using a target capture panel (106 cancer related genes). Captured DNA libraries were sequenced using an Illumina NextSeq 550 platform (Illumina, San Diego, CA, USA) in 2 × 150 bp paired-end mode. All sequencing reads from the samples were generated as fastq format. Filtered fastq files were aligned to the human reference genome (hg38) using Burrows-Wheeler Aligner (v0.7.17) "mem" algorithm (35). Reads mapped on the target regions were extracted, and collapsing was carried out using Genecore (36). In order to variant calling, Initial variant calls were compiled using VarDict (37), then a series of in-house filtering steps were applied. The remaining calls were annotated using

SnpEff (38), SnpSift (39), and VEP (40) for functional effect prediction and tagging information from various databases. In order to distinguish germline mutation from somatic mutation, the database called for GNOMAD was used for germline mutation. To analyze the fragment lengths of cfDNA molecules, I sorted that each read pair from a cfDNA molecule had a Phred quality score ≥30 using Samtools (41). Then, I collected the read pairs contained the mutated (or wild-type) allele at the given genomic position. This was done using Bedtools and Pysam. Finally, for each read pairs, the fragment length was calculated by from the end of R1 tag to end of R2 tag. The performance of this was not different from Picard tools 'CollectInsertSizeMetrics'(Figure 1). This step was performed total three times for total regions of panel, regions of patient specific somatic and germline mutation. Using cfDNA fragment size, the distribution curve was calculated and the area under curve (AUC) for two regions was calculated (P1: 100 – 155 bp, P2: 160 – 180 bp). These regions were reported as representative regions for tumor fraction enriched (P1) and normal-like (P2) regions (30).

4. ROC analysis

In order to determine the optimal threshold for classifying clinical response group, ROC analysis was performed using R package 'pROC'. Through this step, the ROC curve and AUC were calculated for each variable. And the optimal threshold for classifying two group was calculated (42).

5. Survival analysis

Using R package 'survminer', the survival rate for OS and PFS were calculated. And Kaplan-Meier plot were visualized and p value were calculated by Log-rank test.

6. Statistical test

Using R, the following distribution of cfDNA/ctDNA size were evaluated by Kolmogorov-Smirnov test and Q-Q plot: "Colorectal cancer patients vs. Healthy controls" and "read fragments harboring mutant allele vs. read fragments harboring wild-type allele". The of Progression-free survival (PFS) was evaluated as outcome measures for each marker related to fragment size. The following statistical significance for Kaplan-Meier analysis were evaluated by Log-rank test: "short fragment vs. long fragment", "AUC_{P1} high vs. AUC_{P1} low", "AUC_{P2} high vs. AUC_{P2} low" and "fragmentation ratio high vs. fragmentation ratio low". The ANOVA test was performed for significance of the fragment size among time point, response group.

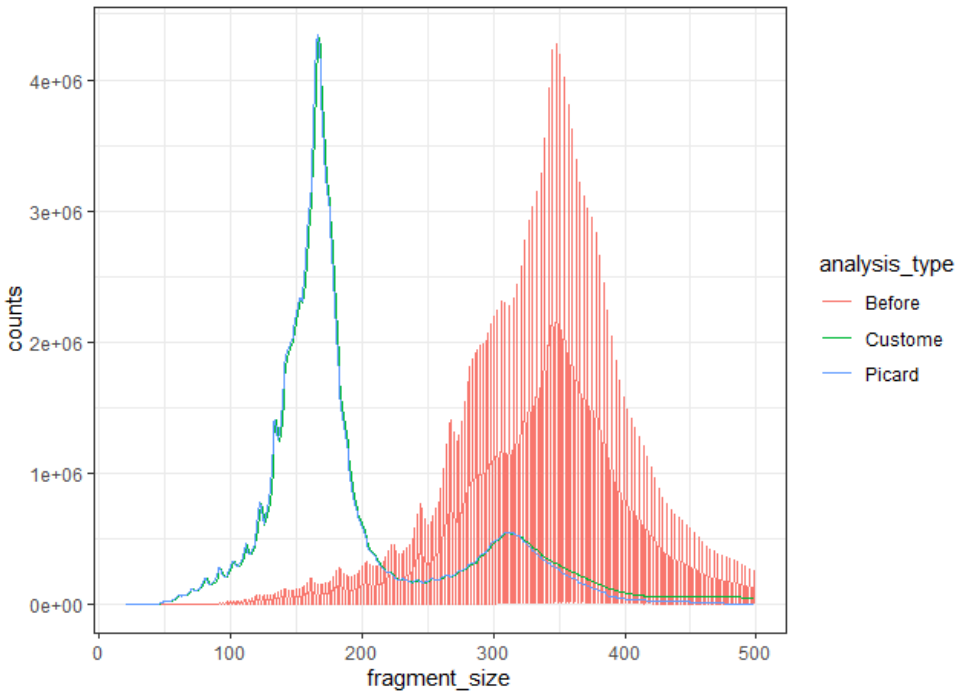

FIGURE 1. The benchmarking of the DNA fragment size calculation

A distribution curve for patients with colorectal cancer was generated before correction (red) and after correction using the Picard tool (blue) and custom script (green).

RESULTS

1. Patient characteristics

Two hundred eighty plasma samples were collected from 62 colorectal cancer patients (Table 1). Median age of cohort was 62 (37 - 79) and proportion of male and female was 61.3% (n=38) and 38.7% (n=24). Anatomical sites of primary tumor were ascending colon (9.7%; n=6), transverse colon (1.6%; n=1), sigmoid colon (53.2%; n=34), cecum (6.5%; n=4) and rectum (29.0%; n=18). Pathological subtypes were adenocarcinoma W/D (3.2%; n=2), adenocarcinoma M/D (87.1%; n=54) and adenocarcinoma P/D (9.7%; n=6). The status of microsatellite instability was MSI-H (1.6%; n=1), MSI-L (11.3%; n=7), MSS (82.3%; n=51) and NA (4.8%; n=3). Forty-one (66.1%) patients received FOLFIRI based cytotoxic chemotherapy and 20 (32.3%) patients received FOLFOX based cytotoxic chemotherapy. There were 57 patients who received the targeted therapy (Cetuximab; n=35, Bevacizumab; n=22). For each patient, I generated targeted deep sequencing data and assessed the utility of ctDNA fragment size as prognostic marker.

Categories		Number of patients (%)
Age at diagnosis, median (range)		62 (37 – 79)
Sex	Male	38 (61.3 %)
	Female	24 (38.7 %)
Disease presentation at enrollment	Metastasis	46 (74.2 %)
	Recurrence (metastatic)	16 (25.8 %)
Primary tumor site	Cecum	4 (6.5 %)
	Ascending colon	6 (9.7 %)
	Transverse colon	1 (1.6 %)
	Sigmoid colon	34 (53.2 %)
	Rectum	18 (29.0 %)
Metastasis site	Liver	46 (74.2%)
	Lung	20 (32.3%)
	Peritoneal seeding	12 (19.4%)
	Lymph nodes	12 (19.4%)
	Other organs	7 (11.3%)
Pathology	ADC, W/D	2 (3.2 %)
	ADC, M/D	54 (87.1 %)
	ADC, P/D	6 (9.7 %)
Microsatellite instability	MSS	51 (82.3 %)
	MSI-L	7 (11.3 %)
	MSI-H	1 (1.6 %)
	Not available	3 (4.8 %)
Cytotoxic chemotherapy*	FOLFIRI	41 (66.1 %)
	FOLFOX	20 (32.3 %)
Targeted therapy**	Cetuximab	35 (56.5 %)
	Bevacizumab	22 (35.5 %)

* 1 loss to follow up after pre-treatment evaluation
** 4 treated without targeted therapy

Table 1. Clinicopathological information of the prospective patient cohort.

2. Differences in cfDNA size in healthy controls

Recently, it was reported that cfDNA fragment size in cancer patients is shorter than that in healthy individuals (43). In our cohort, the cfDNA in 280 plasma samples from patients with colorectal cancer and 50 plasma samples from healthy controls was subjected to targeted deep sequencing, and the cfDNA fragment size was estimated. As expected, the size of the cfDNA fragments from patients with colorectal cancer tended to be shorter than those from healthy controls (169.6 bp vs. 174.0 bp; P = 1.119e-09; Kolmogorov-Smirnov test) (Figure 2). One hypothesis for this phenomenon is that plasma cfDNA from patients with colorectal cancer harbor a highly aberrant fraction of DNA from the primary tumor. The size distribution between the cfDNA fragments harboring mutant alleles and wild-type alleles from patients with colorectal cancer (METHOD) was compared. The size of cfDNA fragments harboring mutant alleles was significantly shorter than the size of fragments harboring wild-type alleles (P = 0.0007829; Kolmogorov-Smirnov test) (Figure 3). The shorter cfDNA fragments from patients with colorectal cancer appear to be the indirect result of ctDNA harboring mutant alleles.

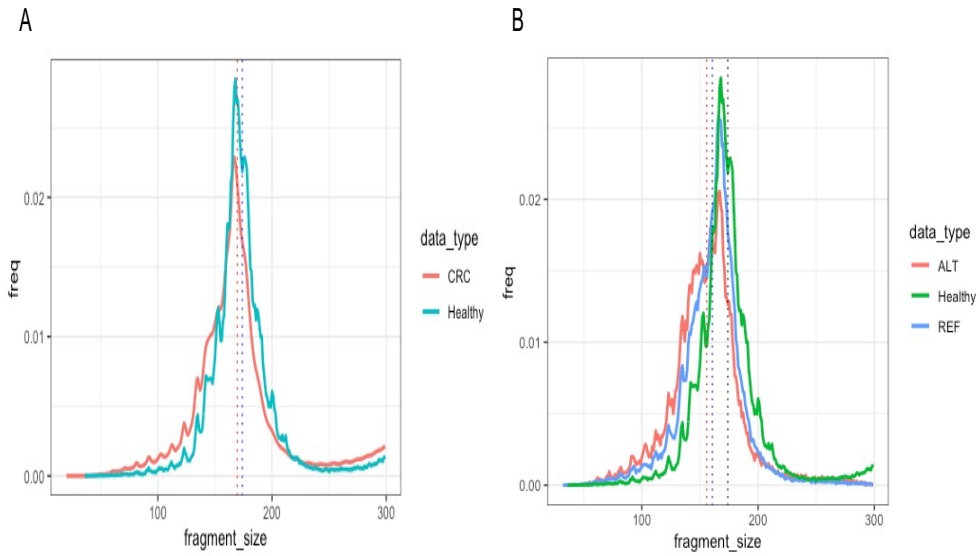

FIGURE 2. Distribution curve of cfDNA fragment size in patients with colorectal cancer (n=62) and in healthy controls (n=50).

A. The total fragment sizes were calculated and plotted. The mean fragment sizes from patients with colorectal cancer and healthy controls were 169.585 bp and 173.964 bp, respectively; P = 1.119e-09; Kolmogorov-Smirnov test. B. For patients with colorectal cancer, read fragments were separated into those harboring mutant alleles (ALT) and wild-type alleles (REF) and plotted. The mean ALT and REF sizes were 155.853 bp and 160.613 bp, respectively.

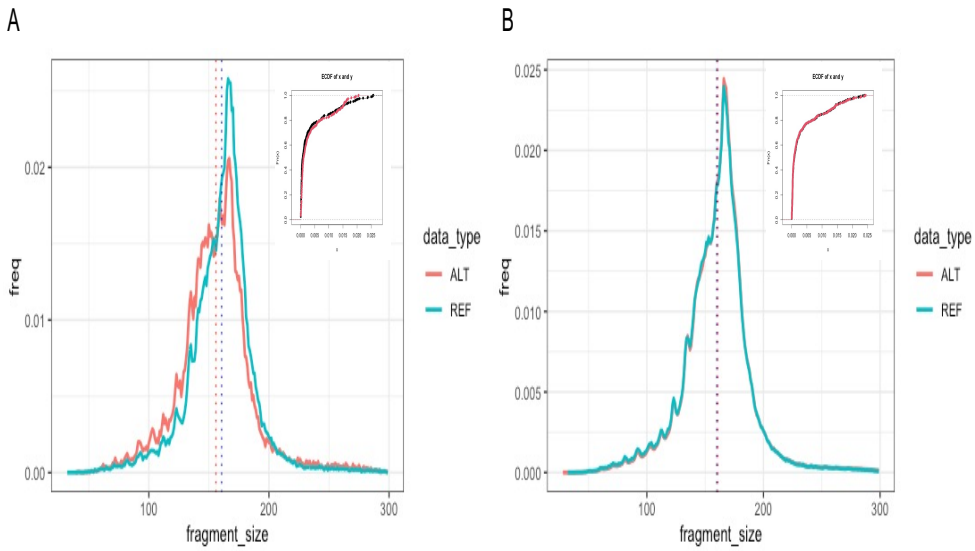

FIGURE 3. Distribution curve of cfDNA fragments by mutation type.

Read fragments from DNA harboring somatic (A) or germline (B) mutations were estimated and plotted. The mean sizes of somatic mutation fragments harboring ALT or REF were 155.853 bp and 160.613 bp, respectively (Kolmogorov–Smirnov test, P = 0.0007829). The mean sizes of germline mutation fragments harboring ALT or REF were 160.911 bp and 159.889 bp, respectively; P = 0.992; Kolmogorov–Smirnov test.

3. The association of ctDNA size with clonality

Various types of mutations derived from PBMCs, tumor tissues, and normal tissues are normally detected in plasma cfDNA (44). To discern whether cfDNA fragment size depends on mutation type, ctDNA fragments harboring somatic mutations were separated from cfDNA fragments harboring germline mutations. Then, the size of each fragment was estimated. Although the ctDNA fragment size was significantly different between those harboring somatic mutant alleles and those harboring wild-type alleles, cfDNA fragment size was not affected by the presence of a germline mutation (somatic mutation; P = 0.0007829, germline mutation; P = 0.992) (Figure 3). This is the one of the features which could be distinguished somatic variants from germline variants. Next, the association of the variant allele frequency (VAF) of somatic mutations with ctDNA fragment size was determined. As the VAF of mutations, which represent either the clonality or purity of the samples, is roughly based on the cancer genome (45), three groups of clonality based on the maximum VAF of somatic mutations were defined and identified in patients with colorectal cancer. Very low clonality (max VAF < 10%), low clonality (10% < max VAF < 40%), and high clonality (max VAF > 40%) were identified in 29, 9, and 24 patients, respectively (Figure 4). Additionally, the mean size of all fragments, the mean size of fragments harboring mutant alleles, the mean size of fragments harboring wild-type alleles, the proportion of short fragments (P1; 100–155 bp), and the proportion of reference cfDNA fragments (P2; 160–180 bp) were calculated. Interestingly, the proportion of short fragments was significantly correlated with the maximum VAF of

somatic mutations (Pearson correlation r = 0.86; P = 2.9e-13) and was greater in the high clonality group (P < 0.001) (Figure 5 and 6). Among patients with a VAF of somatic mutations above 10%, ctDNA fragments with mutant alleles were shorter than those harboring wild-type alleles in 12 out of 33 patients (Figure 7). In summary, shorter ctDNA fragments were more frequently found in patients in the high clonality group.

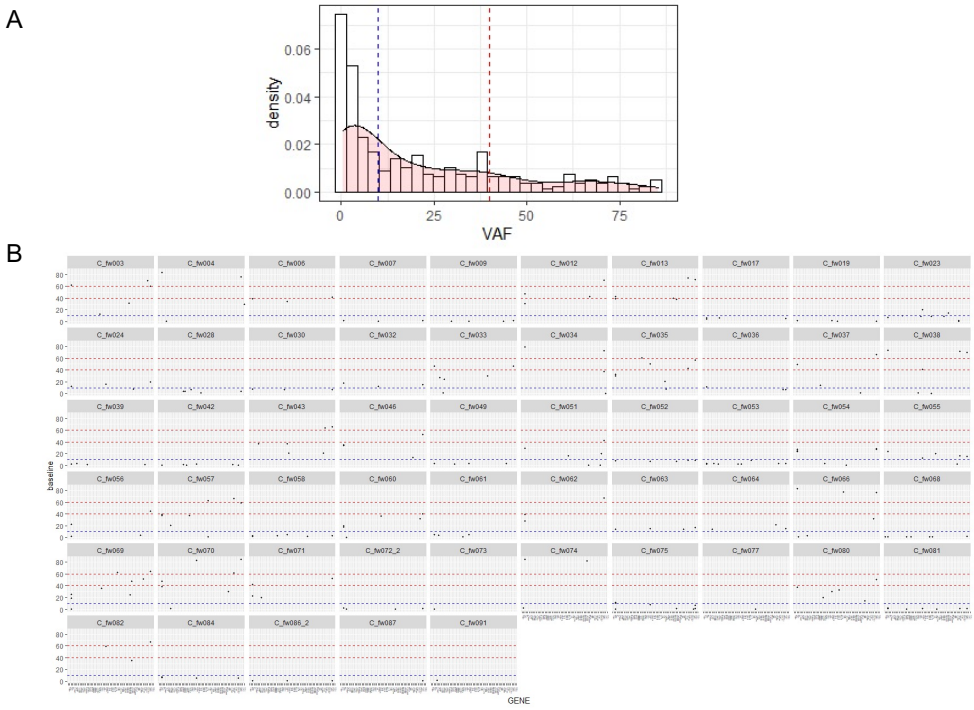

FIGURE 4. Distribution curve of the VAF of somatic mutations detected in plasma cfDNA.

A. The VAFs of somatic mutations were plotted. B. The VAF distribution was plotted for each patient, individually (10 %, 40%, and 60% VAF are annotated by blue, red, and red).

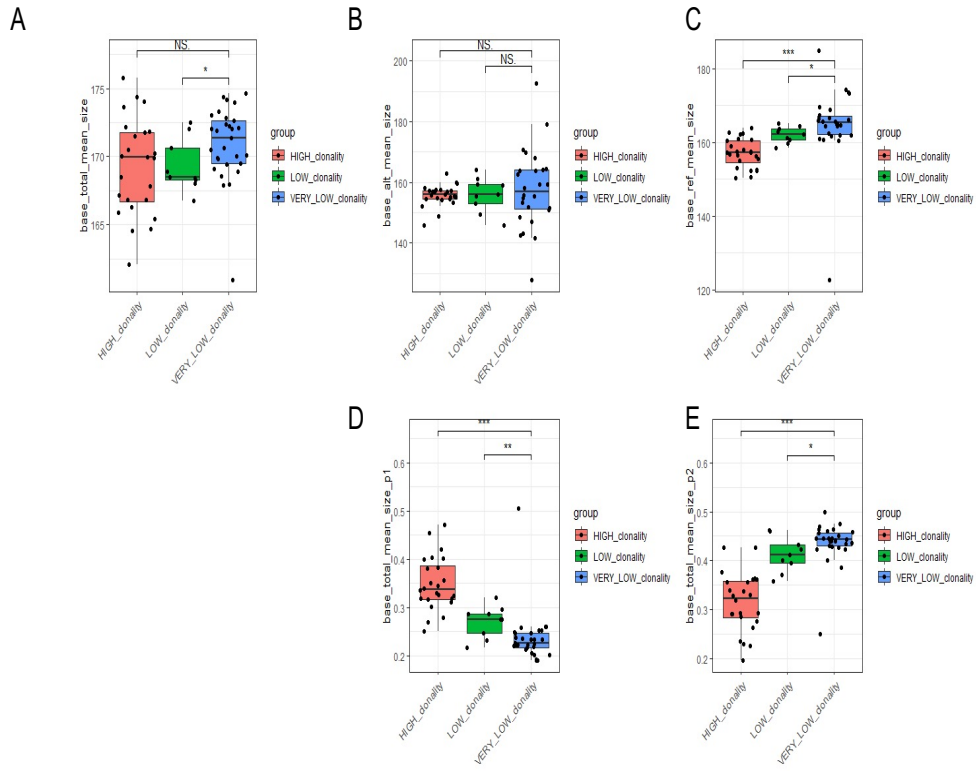

FIGURE 5. The association between clonality and ctDNA fragment size.

Variables were plotted along with clonality group: A. Mean total fragment size in baseline samples. B. Mean ALT size in baseline samples. C. Mean REF size in baseline samples. D. AUC for short fragments (100–155 bp). E. AUC for reference fragments (160–180 bp). NS, not significant; *, $P < 0.05$; **, $P < 0.01$, ***, $P < 0.001$; Student's t test.

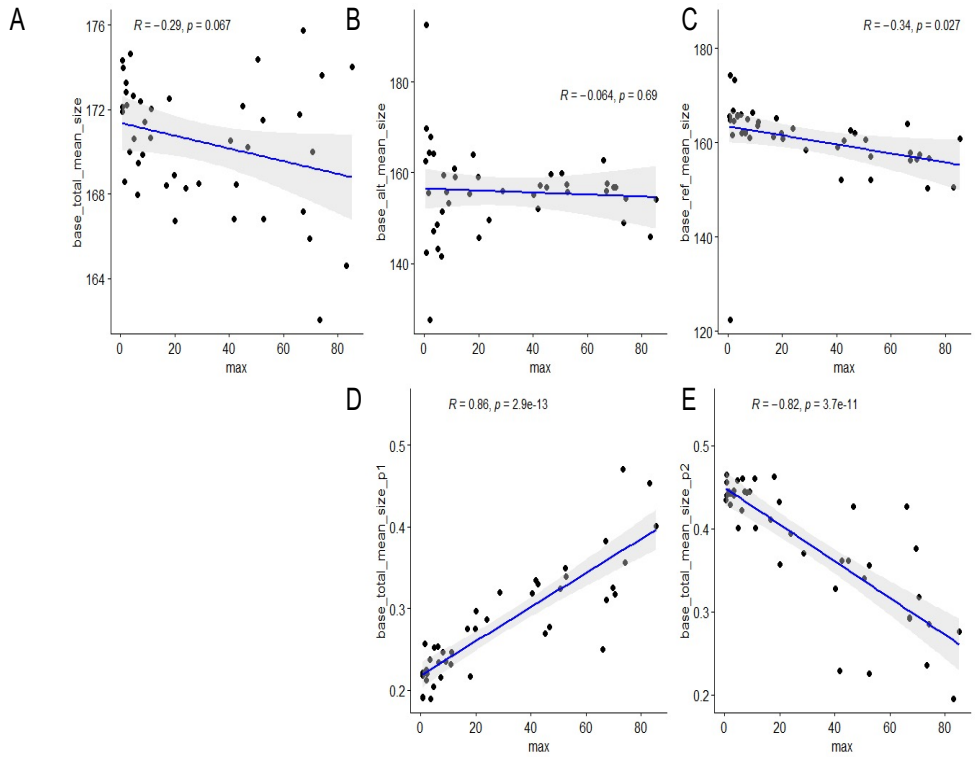

FIGURE 6. Correlation between the maximum VAF and ctDNA fragment size.

Variables were plotted along with maximum VAF. A. Total mean fragment size in baseline samples. B. Mean ALT size in baseline samples. C. Mean REF size in baseline samples. D. AUC for short fragments (100–155 bp). E. AUC for reference fragments (160–180 bp). Pearson coefficients for A, B, C, D, and E were −0.29, −0.064, −0.034, 0.86, and −0.82; P values were 0.067, 0.69, 0.027, 2.9e−13, and 3.7e−11; Pearson correlation analysis.

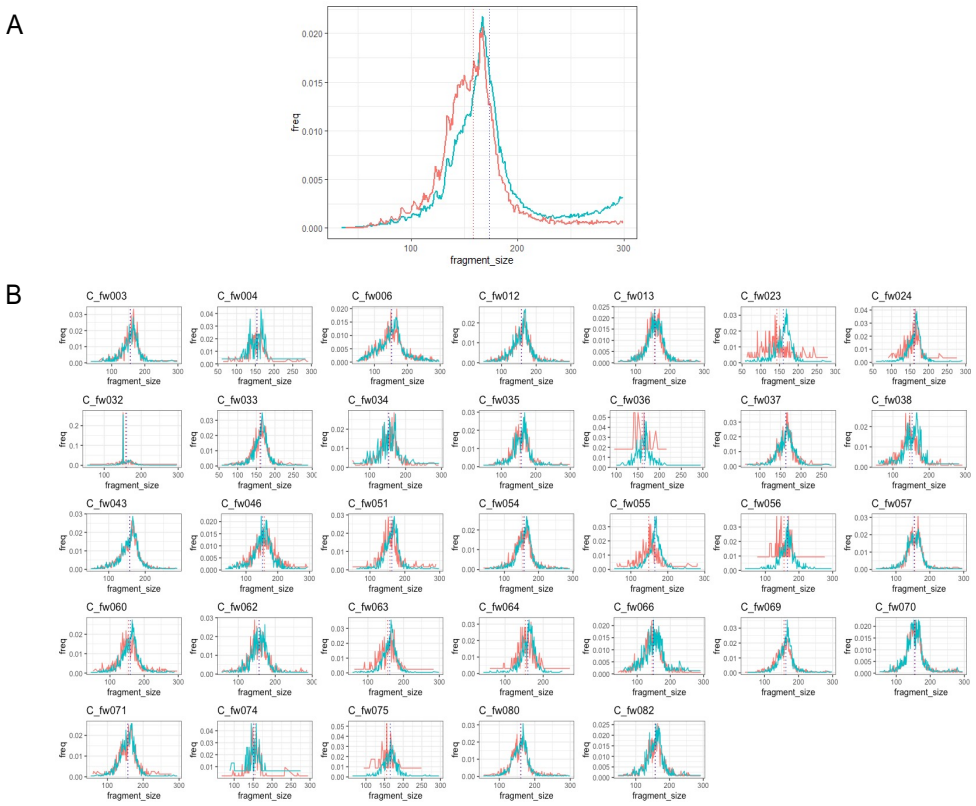

FIGURE 7. Distribution curves for ctDNA fragments from patients with more than 10% somatic mutations detected in plasma (n=33).

A. Fragment sizes with a VAF greater than 10% were plotted. The mean ALT and REF sizes were 158.241 bp and 173.669 bp, respectively. The red and green lines indicate the ALT and REF read fragments, respectively. B. A distribution curve was plotted for each patient, individually.

4. The fragmentation ratio as prognostic marker

Next, I wondered the ctDNA fragment size could predict the clinical response of colorectal cancer patients. For fifty-two out of 62 patients, clinical response was analyzed by RECIST 1.1 (46). Fifty-two patients were separated by two group which are responder (PR; n = 31) and non-responder (PD; n = 1 & SD; n = 20). The median survival was 269 days in non-responder group and 442 days in responder (P = 0.00059; Log-rank test) (Figure 8). The average 4.51 plasma samples of 62 colorectal cancer patients were collected along with the clinical response from baseline to end point. Among these samples, the data of four time points were analyzed (the baseline, first follow up; median 75 days, before end point; median 261 days and end point; median 339.5 days). In order to assess the utility of ctDNA fragment size as prognostic marker, I defined the ratio which is the proportion of the short fragment (P1; 100 – 155 bp) divided by the proportion of the reference cfDNA fragment lengths (P2; 160 – 180 bp) as 'the fragmentation ratio'. The fragmentation ratio and mean size of total read fragment was calculated every time point each patient. Then, the threshold of each variable and each time point was calculated by ROC analysis (Figure 9). Clinical sub-group (responder vs. non-responder) could be separated by the optimal threshold as mentioned. With these thresholds, the group for prognosis analysis were separated by two group. As a result, the high fragmentation ratio group predicted poor prognosis than the low fragmentation ratio group at before end point (P = 0.016; Log-rank test) (Figure 10). The other plasma samples which were collected was not shown significant prediction. Furthermore, the

fragment ratio was increased in non-responder group from first follow up to end point (P = 0.0195; two-way ANOVA test) (Figure 11). In conclusion, with the fragmentation ratio, it showed that monitoring and prognosis for colorectal cancer patients were possible since the specific sampling time point (before end point).

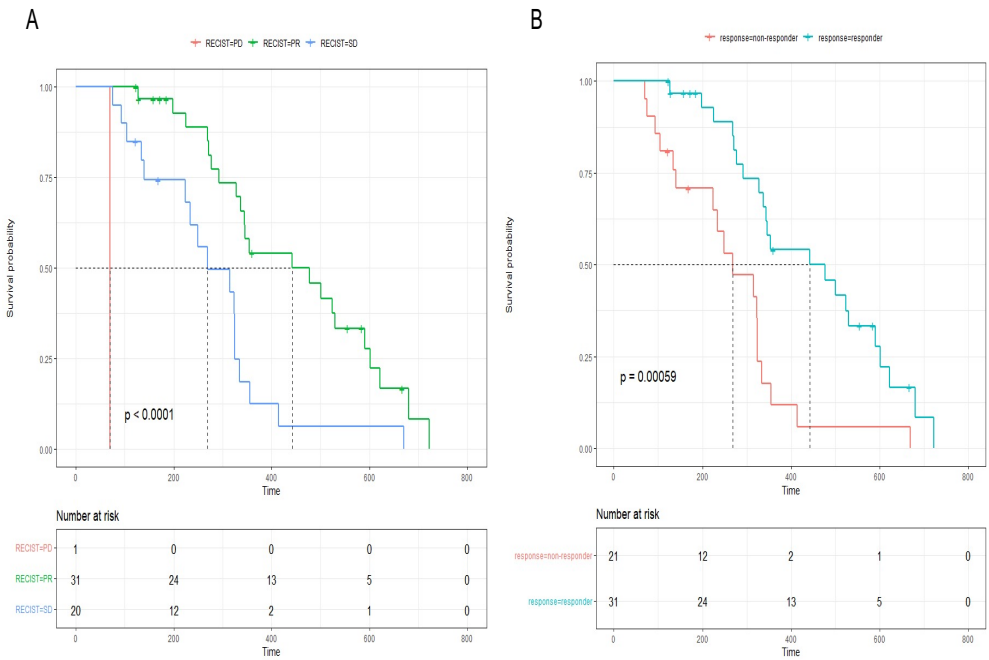

FIGURE 8. Calculation of PFS according to the RECIST 1.1 guideline.

A. A Kaplan-Meier plot was used to analyze median survival time of patients grouped into those with PD, SD, or PR; P < 0.0001; log-rank test. B. A Kaplan-Meier plot was used to analyze median survival time of patients grouped into responder and non-responder groups; P = 0.00059.

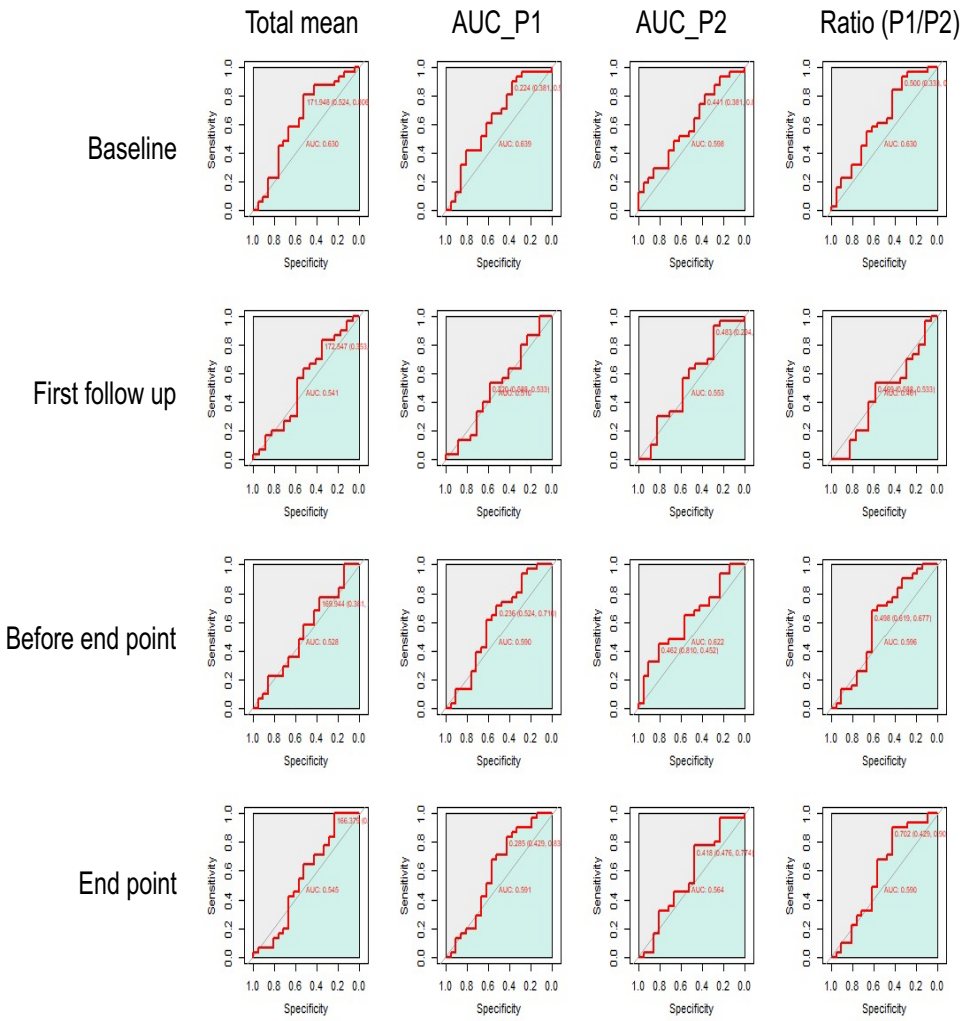

FIGURE 9. ROC analysis for calculating the optimal cutoff values used to classify patients into the responder and non-responder groups.

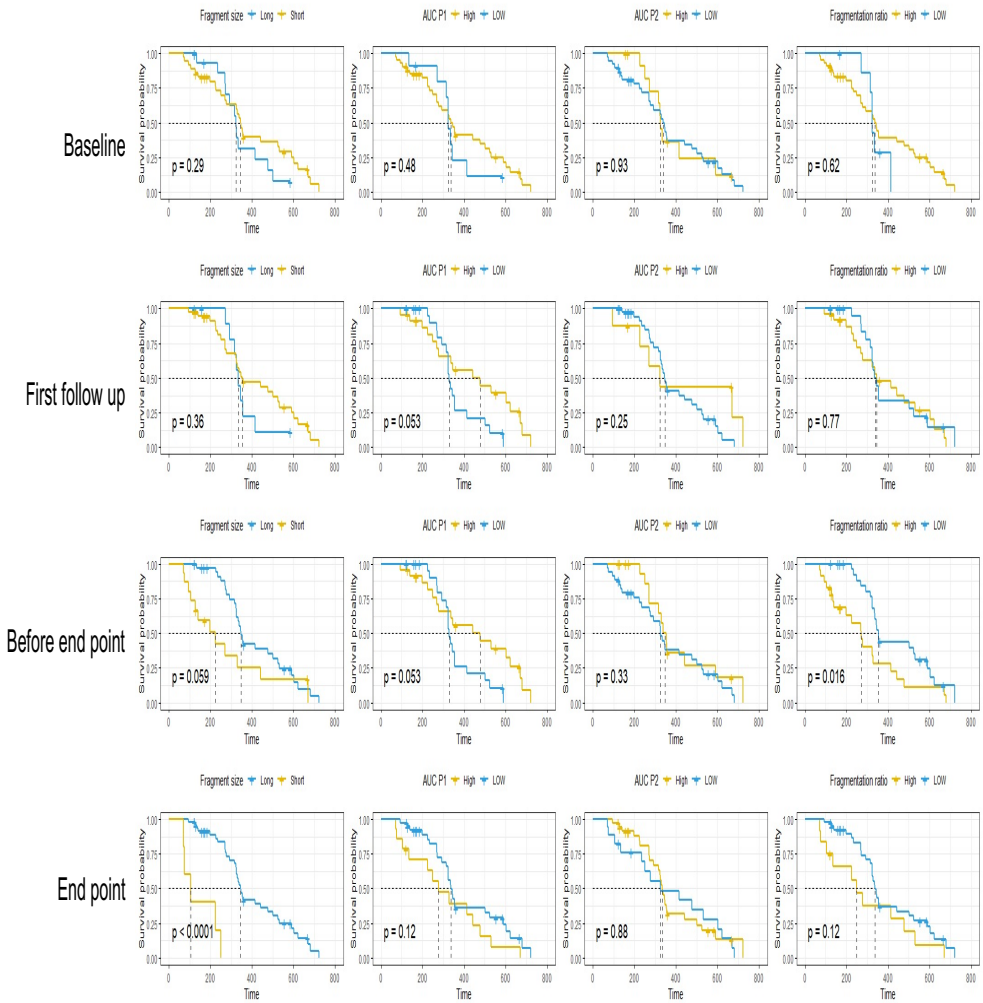

FIGURE 10. Survival plot for each sampling time point and variables.

Kaplan-Meier analysis was performed using optimal cutoff values for clinical responses. A log-rank test was used for analysis.

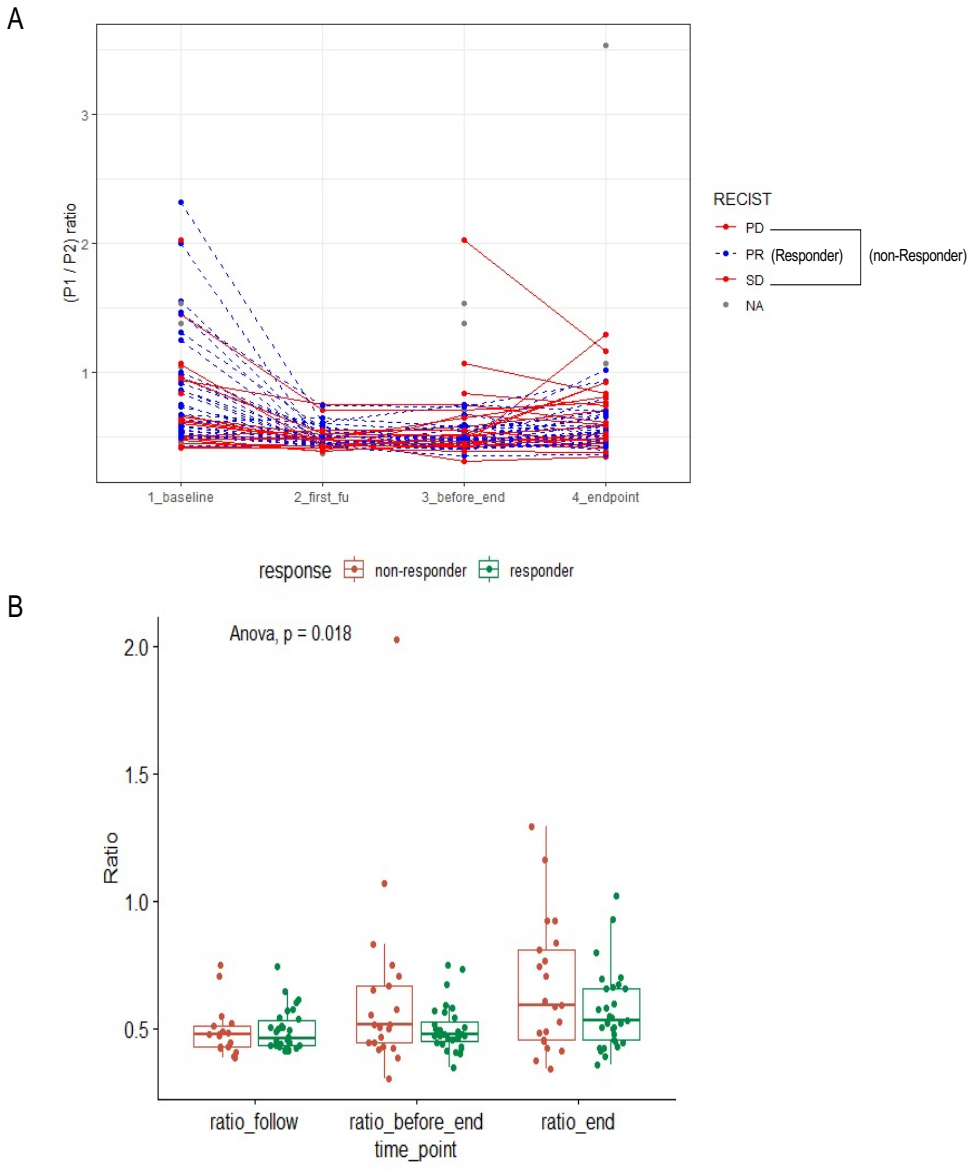

FIGURE 11. Clinical response monitoring using the fragmentation ratio (AUCp1 / AUCp2).

A. Plot of sampling time and clinical response. B. Plot of fragmentation ratio and sampling time; P = 0.0195; ANOVA test.

DISCUSSION

The results of this study show the potential of cfDNA fragment size as a prognostic marker for patients with colorectal cancer. Two hundred eighty plasma samples from 62 patients with colorectal cancer who received targeted therapy and 50 plasma samples from healthy individuals were analyzed using targeted deep sequencing of 106 cancer-related genes, and a comparison of cfDNA fragment sizes was performed. The representative size of cfDNA was reported to be 166 bp, likely as a result of histone packaging. On the other hand, highly fragmentated cfDNA (< 100 bp) is associated with transcription factor binding. To investigate this further, a platform able to detect nucleosome footprinting would be needed, such as MNase-seq, ATAC-seq, or DNase-seq (47) (48). According to previous reports, DNA fragmentation is related to cell death (25), as caspase-activated DNase activity has been shown to be associated with DNA fragmentation (49). Even though patient-derived ctDNA fragments harboring mutant alleles were shorter than those from healthy controls, the molecular mechanism is still unclear.

Because the size of ctDNA fragments was shorter than cfDNA from normal cells, the region used for determining fragment size is important. Clonality is highly correlated with ctDNA fragment size. Concordance between primary tissues and ctDNA was previously calculated to be 93% in patients with colorectal cancer (50). The steps for detecting somatic mutations and ctDNA fragment size thus followed this previous report.

There are two primary approaches for analyzing the size of ctDNA fragments: the use of either a double-stranded DNA (dsDNA) or single-stranded (ssDNA) library. Snyder et al. described a diagnostic model for cancer based on analyzing the size of cfDNA fragments using an ssDNA library prepared via WGS (27). In that paper, the small cfDNA size was associated with transcription factor binding. Even though a dsDNA library was used to analyze cfDNA, the fragment data were sufficient for analysis. A different study compared the use of ssDNA and dsDNA libraries for cfDNA analysis (51), finding that fragments larger than 100 bp could be detected using either library. However, fragments smaller than 100 bp could only be detected using an ssDNA library. For analysis of either genome-wide fragment patterns or nucleosome footprinting, both an ssDNA library and WGS are needed.

In a study monitoring the treatment response of patients with colorectal cancer who received an anti-EGFR therapy (52), patients whose average VAF was less than 1% during the first evaluation had significantly better PFS than those with a higher VAF ($P < 0.001$). In this study, responders and non-responders were categorized, according to the RECIST 1.1 guideline, by sampling time as determined by the ROC analysis. At time points prior to PD, the fragmentation ratio (AUCp1/AUCp2) was confirmed to be a prognostic marker for patients in the colorectal cancer cohort, increasing after treatment in the non-responder group. Thus, the fragmentation ratio can be used as a prognostic marker of treatment success in patients with colorectal cancer. To validate these findings, a larger cohort with various types of cancer should be studied.

CONCLUSION

This is a study on the discovery of non-genetic markers for predicting the prognosis of Korean colorectal cancer for utilizing liquid biopsy.

In part 1, based on machine learning, prediction model with genome-wide DNA methylation markers for diagnosis and risk score for predicting prognosis were devised. Seven-hundred ninety methylation data were processed as COPM dataset (n=709; 330 normal adjacent colon tissues & 379 colorectal cancer tumor tissues). Using machine learning algorithm, the optimized predictive modeling was set and the final 305 probes were defined as diagnostic marker. The performance of my prediction model was 0.997 (training cohort) and 0.976 (validation cohort) AUC, each. With my prediction model, the colorectal cancer patient in TCGA and GEO dataset (n=9,660 tissues and 59 plasma cfDNA) were significantly predicted as the positive for colorectal cancer. A subset of diagnostic markers was chosen to define a risk score for predicting prognosis (OS; 20 & PFS; 133 probes). The prognostic markers were enriched in transcription regulatory regions and some of the probe set were correlated with mRNA expression (41%; 9 / 22).

In part 2, using next-generation sequencing data, the prognostic system with the fragment size of ctDNA was devised. Two-hundred eighty plasma samples from 62 colorectal cancer patients were collected along with 50 plasmas from healthy donors and analyzed by targeted deep sequencing including cancer related 106 genes. The ctDNA fragment size was shorter than cfDNA from non-cancerous cell. The clonality was highly correlated with

the ctDNA fragment size. The responders and non-responders were separated by RECIST 1.1. Using ROC analysis, the optimal cut-off values for various sampling time points, each. At the time points before PD, the ratio of AUC (AUC_{p1}/AUC_{p2}) could be the prognostic markers for colorectal cohort. Assessing the utility of ratio of AUC (AUC_{p1}/AUC_{p2}) at various sampling time point, the ratio was increased after treatment in non-responder group, consecutively.

In conclusion, I focused on validating the utility of non-genetic signatures(methylation markers and fragmentomics) in Liquid biopsy.

REFERENCES

I. Use of an optimized machine learning algorithm to discover DNA methylation markers from Korean colorectal cancer patients

1. Dekker E, Tanis PJ, Vleugels JLA, Kasi PM, Wallace MB. Colorectal cancer. Lancet. 2019;394(10207):1467-80.

2. Keum N, Giovannucci E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. Nat Rev Gastroenterol Hepatol. 2019;16(12):713-32.

3. 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(3):209-49.

4. Agrawal K, Das V, Vyas P, Hajduch M. Nucleosidic DNA demethylating epigenetic drugs - A comprehensive review from discovery to clinic. Pharmacol Ther. 2018;188:45-79.

5. Erratum for the Research Article: "Circulating tumor DNA methylation profiles enable early diagnosis, prognosis prediction, and screening for colorectal cancer" by H. Luo, Q. Zhao, W. Wei, L. Zheng, S. Yi, G. Li, W. Wang, H. Sheng, H. Pu, H. Mo, Z. Zuo, Z. Liu, C. Li, C. Xie, Z. Zeng, W. Li, X. Hao, Y. Liu, S. Cao, W. Liu, S. Gibson, K. Zhang, G. Xu, R.-h. Xu. Sci Transl Med. 2020;12(540).

6. Liu MC, Oxnard GR, Klein EA, Swanton C, Seiden MV,

Consortium C. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. Ann Oncol. 2020;31(6):745-59.

7. Galanter JM, Gignoux CR, Oh SS, Torgerson D, Pino-Yanes M, Thakur N, et al. Differential methylation between ethnic sub-groups reflects the effect of genetic ancestry and environmental exposures. Elife. 2017;6.

8. Heyn H, Esteller M. DNA methylation profiling in the clinic: applications and challenges. Nat Rev Genet. 2012;13(10):679-92.

9. Wolf J, Willscher E, Loeffler-Wirth H, Schmidt M, Flemming G, Zurek M, et al. Deciphering the Transcriptomic Heterogeneity of Duodenal Coeliac Disease Biopsies. Int J Mol Sci. 2021;22(5).

10. Fortin JP, Triche TJ, Jr., Hansen KD. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. Bioinformatics. 2017;33(4):558-60.

11. Heitzer E, Haque IS, Roberts CES, Speicher MR. Current and future perspectives of liquid biopsies in genomics-driven oncology. Nat Rev Genet. 2019;20(2):71-88.

12. Saghafinia S, Mina M, Riggi N, Hanahan D, Ciriello G. Pan-Cancer Landscape of Aberrant DNA Methylation across Human Tumors. Cell Rep. 2018;25(4):1066-80 e8.

13. Flanagan JM. Epigenome-wide association studies (EWAS): past, present, and future. Methods Mol Biol. 2015;1238:51-63.

14. Ernst J, Kellis M. ChromHMM: automating chromatin-state

discovery and characterization. Nat Methods. 2012;9(3):215-6.

15. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38(4):576-89.

16. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun. 2019;10(1):1523.

17. Kulis M, Esteller M. DNA methylation and cancer. Adv Genet. 2010;70:27-56.

18. Modhukur V, Sharma S, Mondal M, Lawarde A, Kask K, Sharma R, et al. Machine Learning Approaches to Classify Primary and Metastatic Cancers Using Tissue of Origin-Based DNA Methylation Profiles. Cancers (Basel). 2021;13(15).

19. Holmes EE, Jung M, Meller S, Leisse A, Sailer V, Zech J, et al. Performance evaluation of kits for bisulfite-conversion of DNA from tissues, cell lines, FFPE tissues, aspirates, lavages, effusions, plasma, serum, and urine. PLoS One. 2014;9(4):e93933.

20. Olova N, Krueger F, Andrews S, Oxley D, Berrens RV, Branco MR, et al. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. Genome Biol. 2018;19(1):33.

21. Shen SY, Singhania R, Fehringer G, Chakravarthy A, Roehrl

MHA, Chadwick D, et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. Nature. 2018;563(7732):579-83.

22. Liu Y, Siejka-Zielinska P, Velikova G, Bi Y, Yuan F, Tomkova M, et al. Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. Nat Biotechnol. 2019;37(4):424-9.

23. Fatemi M, Pao MM, Jeong S, Gal-Yam EN, Egger G, Weisenberger DJ, et al. Footprinting of mammalian promoters: use of a CpG DNA methyltransferase revealing nucleosome positions at a single molecule level. Nucleic Acids Res. 2005;33(20):e176.

II. Combined analysis of ctDNA mutation and fragment size for predicting prognosis of colorectal cancer

24. Wan JCM, Massie C, Garcia-Corbacho J, Mouliere F, Brenton JD, Caldas C, et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. Nat Rev Cancer. 2017;17(4):223-38.

25. Rostami A, Lambie M, Yu CW, Stambolic V, Waldron JN, Bratman SV. Senescence, Necrosis, and Apoptosis Govern Circulating Cell-free DNA Release Kinetics. Cell Rep. 2020;31(13):107830.

26. Vierstra J, Wang H, John S, Sandstrom R,

Stamatoyannopoulos JA. Coupling transcription factor occupancy to nucleosome architecture with DNase-FLASH. Nat Methods. 2014;11(1):66-72.

27. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. Cell. 2016;164(1-2):57-68.

28. Corcoran RB, Chabner BA. Application of Cell-free DNA Analysis to Cancer Treatment. N Engl J Med. 2018;379(18):1754-65.

29. Shi J, Zhang R, Li J, Zhang R. Size profile of cell-free DNA: A beacon guiding the practice and innovation of clinical testing. Theranostics. 2020;10(11):4737-48.

30. Mouliere F, Chandrananda D, Piskorz AM, Moore EK, Morris J, Ahlborn LB, et al. Enhanced detection of circulating tumor DNA by fragment size analysis. Sci Transl Med. 2018;10(466).

31. Chabon JJ, Hamilton EG, Kurtz DM, Esfahani MS, Moding EJ, Stehr H, et al. Integrating genomic features for non-invasive early lung cancer detection. Nature. 2020;580(7802):245-51.

32. Cristiano S, Leal A, Phallen J, Fiksel J, Adleff V, Bruhm DC, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature. 2019;570(7761):385-9.

33. Zhu G, Guo YA, Ho D, Poon P, Poh ZW, Wong PM, et al. Tissue-specific cell-free DNA degradation quantifies circulating tumor DNA burden. Nat Commun. 2021;12(1):2229.

34. Serpas L, Chan RWY, Jiang P, Ni M, Sun K, Rashidfarrokhi A,

et al. Dnase1l3 deletion causes aberrations in length and end-motif frequencies in plasma DNA. Proc Natl Acad Sci U S A. 2019;116(2):641-9.

35. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.

36. Chen S, Zhou Y, Chen Y, Huang T, Liao W, Xu Y, et al. Gencore: an efficient tool to generate consensus reads for error suppressing and duplicate removing of NGS data. BMC Bioinformatics. 2019;20(Suppl 23):606.

37. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016;44(11):e108.

38. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6(2):80-92.

39. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, et al. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front Genet. 2012;3:35.

40. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol.

2016;17(1):122.

41. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.

42. Linden A. Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. J Eval Clin Pract. 2006;12(2):132-9.

43. Guo J, Ma K, Bao H, Ma X, Xu Y, Wu X, et al. Quantitative characterization of tumor cell-free DNA shortening. BMC Genomics. 2020;21(1):473.

44. Razavi P, Li BT, Brown DN, Jung B, Hubbell E, Shen R, et al. High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants. Nat Med. 2019;25(12):1928-37.

45. Graham TA, Sottoriva A. Measuring cancer evolution from the genome. J Pathol. 2017;241(2):183-91.

46. Schwartz LH, Litiere S, de Vries E, Ford R, Gwyther S, Mandrekar S, et al. RECIST 1.1-Update and clarification: From the RECIST committee. Eur J Cancer. 2016;62:132-7.

47. Chereji RV, Bryson TD, Henikoff S. Quantitative MNase-seq accurately maps nucleosome occupancy levels. Genome Biol. 2019;20(1):198.

48. Ulz P, Perakis S, Zhou Q, Moser T, Belic J, Lazzeri I, et al. Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. Nat Commun.

2019;10(1):4666.

49. Lu H, Hou Q, Zhao T, Zhang H, Zhang Q, Wu L, et al. Granzyme M directly cleaves inhibitor of caspase-activated DNase (CAD) to unleash CAD leading to DNA fragmentation. J Immunol. 2006;177(2):1171-8.

50. Kang JK, Heo S, Kim HP, Song SH, Yun H, Han SW, et al. Liquid biopsy-based tumor profiling for metastatic colorectal cancer patients with ultra-deep targeted sequencing. PLoS One. 2020;15(5):e0232754.

51. Sanchez C, Roch B, Mazard T, Blache P, Dache ZAA, Pastor B, et al. Circulating nuclear DNA structural features, origins, and complete size profile revealed by fragmentomics. JCI Insight. 2021;6(7).

52. Lim Y, Kim S, Kang JK, Kim HP, Jang H, Han H, et al. Circulating tumor DNA sequencing in colorectal cancer patients treated with first-line chemotherapy with anti-EGFR. Sci Rep. 2021;11(1):16333.

국문 초록

암을 진단하고 모니터링하고 예후를 예측하는 것에 있어서 액체생검은 매우 중요한 한가지 방법으로써 주목받고 있다. 특히나 새로운 마커로써 비유전적 시그니처 들은 더욱 대두되고 있다. 그러한 이유는 암환자의 혈액종양DNA는 다른 어떠한 마커보다 종합적으로 신체를 반영하고 있고, 원발암을 대표하는데 있어서 많은 정보를 갖는다 것에 있다. 이러한 혈액종양DNA는 유전적 마커뿐만 아니라, 비유전적 마커 즉, DNA 메틸레이션 or DNA 프래그먼트 크기 등 다양한 분자적 특성들을 반영한다. DNA 메틸레이션 은 조직에 대한 특이한 패턴을 갖고 있으며, DNA 프래그먼트 크기에 대한 특이성은 무세포핵산 자체의 특징 중 하나고, 이를 활용하려는 노력들이 많아지고 있다. 이러한 특성을 포괄적으로 활용하기 위하여, 통합적인 분석이 필요하고 새로운 마커의 발굴이 필요하다.

본 논문에서는 1) 기존에 DNA 메틸레이션 은 많이 보고 되어있지만, 단일마커 그리고 서양인들 중심으로 보고가 되어왔다. 하지만, 메틸레이션 패턴은 인종간의 차이도 어느정도 있고, 조직의 특이성을 반영하기 위해서는 단일마커보다는 다양한 마커를 활용하여 예측력을 높이는 것이 중요하다. 따라서 나는 709개의 한국인 대장암 조직을 이용하여 얻은 메틸레이션 데이터를 이용하여 머신러닝 기반 305개 마커를 활용하는 진단 예측 모델을 구축하였다. 구축한 모델은 조직 데이터뿐 만아니라 혈장 무세포핵산 메틸레이션 데이터에서도 또한 높은 예측력을 보였으며, 마커의 서브셋을 이용한 예후 예측도 또한 가능하였다.

다음으로 2) 무세포핵산의 프래그먼트 크기는 무세포핵산 만이 갖는 분자적 특성이다. 최근에 암환자에서 유래한 무세포핵산의 크기는 체성변이에서 특이적으로 사이즈 차이가 난다는 점을 이용하는 연구들이 주되었다. 유전체 전체를 이용하여 암 특이적 진단 마커를 발굴하는 내용 그리고 패널 시퀀싱을 이용하여 특정 변이들에서 크기의 차이를 이용하여 변이의 검출확률을 높이는 방법등이 대표적인 예이다. 하지만 진단 이외의 활용측면에서는 아직 연구할 부분이 많다. 이러한 간극을 매꾸기 위하여 혈액종양DNA의 프래그먼트 크기 분석을 진행하였다. 우리는 paired end 시퀀싱 기반의 패널 시퀀싱 데이터를 활용하여 핵산 분자의 실제 크기를 계산하였고, 이러한 크기가 원발암 유래에 의함이

라는 것을 데이터상으로 증명했다. 나아가, 한환자로부터 유래한 다양한 치료 전/후 대장암 혈액 샘플에서 특정 시점에서 크기를 활용한 마커가 예후 예측에 통계적으로 유의미한 파워를 갖는 것을 확인하였다.

주 요 어 : 대장암, 후성유전체, 액체생점, 혈액 종양 DNA, 예후 예측, 차세대 염기서열분석
학 번 : 2018-37966