creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

연세대학교
YONSEI UNIVERSITY

Development of a multi-modal immunotherapy response predictor for colorectal cancer based on clinical, genomic, imaging, and pathological data

Hong, Jiyun

Department of Medical Science

Graduate School

Yonsei University

연세대학교
YONSEI UNIVERSITY

Development of a multi-modal immunotherapy response predictor for colorectal cancer based on clinical, genomic, imaging, and pathological data

Advisor Kim, Sangwoo

A Dissertation Submitted
to the Department of Medical Science
and the Committee on Graduate School
of Yonsei University in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy in Medical Science

Hong, Jiyun

June 2025

연세대학교
YONSEI UNIVERSITY

Development of a multi-modal immunotherapy response predictor for colorectal cancer based on clinical, genomic, imaging, and pathological data

This Certifies that the Dissertation of Hong, Jiyun is Approved

Committee Chair	Kim, Han Sang
Committee Member	Kim, Sangwoo
Committee Member	Kim, Jin Sung
Committee Member	Kim, Hyun Seok
Committee Member	Jung, Minsun

Department of Medical Science
Graduate School
Yonsei University
June 2025

연세대학교
YONSEI UNIVERSITY

ACKNOWLEDGEMENTS

First and foremost, my heartfelt thanks go to my advisor, Professor Sangwoo Kim. His unwavering support and insightful guidance have been instrumental in the successful completion of my research. The six years I spent at TGIL, including my internship, have provided invaluable experiences that I will cherish for a lifetime.

I extend my sincere gratitude to all the professors who offered valuable insights during the thesis review process. In particular, I would like to thank Professor Hansang Kim for his passionate mentorship, Professor Jinsung Kim for his invaluable advice on imaging data, Professor Hyunseok Kim for his meticulous attention to detail, and Professor Minsun Jung for her generous assistance with pathology data.

I am deeply grateful to my fellow researchers and mentors at TGIL for fostering a collaborative and supportive environment. The friendships and shared experiences we built together will always remain close to my heart. Special thanks to my senior colleagues Mikyung, Hyeonduk, and Beumjin for helping me settle into the lab, and to my junior colleagues Eunwoo and Chaejoo for the joy of working together on various projects. I would also like to warmly thank Jeongsoo, Seunghee, and Seungseok for the lively discussions and memorable lab moments. Your camaraderie made my time here truly meaningful.

연세대학교
YONSEI UNIVERSITY

I sincerely thank the Heo Ji-young Cancer Scholarship Foundation for their generous support, which enabled me to continue my studies and pursue my aspirations. I am truly honored and committed to making the most of this opportunity.

Lastly, I am profoundly grateful to my family for their unwavering support and belief in me. Their love and sacrifices have been the bedrock of my journey. A very special thanks to my best friend and fiancée, Eunyoung Choi, for her boundless love, encouragement, and steadfast support throughout my Ph.D. journey. Her presence has made this experience truly extraordinary.

연세대학교
YONSEI UNIVERSITY

TABLE OF CONTENTS

연세대학교
YONSEI UNIVERSITY

연세대학교
YONSEI UNIVERSITY

연세대학교
YONSEI UNIVERSITY

LIST OF FIGURES

연세대학교
YONSEI UNIVERSITY

연세대학교
YONSEI UNIVERSITY

LIST OF TABLES

연세대학교
YONSEI UNIVERSITY

ABSTRACT

Development of a multi-modal immunotherapy response predictor for colorectal cancer based on clinical, genomic, imaging, and pathological data

Colorectal cancer (CRC) is the third most prevalent cancer globally, accounting for approximately 10% of all cancer cases and serving as the second leading cause of cancer-related deaths. While mismatch repair-deficient and microsatellite instability-high CRCs have shown favorable responses to immunotherapy, particularly with FDA-approved immune checkpoint inhibitors (ICIs), significant unmet medical needs remain, especially for mismatch repair-proficient and microsatellite instability-low CRCs that exhibit resistance to current treatments. Although therapeutic options exist for advanced-stage patients, there is an urgent need for more effective and systematic approaches. To address these needs, the identification of suitable biomarkers for drug selection and the prediction of individual patient responses is crucial. Recent efforts have focused on integrating diverse data types—including genomic, clinical, digital imaging, and digital pathology—to enhance predictive accuracy. This study aims to develop a multimodal immunotherapy response predictor for CRC that incorporates whole exome sequencing (WES), RNA sequencing, metagenomic sequencing, clinical data, pathology, and radiology images. We collected samples from 106 patients at Severance Hospital, including 36 responders and 70 non-responders. Our findings indicate that features derived from individual modalities are insufficient for accurately distinguishing between responders and non-responders. However, integrating these modalities produces a synergistic effect that significantly enhances classification accuracy. Our machine learning-based multimodal models, utilizing techniques such as Random Forest, Logistic Regression, Naïve Bayes, and XGBoost, demonstrated improved performance with the incorporation of additional modalities, leading to substantial enhancements in accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Notably, the tool we developed offers a significant advantage by providing response prediction scores even in cases of missing modalities or missing features within a modality. This

연세대학교
YONSEI UNIVERSITY

robustness enhances its applicability in clinical settings where complete data may not always be available. We validated our model's generalization performance through stratified k-fold cross-validation, an independent external validation set, and comparisons with single-modality immune checkpoint inhibitor response predictors. This research represents a pioneering effort to develop a multimodal predictive model for immune checkpoint inhibitor responses in colorectal cancer. Furthermore, we demonstrate that predictive performance improves incrementally with the combination of modalities, offering the potential to enhance treatment strategies and patient outcomes.

Key words: colorectal cancer, immunotherapy response predictor, predictive biomarker, multi-omics, multimodal integration

연세대학교
YONSEI UNIVERSITY

1. INTRODUCTION

Colorectal cancer (CRC) is referred cancers arise from the colon and rectum, making up the large bowel. CRC accounts for about the 10% of cancer types and second deleterious cancer worldwide[1]. Although death and incidence rate are declined over the past 10 years due to regular diagnosis and screening test, they are not still favorable cancer about the advanced stage patients, accounting only 40% of CRCs are early stage and survival rate is highly beneficial over 90%[2]. Available treatment options for CRC include chemotherapy, radiation therapy, surgery, and immunotherapy. Among them, immunotherapy is type of treatment which manipulating own tumor immune microenvironments to reduce cancer cells. Specifically, there are various FDA-approved treatments for the high microsatellite instability (MSI-H), and deficient DNA mismatch repair (dMMR), including monoclonal antibody that targets VEGF, EGFR pathway, and immune checkpoint inhibitors (ICIs)[2]. Many promise immunotherapies are ongoing clinical test for the CRC, some limitations are not still resolved. In the clinical field, several unmet needs are evident. Currently, metastatic CRCs with proficient DNA mismatch repair (pMMR) or microsatellite stability (MSS) types, which are typically considered immune-cold, do not respond to immunotherapy agents[3]. While there have been some improvements observed in MSI-H cases through several studies, the response rate has remained within the range of 30% to 50%, indicating the necessity for more promising therapeutics[4]. Therefore, predicting treatment responses and recommending the best actionable drugs based on individual patient characteristics could prove invaluable in prescribing optimal therapeutics. Recent advances in machine learning approaches have led to studies focusing on the development of drug response prediction models using various biomarkers in different cancer types [5-8]. Moreover, in contrast to previous studies, multimodal prediction models that integrate radiology, pathology, and various genomics data have shown higher precision and recall scores[9-10].

To address these challenges, we have already gathered multi-institutional and public datasets that include perfectly or partially combined multimodal data, encompassing genomics, clinical information, images, and pathological digital data. We aim to construct a prediction model and validate it using independent datasets to enhance confidence in our approach. In summary, our pursuit of a multimodal immunotherapy response predictor for CRC is driven by the imperative need to address the unmet needs in clinical practice. We aspire to provide oncologists with a powerful

연세대학교
YONSEI UNIVERSITY

tool that not only identifies the most sensitive biomarkers for diverse treatments but also recommends the optimal drug tailored to individual patient profiles. Through the integration of cutting-edge machine learning techniques and extensive multimodal datasets, we aim to make significant strides in enhancing the efficacy of CRC treatment strategies, ultimately advancing the prospects for improved patient outcomes.

연세대학교
YONSEI UNIVERSITY

2. MATERIALS AND METHODS

2.1. Clinical cohorts and tissue collection

This research was conducted in compliance with ethical standards and received approval from the Severance Hospital Institutional Review Board (IRB 2019-1690-036), ensuring adherence to the Declaration of Helsinki throughout the study process. A total of 106 patients were enrolled in this study, and their characteristics are summarized in Table 1. The samples were collected based on clinical data from patients who received immune checkpoint inhibitors at Severance Hospital. The study aimed to analyze DNA, RNA, and gut microbial profiles from formalin-fixed paraffin-embedded (FFPE) tissue samples. According to the Response Evaluation Criteria in Solid Tumors (RESIST), patients were categorized into immune checkpoint inhibitor responders (complete response [CR], partial response [PR], or progression-free survival [PFS] of 180 days or more) and non-responders (progressive disease [PD]) based on their best response. The immune checkpoint inhibitors administered to the participants included three types: Durvalumab (anti-PD-L1), and two anti-PD-1 agents, Nivolumab and Pembrolizumab.

연세대학교
YONSEI UNIVERSITY

Table 1. Patient characteristics of severance CRC cohort

Clinical features	***n* = 106**
Age.	
mean±sd	55±14
Range	21-85
Best overall response.	
CR/PR	22
SD/PD	74
NE	10
Progression free survival: months	
mean±sd	7.5±11.2
Range	0.1-58.2
Overall survival: months	
mean±sd	18.7±19.3
Range	0.1-82.6
Sex: n (%).	
M	63 (59.4%)
F	43 (40.6%)
Stage: n (%).	
I, II, III	57 (53.8%)
IV	48 (45.3%)
unknown	1 (0.9%)
Sidedness: n (%).	
Right	64 (60.4%)
Left	41 (38.7%)
unknown	1 (0.9%)
Microsatellite: n (%).	
MSI-H	45 (42.5%)
MSS	42 (39.6%)
unknown	19 (17.9%)
Drug: n (%).	
pembrolizumab	89 (84.0%)
durvalumab	7 (6.6%)
nivolumab	6 (5.7%)
nivolumab + ipilimumab	4 (3.7%)

연세대학교
YONSEI UNIVERSITY

2.2. Study design

We classified 106 patients with colorectal cancer who received immune checkpoint inhibitors at Severance Hospital based on their clinical information and the best response according to RECIST criteria. From their FFPE tissue samples, we collected Whole Exome Sequencing (WES) data for 52 patients, RNA sequencing data for 53 patients, and metagenomic sequencing (MGS) data for 57 patients. Additionally, we obtained pathology and radiology information for all 106 patients, as illustrated in Figure 1. In our discovery set, there are 36 responders and 70 non-responders, among whom 7 received the PD-L1 inhibitor durvalumab, 89 received the PD-1 inhibitor pembrolizumab, 10 received nivolumab, and 4 received a combination of nivolumab and the CTLA-4 inhibitor ipilimumab. Our results were validated using a total of two independent public external sets with immune checkpoint inhibitor response information: the first is the IMMUNOMSI dataset, which includes 71 patients (67 responders and 4 non-responders), and the second is the NIPICOL dataset, which consists of 47 patients (34 responders and 13 non-responders). The internal dataset utilized clinical, radiologic (image), and pathologic information for all 106 patients, as well as DNA modality data for 52 patients, RNA modality data for 53 patients, and MGS modality data for 57 patients. Among these, 51 patients had complete information across all six modalities. The IMMUNOMSI dataset employed clinical information, DNA, RNA, and pathologic modalities, while NIPICOL utilized clinical information, DNA, and pathologic modality data (Figure 1).

We developed a model based on clinical information (C), Whole Exome Sequencing (WES) (D), RNA sequencing (R), metagenomic sequencing (MGS) (M), pathologic information (P), and imaging information (I) (Figure 2). In the internal dataset, we split the data into training and testing sets in an 80:20 ratio and performed stratified k-fold cross-validation with k=5 to address the issue of data imbalance. The ratio of responders (R) to non-responders (NR) in each training and testing set was set to 3:7. Feature selection was conducted within the training set based on Random Forest-based importance. Using the selected features from each modality, we constructed ensemble model objects for logistic regression (LR), random forest (RF), XGBoost (XG), and naive Bayes (NB) for each modality, and calculated the immune checkpoint inhibitor response score using ensemble techniques based on combinations of modalities.

연세대학교
YONSEI UNIVERSITY

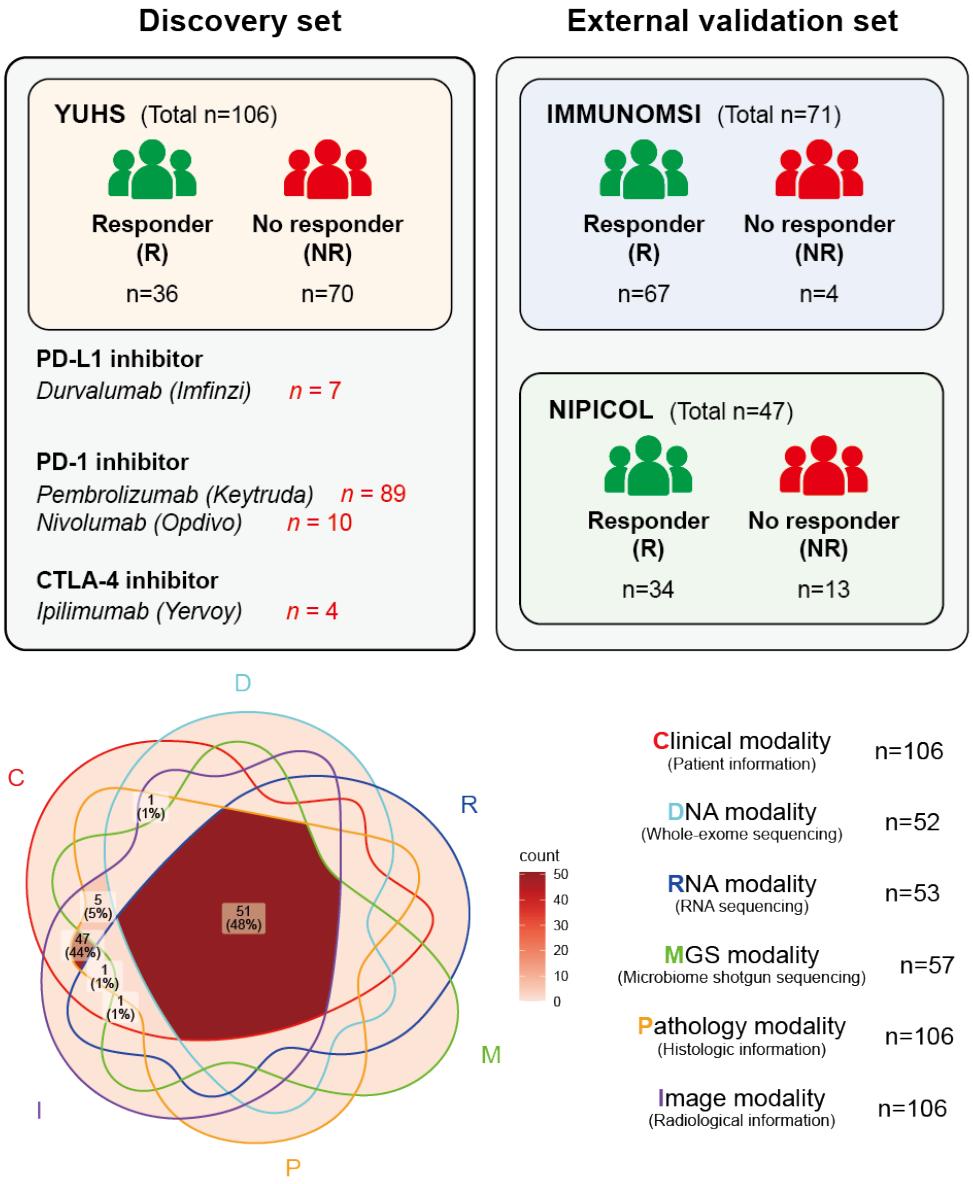

Figure 1. Overview of patient cohorts and data modalities in the study. This figure presents the demographics of the 106 colorectal cancer patients treated with immune checkpoint inhibitors, including 36 responders and 70 non-responders. It outlines the data collected from FFPE tissue samples: Whole Exome Sequencing (WES) for 52 patients, RNA sequencing for 53 patients, and metagenomic sequencing (MGS) for 57 patients. Additionally, it summarizes available pathology and radiology information, as well as the two independent validation datasets: IMMUNOMSI (n=71) and NIPICOL (n=47).

연세대학교
YONSEI UNIVERSITY

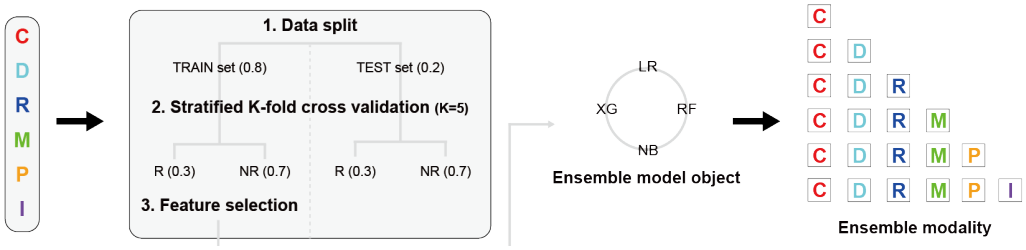

Figure 2. Workflow for developing an ensemble model using multi-modal data. This figure illustrates the workflow for model development utilizing clinical, genomic, pathologic, and imaging data. The data was split into training and testing sets, followed by stratified k-fold cross-validation and feature selection using Random Forest importance. Ensemble models were constructed for each modality to calculate the immune checkpoint inhibitor response score.

2.3. Library preparation

Fresh frozen tissues were collected for RNA sequencing library preparation following the SureSelect RNA Direct Human Sample Preparation protocol. Concurrently, tumor samples underwent paired-end whole exome sequencing (WES) using the SureSelect V6 Post-FFPE kit, which targets 60 Mb of the human exome to provide a comprehensive analysis of genetic variations. To enhance this analysis and investigate the associated microbial profiles, metagenomic shotgun sequencing was performed with the TruSeq Nano DNA 350 META kit. The quality and integrity of the libraries were assessed using the Agilent Technologies 2100 Bioanalyzer to ensure optimal sequencing performance. Finally, sequencing was conducted on the Illumina NovaSeq X platform.

2.4. Bioinformatical pipeline

2.4.1. Clinical feature analysis

To assess the impact of clinical variables on the probability of being classified as a responder compared to a non-responder, we performed a logistic regression analysis utilizing the generalized linear model (GLM) framework[11]. The outcome variable for our analysis was a binary indicator of response status, where responders were coded as 1 and non-responders as 0. The independent variables included sex, age group (mean age), disease stage, type of drug administered, laterality (sidedness), and microsatellite instability (MSI) status. The model summary yields coefficients that reflect the log odds of being classified as a responder in relation to each clinical variable. When exponentiated, these coefficients represent odds ratios, providing valuable insights into how the

연세대학교
YONSEI UNIVERSITY

likelihood of treatment response is affected by each variable.

2.4.2. Data preprocessing and alignment

To initiate our analysis, we performed a thorough quality assessment and preprocessing of all sequencing reads using fastp (v0.21.0)[12], ensuring that only high-quality reads were retained for further investigation. The filtered reads were aligned to the human reference genome (GRCh38), utilizing BWA-MEM (v0.7.17) for DNA sequencing[13] and STAR (v2.7.3a) in a two-pass mode for RNA sequencing[14]. For genomic variant calling, we corrected base quality scores using GATK's BaseRecalibrator and ApplyBQSR modules[15]. For our metagenome sequencing analysis, we commenced by aligning reads with Bowtie2 (v2.3.5.1)[16], followed by the removal of host reads. The remaining unmapped reads were utilized for downstream analysis.

2.4.3. Genomics feature analysis

Somatic variants were filtered to eliminate artifacts through Mutect2, which compared matched normal DNA samples against public variant databases. The genomic variants obtained were then annotated with the Ensembl Variant Effect Predictor (VEP)[17] to elucidate their biological implications and converted to MAF format using vcf2maf (v1.6.20). Our detailed analysis focused on non-synonymous genomic variants, where we calculated the Tumor Mutation Burden (TMB) as the count of non-synonymous somatic mutations per megabase (Mb).

2.4.4. Transcriptomics feature analysis

In our transcriptomic analysis, we quantified gene expression through read counts using HTSeq (v0.11.1)[18]. Lowly expressed genes were filtered out to mitigate bias, and batch effects were adjusted using Combat-seq[19]. The resulting normalized gene expression matrix allowed for the identification of differentially expressed genes (DEGs) through the DESeq2 R package (v1.26.0)[20], applying the following selection criteria: (1) adjusted p-value < 0.05 and (2) absolute Log2 fold change > 2. Moreover, the normalized expression data enabled us to classify consensus molecular subtypes (CMS) of colorectal cancer using the nearest-centroid single-sample predictor algorithm from the CMSclassifier R package (v1.0.0)[21].

2.4.5. Metagenomics feature analysis

Taxonomic profiling was performed against a standard database using Kraken2 (v2.1.1)[22]

연세대학교
YONSEI UNIVERSITY

and Bracken (v2.9)[23], successfully classifying 2,329 genera, which were subsequently refined to a total of 1,590 genera for analysis. We constructed a genus-by-sample matrix through Counts Per Million (CPM) normalization and employed the LEfSe method from the microbiomeMarker R package (v1.8.0)[24] to conduct differential abundance analyses. Additionally, we assessed microbial diversity by calculating alpha and beta diversity metrics, including Shannon, Simpson, and inverse Simpson indices.

2.4.6. Machine learning techniques

To integrate modalities, we first conducted data curation for each modality. We removed NA values, and for continuous values that were not applicable or categorical values with insufficient sample sizes, we applied one-hot encoding. To prevent overfitting, we limited the features used in the individual analyses of each modality to a maximum of two. During the feature extraction process, we utilized filter methods (correlation, variance), wrapper methods (stepwise selection), and embedded methods (elastic net). Using the selected features, we developed a multi-modality-based response predictive model employing three algorithms: logistic regression, XGBoost, and random forest. For data splitting, we allocated the dataset into training and testing sets in a 70:30 ratio, ensuring that each subset contained a minimum number of responder and non-responder samples. The test set maintained the distribution of raw data, and if the training set displayed data imbalance, we performed up-sampling. Our model's performance was evaluated using 5-fold cross-validation.

2.4.7. Statistical analysis

We used statistical tests to assess differences between groups and determine significance. In this study, comparisons between groups were performed using the Wilcoxon Rank Sum test. For multiple group comparisons, we applied the Kruskal-Wallis test. Additionally, we employed Fisher's exact test for evaluating the association between categorical variables. To manage multiple comparisons, we implemented corrections using both the False Discovery Rate (FDR) methods. The significance of differences between groups was determined based on the following thresholds: *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001.

연세대학교
YONSEI UNIVERSITY

3. RESULTS

3.1. Basic clinical characteristics

We identified colorectal cancer patients from the Severance cohort who received immune checkpoint inhibitors between 2017 and 2024, with known best responses and outcomes. The best response was measured based on RECIST criteria, resulting in 1 patient classified as a complete responder (CR), 21 as partial responders (PR), 28 as stable disease (SD), and 46 as progressive disease (PD), with 10 classified as not estimable (NE). Due to data imbalance in our analysis, we binarized the cohort, classifying CR and PR, or patients with progression-free survival (PFS) of 180 days or more, as responders, while the remaining patients were classified as non-responders. Additionally, to account for the distribution of the data in logistic regression, we also binarized variables such as stage, sidedness, microsatellite instability (MSI), and drug type. Detailed information about these variables is provided in Table 1. The comparison of clinical variables between the responder and non-responder groups revealed significant differences in progression-free survival (PFS) and overall survival (OS) (Table 2). The cohort consisted of 36 responders and 70 non-responders. It was confirmed that age, sex, stage, and sidedness do not statistically contribute to the classification of response to immune checkpoint blockade (ICB). However, microsatellite instability (MSI) status showed a highly significant difference ($p < 0.001$) between responders and non-responders, with MSI-H patients being more likely to be responders. Additionally, responders had statistically significantly higher progression-free survival and overall survival compared to non-responders (both $p < 0.001$). When comparing the two groups using univariable and multivariable analyses, there were no significant differences in binarized sidedness (right and left), age (old and young), sex (male and female), and stage (I, II, III, and IV) (Figure 3).

연세대학교
YONSEI UNIVERSITY

Table 2. Comparison of clinical features between R (*n* = 36) and NR (*n* = 70) group

		Responder	Non-responder	*p-value*
Age	Mean (SD)	54 (14.9)	55.7 (14.1)	0.655
Progression free survival	Mean (SD)	18.2 (13.4)	1.7 (1.1)	<0.001
Overall survival	Mean (SD)	31.4 (22)	12 (13.8)	<0.001
Sex	M (%)	19 (17.9)	44 (41.5)	0.428
	F (%)	17 (16)	26 (24.5)	
Stage	I, II, III (%)	24 (22.6)	33 (31.1)	0.102
	IV (%)	12 (11.3)	36 (34)	
Sidedness	Right (%)	17 (16)	24 (22.6)	0.229
	Left (%)	18 (17)	46 (43.4)	
Microsatellite	MSI-H (%)	25 (23.6)	20 (18.9)	<0.001
	MSS (%)	6 (5.7)	36 (34)	

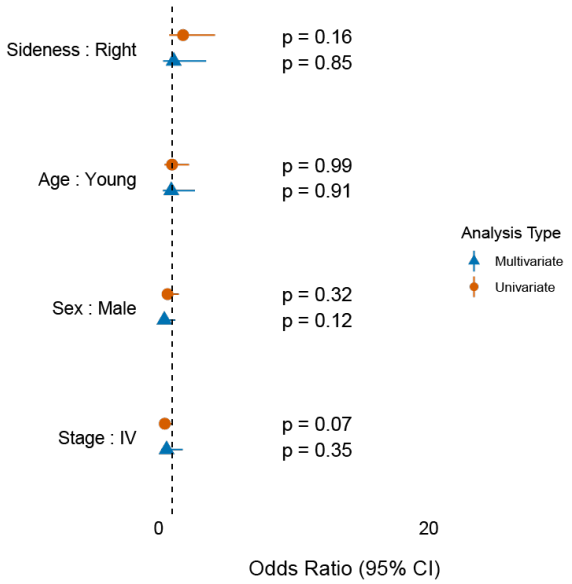

Figure 3. Comparison of clinical variables between clinical groups. This section presents the hazard ratios of clinical variables based on the estimate values calculated through logistic regression for the best response.

연세대학교
YONSEI UNIVERSITY

3.2. The impact of clinical variables on ICB response

3.2.1. Comparison of ICB markers between ICB response groups

In our cohort, we compared the expression information of ICB markers (*CD274*, *PDCD1*, *CTLA4*) in 53 individuals, consisting of 13 responders and 38 non-responders (Figure 4). When comparing PD-1 (*PDCD1*), *CTLA4*, and PD-L1 (*CD274*) between the two groups, the p-values were 0.52, 0.96, and 0.82, respectively, indicating no significant differences. Furthermore, when comparing the two groups for the additional inhibitory markers *B7-1*, *GAL9*, *PD-L2*, and *TIM3*, *B7-1* and *GAL9* showed no significant differences with p-values of 0.59 and 0.84, respectively. However, *PD-L2* and *TIM3* had p-values of less than 0.01, indicating higher expressions in the responder group (Figure 5).

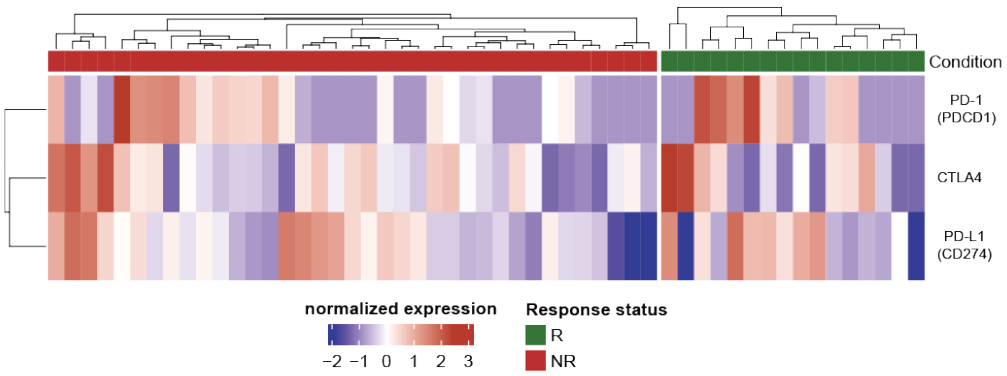

Figure 4. Comparison of expressions of conventional ICB markers. The figure illustrates the expression levels of ICB markers (*CD274*, *PDCD1*, *CTLA4*) in the cohort of 53individuals, consisting of 13 responders and 38 non-responders. The green color represents the ICB responder group, while the red color indicates the non-responder group.

연세대학교
YONSEI UNIVERSITY

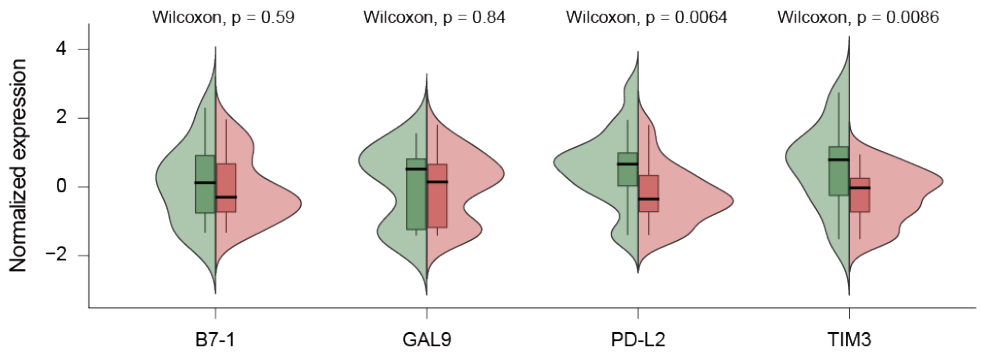

Figure 5. Comparison of expressions of other ICB markers. The figure illustrates the expression levels of other ICB markers (*B7-1*, *GAL9*, *PD-L2*, *TIM3*) in the cohort of 53 individuals, consisting of 13 responders and 38 non-responders. The green color represents the ICB responder group, while the red color indicates the non-responder group.

3.2.2. The impact of MSI-H status (MSI-H vs MSS/MSI-L) on prognosis

In a cohort of 103 patients, a comparison of progression-free survival (PFS) and overall survival (OS) between 45 patients with MSI-H and 41 patients with MSS/MSI-L confirmed that MSI status significantly contributes to prognosis (Figures 6 and 7). Patients with MSI-H exhibited statistically longer PFS and OS compared to those with MSS/MSI-L (both p-value < 0.001).

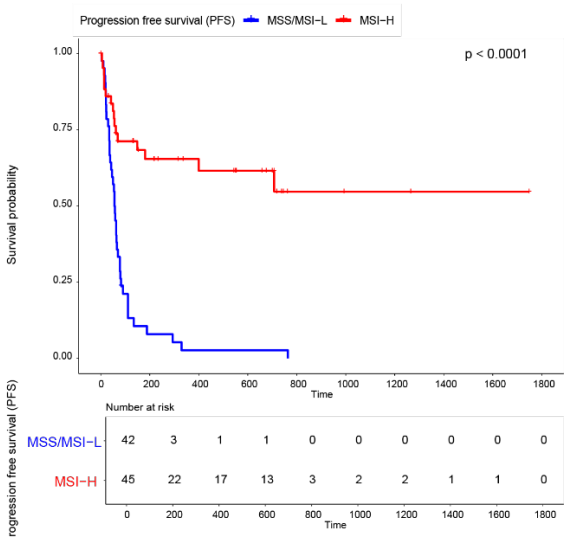

Figure 6. PFS comparison between MSI-H and MSS/MSI-L groups. This Kaplan-Meier plot illustrates the comparison of PFS between MSI-H and MSS/MSI-L. The red line represents MSI-H, while the blue line represents MSS/MSI-L.

연세대학교
YONSEI UNIVERSITY

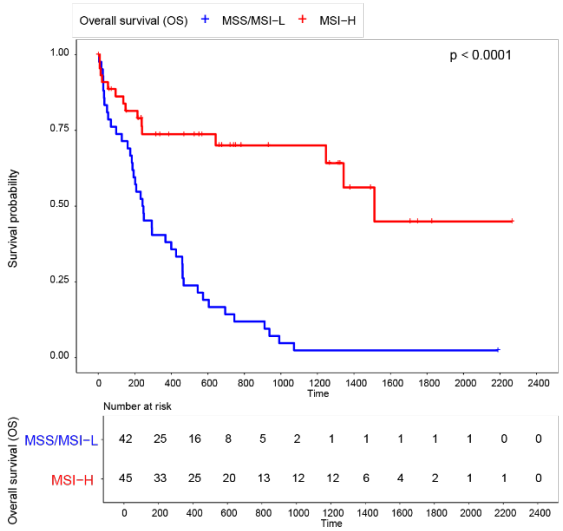

Figure 7. OS comparison between MSI-H and MSS/MSI-L groups. This Kaplan-Meier plot illustrates the comparison of OS between MSI-H and MSS/MSI-L. The red line represents MSI-H, while the blue line represents MSS/MSI-L.

3.2.3. The impact of ICB response on prognosis

In a cohort of 106 patients, a comparison of progression-free survival (PFS) and overall survival (OS) between 36 responders and 70 non-responders revealed that responders had significantly better prognoses (Figures 8 and 9). Both PFS and OS showed p-values of less than 0.001, indicating that responders experienced delayed disease progression and longer survival.

연세대학교
YONSEI UNIVERSITY

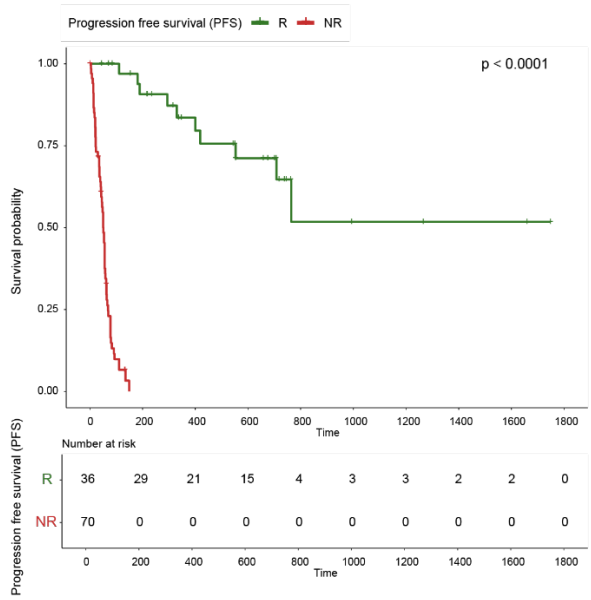

Figure 8. PFS comparison between R and NR groups. This Kaplan-Meier plot illustrates the comparison of PFS between Responder and Non-responder. The green line represents R group, while the red line represents NR group.

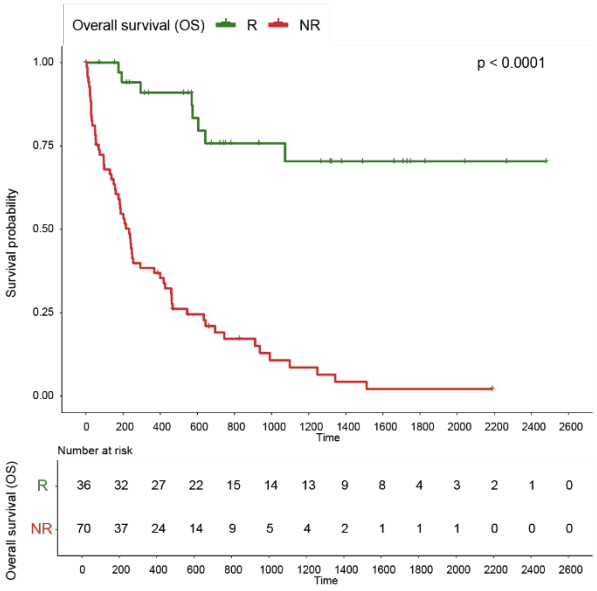

Figure 9. OS comparison between R and NR groups. This Kaplan-Meier plot illustrates the comparison of OS between Responder and Non-responder. The green line represents R group, while the red line represents NR group.

연세대학교
YONSEI UNIVERSITY

3.2.4. The impact of ICB response and MSI status on prognosis

When comparing PFS across the four conditions, considering both ICB response and MSI status revealed statistically significant longer PFS (Figure 10). Patients who were ICB responders with MSI-H status experienced the slowest disease progression and the longest survival. Following them, patients with MSS/MSI-L status who were ICB responders also had favorable outcomes. In contrast, patients who did not respond to ICB showed no significant differences when comparing MSI-H and MSS/MSI-L, and they exhibited the poorest prognosis among the four groups, regardless of MSI status. While patients who were ICB responders with MSI-H status showed the longest OS, there were no significant differences among the other groups (Figure 11).

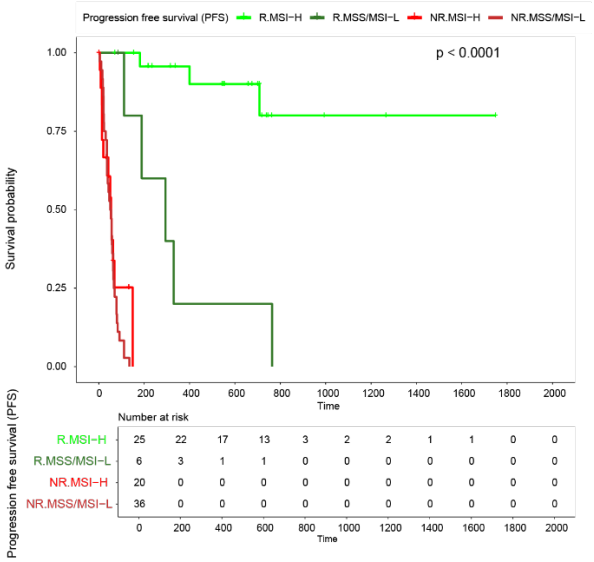

Figure 10. PFS Comparison among ICB response and MSI status groups. This Kaplan-Meier plot illustrates the comparison of PFS among the four groups: (i) ICB responders with MSI-H status (light green line), (ii) ICB responders with MSS/MSI-L status (dark green line), (iii) ICB non-responders with MSI-H status (light red line), and (iv) ICB non-responders with MSS/MSI-L status (dark red line).

연세대학교
YONSEI UNIVERSITY

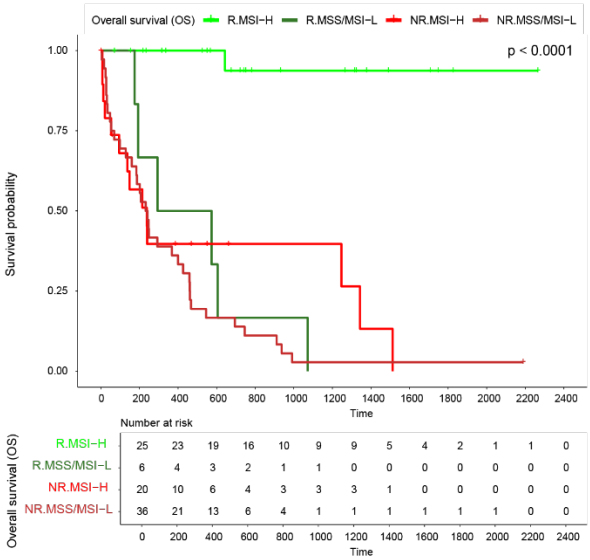

Figure 11. OS Comparison among ICB response and MSI status groups. This Kaplan-Meier plot illustrates the comparison of OS among the four groups: (i) ICB responders with MSI-H status (light green line), (ii) ICB responders with MSS/MSI-L status (dark green line), (iii) ICB non-responders with MSI-H status (light red line), and (iv) ICB non-responders with MSS/MSI-L status (dark red line).

3.2.5. The association between ICB response and MSI status

We investigated the relationship between ICB (immune checkpoint blockade) response and MSI (microsatellite instability) status, and our results confirmed a statistically significant correlation between ICB response and MSI status within our cohort, with a p-value of less than 0.001 (Figure 12). As suggested, higher levels of MSI-H were associated with a greater likelihood of response, and our results replicated this finding, indicating that patients with MSI-H were more likely to be responders. This implies the utility of MSI-H as an independent biomarker for ICB response.

연세대학교
YONSEI UNIVERSITY

Figure 12. Correlation between ICB response and MSI status. This figure illustrates the relationship between ICB response and MSI status within our cohort. The colors represent the ICB response groups; green indicates responders, while red indicates on-responders.

3.3. Genomic profiling

3.3.1. Association between best response group and TMB

In our cohort, there were a total of 52 patients with genomic data, comprising 15 responders (28.8%) and 37 non-responders (71.2%). We classified the relative binarized TMB (tumor mutational burden) groups based on the mean TMB values and found no significant correlation between the TMB groups and the ICB response groups, with a p-value of 0.064 (Figure 13). Furthermore, when comparing TMB values between the ICB response groups, we observed no significant difference, with a p-value of 0.75 (Figure 14).

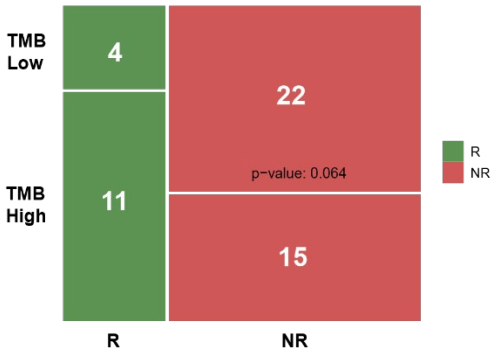

Figure 13. Comparison between best responder group and TMB binarized group. The TMB binarized groups were classified based on the mean TMB value within the cohort, and the p-value was calculated using Fisher's exact test. Red represents non-responders, while green represents responders.

연세대학교
YONSEI UNIVERSITY

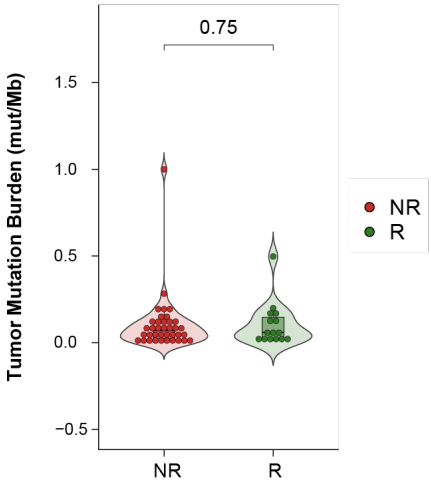

Figure 14. Comparison of TMB between best responder groups. TMB was compared between the two groups based on the number of mutations per megabase. A Wilcoxon test was performed.

3.3.2. Mutation landscape for colorectal cancer

Genomic profiling was conducted utilizing the mutation landscape of driver mutations in colorectal cancer as listed in the IntOGen database. This analysis included the top five genes with the highest frequencies in colorectal cancer: *APC*, *TP53*, *KRAS*, *PIK3CA*, and *SMAD4*. These genes showed no significant differences in mutation frequency compared to the large cohort reported in the IntOGen database for colorectal cancer (APC: 61%; TP53: 56%; KRAS: 40%; PIK3CA: 17%; SMAD4: 11%). Furthermore, the comparison between the two groups showed that most driver mutations exhibited no significant differences (Figure 15). However, among the driver mutations, three genes demonstrated statistically significant differences: *AMER1* ($p = 0.02$), *ATM* ($p = 0.05$), and *CTNNB1* ($p < 0.01$). Notably, *AMER1* (R: 5/15, NR: 2/37) and *CTNNB1* (R: 5/15, NR: 1/37) were more frequent in the responder group, whereas *ATM* was significantly more prevalent in the non-responder group (R: 0/15, NR: 10/37).

연세대학교
YONSEI UNIVERSITY

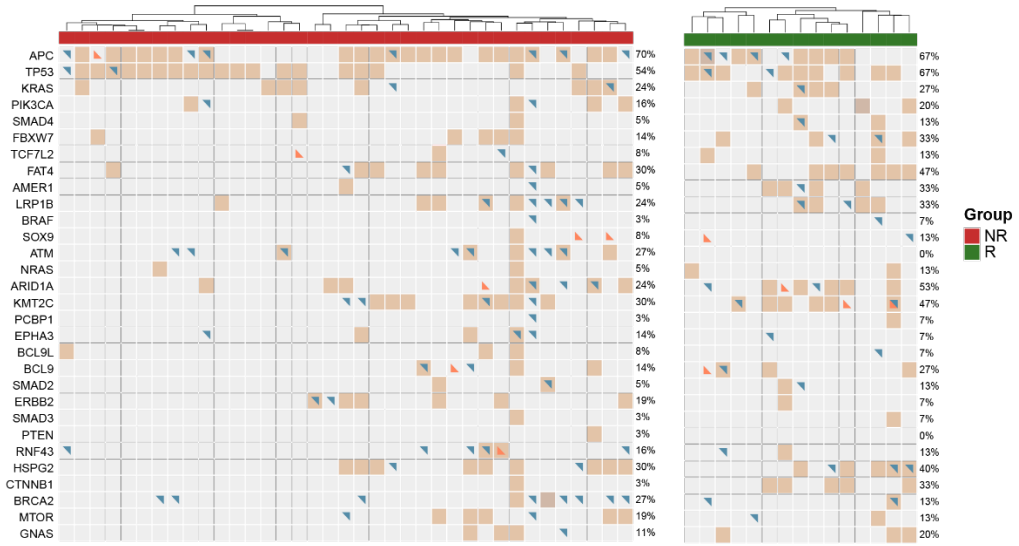

Figure 15. Landscape of driver mutations in colorectal cancer. The landscape of driver mutations in colorectal cancer was visualized using an oncoplot based on the driver mutation list from the IntOGen database. The red asterisks indicate driver mutations with statistically significant differences between the two groups.

3.4. Transcriptomic profiling

3.4.1. Expression-based PCA Analysis

A total of 51 patients with expression data were included (13 responders and 38 non-responders), and PCA was conducted based on the gene by sample matrix to assess the variance and patterns in gene expression data. This analysis demonstrates the absence of any clearly identifiable batch effects within the cohort (Figure 16).

연세대학교
YONSEI UNIVERSITY

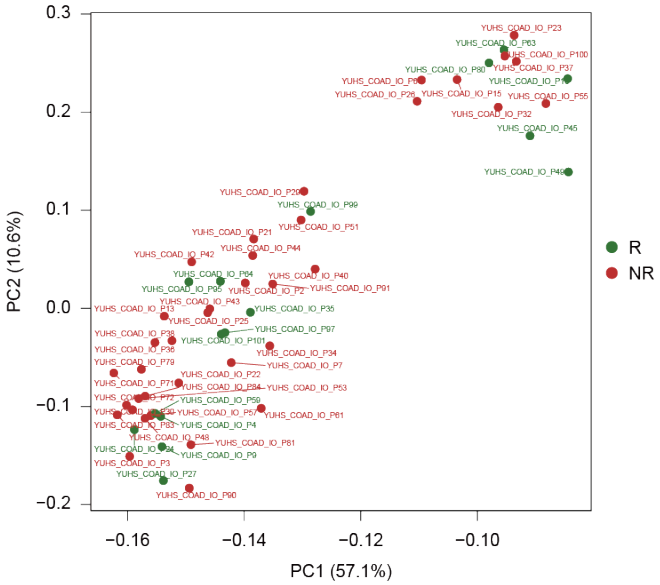

Figure 16. Visualization of PCA based on expression patterns. PCA visualization was performed using the expression matrix based on DESeq2 VST normalization. The red dots represent non-responders, while the green dots indicate responders.

3.4.2. Identification of differentially expressed genes

There are a total of 4 differentially expressed genes (DEGs) between the two groups (Figure 17). All of these are down-regulated DEGs in responders. These genes are OR4N3P (log2FC = -23.2), COL25A1 (log2FC = -4.6), CYP2E1 (log2FC = -4.4), and MUC12 (log2FC = -2.4).

연세대학교
YONSEI UNIVERSITY

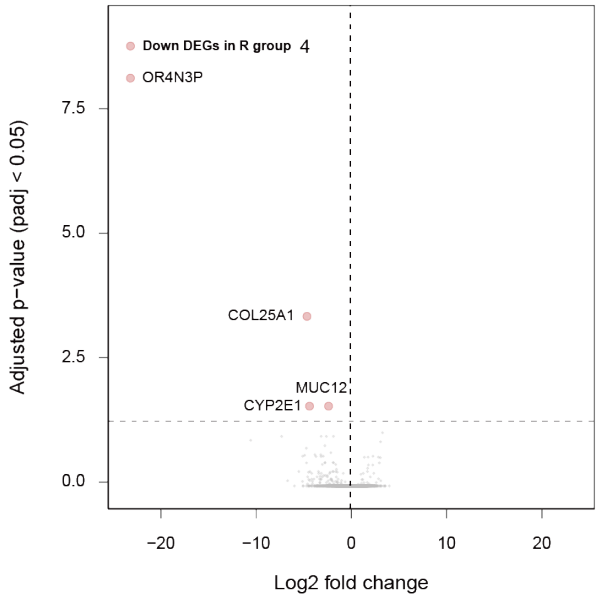

Figure 17. Visualization of DEG volcano plot. The volcano plot is visualized based on differentially expressed genes (DEGs). The x-axis represents the log2 fold change, while the y-axis displays the adjusted p-value. The significance threshold is set at an adjusted p-value of 0.05.

3.4.3. Comparison of tumor microenvironments

To compare the tumor microenvironments between responders and non-responders, we evaluated the tumor microenvironment scores of the two groups by assessing immune cell infiltration (IIS), T cell infiltration (TIS), cytolytic activity (CYT), angiogenesis (ANG), and the antigen presentation mechanism (APM) (Figure 18). As a result, no significant differences were found between the two groups in all tumor microenvironment-related scores. Additionally, we compared interferon gamma (IFNG), T cell dysfunction and exclusion score, cancer-associated fibroblast score, cytotoxic lymphocyte score (CTL), and myeloid-derived suppressor cell score to assess T cell functionality. The results showed that there were no statistically significant differences in T cell functionality-related scores between the two groups (Figure 19).

연세대학교
YONSEI UNIVERSITY

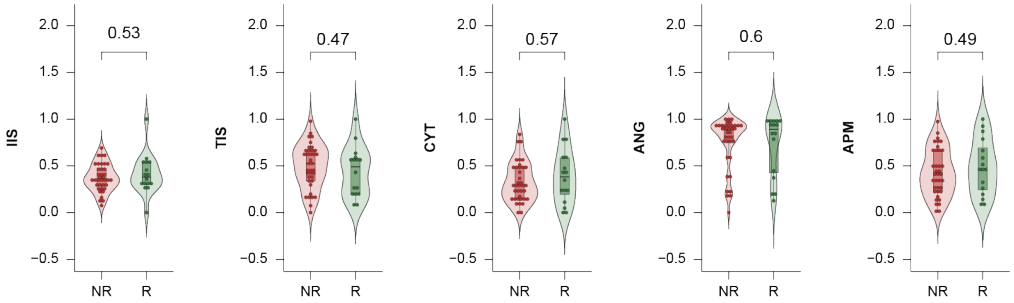

Figure 18. Comparison of tumor microenvironments. Tumor microenvironment scores were evaluated based on immune cell infiltration (IIS), T cell infiltration (TIS), cytolytic activity (CYT), angiogenesis (ANG), and the antigen presentation mechanism (APM).

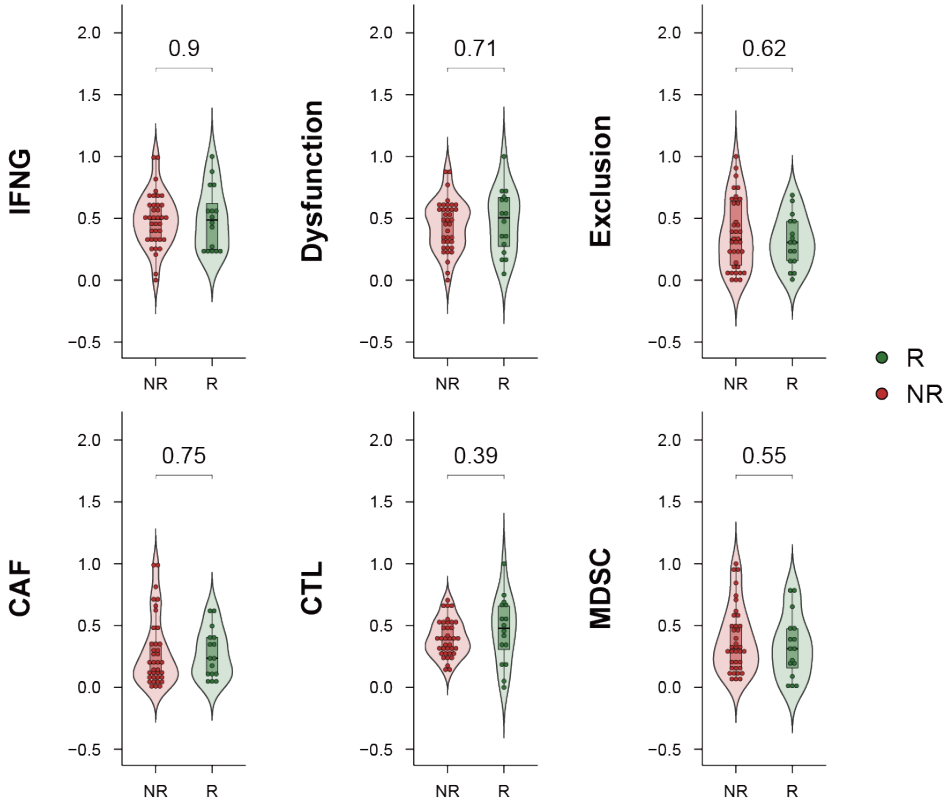

Figure 19. Comparison of T cell functionality related scores. Evaluation of T cell functionality through the comparison of interferon gamma (IFNG) levels, T cell dysfunction and exclusion score, cancer-associated fibroblast score, cytotoxic lymphocyte score (CTL), and myeloid-derived suppressor cell score.

연세대학교
YONSEI UNIVERSITY

3.4.4. Gene set enrichment analysis

When comparing responders and non-responders based on the REACTOME database, several terms were up-regulated in responders, including neutrophil degranulation (FDR < 0.001), and chemokine receptors bind chemokines (FDR < 0.001) (Figure 20). Additionally, terms related to extracellular matrix (ECM) remodeling, such as assembly of collagen fibrils and other multimeric structures (FDR < 0.001), and collagen formation (FDR < 0.01) were also found to be up-regulated in the responder group (Figure 20). Furthermore, when comparing the two groups using the cancer hallmark database, we observed increases in inflammatory response (FDR < 0.01), and allograft rejection (FDR < 0.01) in the responder group (Figure 21). Additionally, using the biological process-related gene ontology database, terms related to antigen processing and presentation (FDR < 0.05) were significantly increased in the responder group (Figure 22).

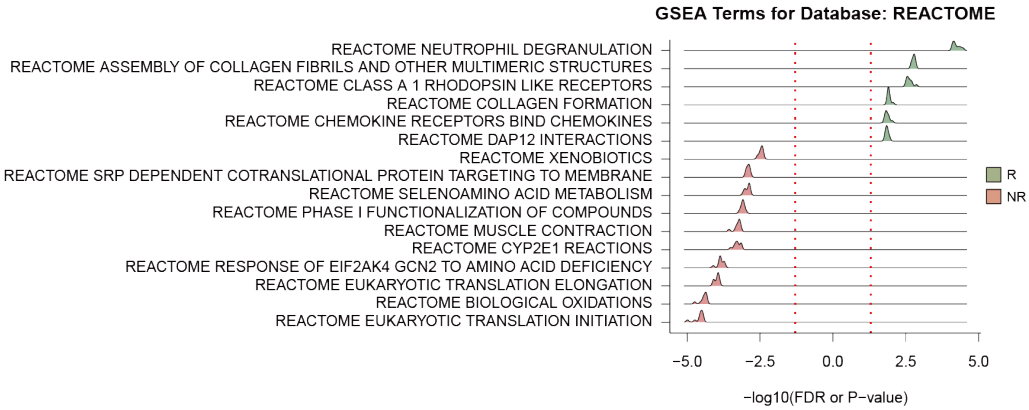

Figure 20. Ridgeline plot of gene set enrichment analysis (REACTOME). Gene set enrichment analysis was conducted through 100 rounds of bootstrapping, selecting the top 10 terms based on FDR values for both responder and non-responder groups. The x-axis represents -log10 FDR, while the y-axis displays individual terms from the REACTOME database. Green indicates up-regulated terms in responders, while red represents down-regulated terms in responders.

연세대학교
YONSEI UNIVERSITY

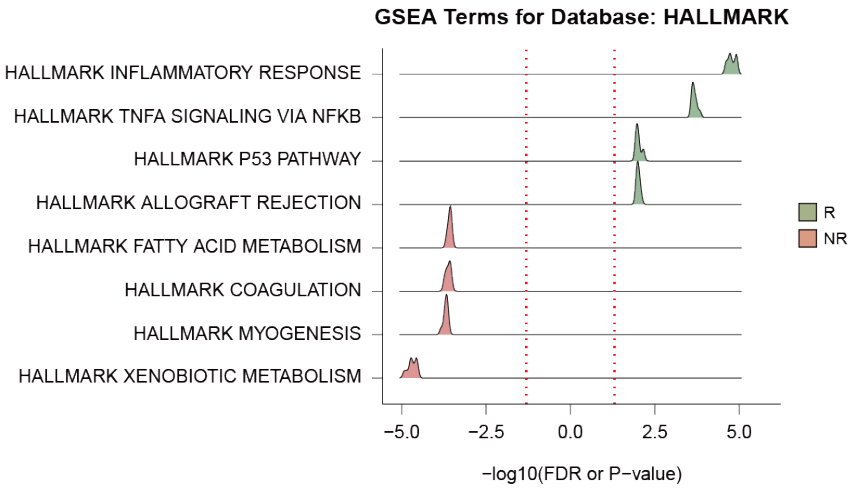

Figure 21. Ridgeline plot of gene set enrichment analysis (HALLMARK). Gene set enrichment analysis was conducted through 100 rounds of bootstrapping, selecting the top 10 terms based on FDR values for both responder and non-responder groups. The x-axis represents -log10 FDR, while the y-axis displays individual terms from the cancer hallmark database.

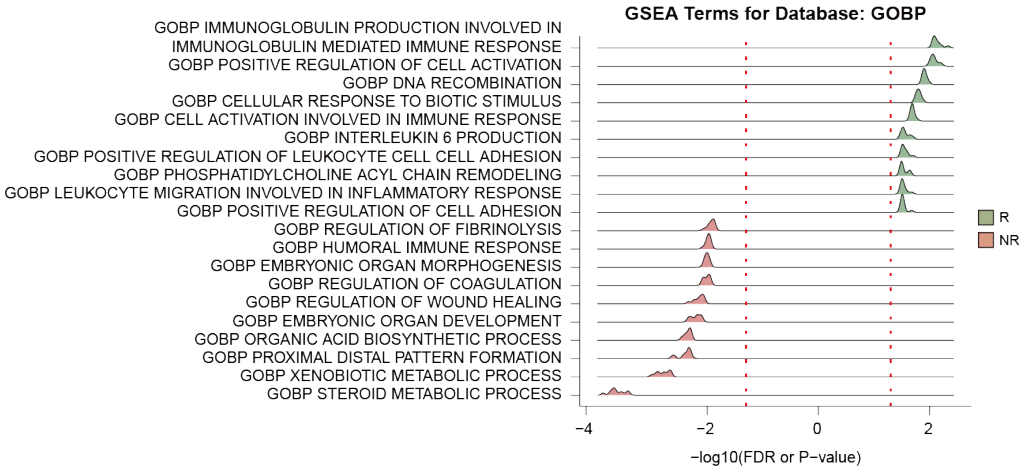

Figure 22. Ridgeline plot of gene set enrichment analysis (GO biological process terms). Gene set enrichment analysis was conducted through 100 rounds of bootstrapping, selecting the top 10 terms based on FDR values for both responder and non-responder groups. The x-axis represents -log10 FDR, while the y-axis displays individual terms from the Gene Ontology database.

연세대학교
YONSEI UNIVERSITY

3.4.5. Identification of enriched cell types

To identify the cell types that were up-regulated or down-regulated in the ICB responder group, we conducted a cell deconvolution analysis (Figure 23). When comparing the two groups, we observed an increase in M2-polarized macrophages ($p < 0.01$), regulatory T cells ($p < 0.01$), and total macrophages ($p < 0.05$) in the responder group. However, there were no cell types that showed a statistically significant decrease in the responder group compared to the non-responder group.

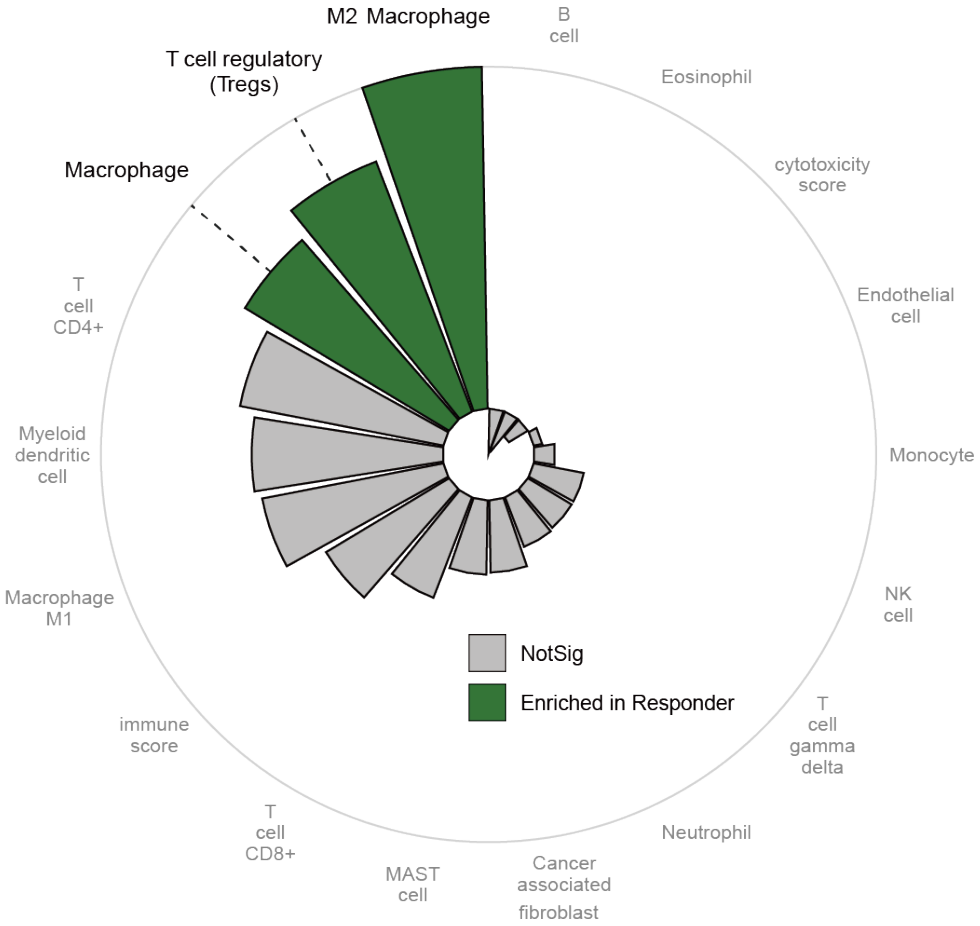

Figure 23. Cell type deconvolution analysis in ICB responder and non-responder groups. Cell deconvolution analysis revealed significant differences in the composition of cell types between the ICB responder and non-responder groups.

연세대학교
YONSEI UNIVERSITY

3.5. Metagenomic profiling

3.5.1. Comparison of taxonomy enrichment between responders and non-Responders

In the cohort, the metagenomic data consisted of a total of 57 individuals, with 18 responders and 39 non-responders. The comparison of these two groups based on the metagenomic data revealed abundant taxonomy in both responders and non-responders. According to the LDA score, the responder group exhibited statistically significant enrichment of the following taxa, which represent the top five significant findings: *Bacteroides, Fusobacterium, Flavonifractor, Enterocloster,* and *Porphyromonas*. Conversely, the following taxa demonstrated a significant decrease in the responder group: *Croceicoccus*, *Rhizorhabdus*, *Burkholderia*, *Streptomyces*, and *Sphingomonas* (Figure 24).

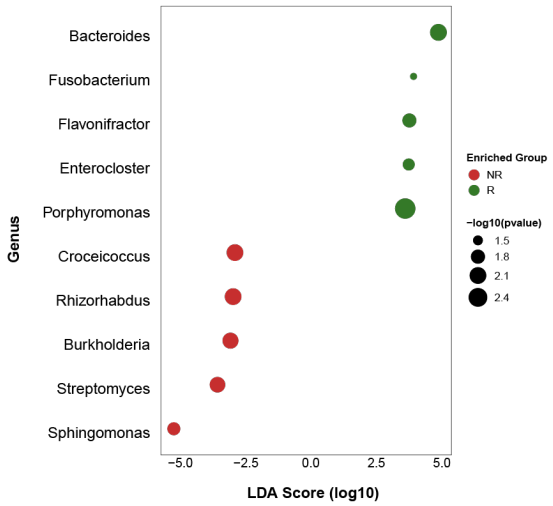

Figure 24. Enrichment of taxonomy using LDA score between two groups. A comparison of taxonomy enrichment between the two groups was performed. The x-axis represents the log10-based LDA score, while the y-axis indicates each taxonomy. Green dots represent taxa enriched in responders, while red dots indicate taxa that decreased in responders.

Additionally, we assessed the enriched microbiome in responders and non-responders using a phylogenetic tree (Figure 25). The analysis identified a total of 110 species at the species level, of which 30 were enriched in responders and 80 were downregulated in responders. Notably, *Bacteroides*, *Fusobacterium*, and *Porphyromonas* showed a significant increase in the responder group. However, the previously identified taxon *Prevotella*, which was reported to increase in responder groups in prior studies, was not replicated in our research. Furthermore, the association

연세대학교
YONSEI UNIVERSITY

of *Fusobacterium* with immune response to immunotherapy yielded results that were contrary to those reported in previous studies.

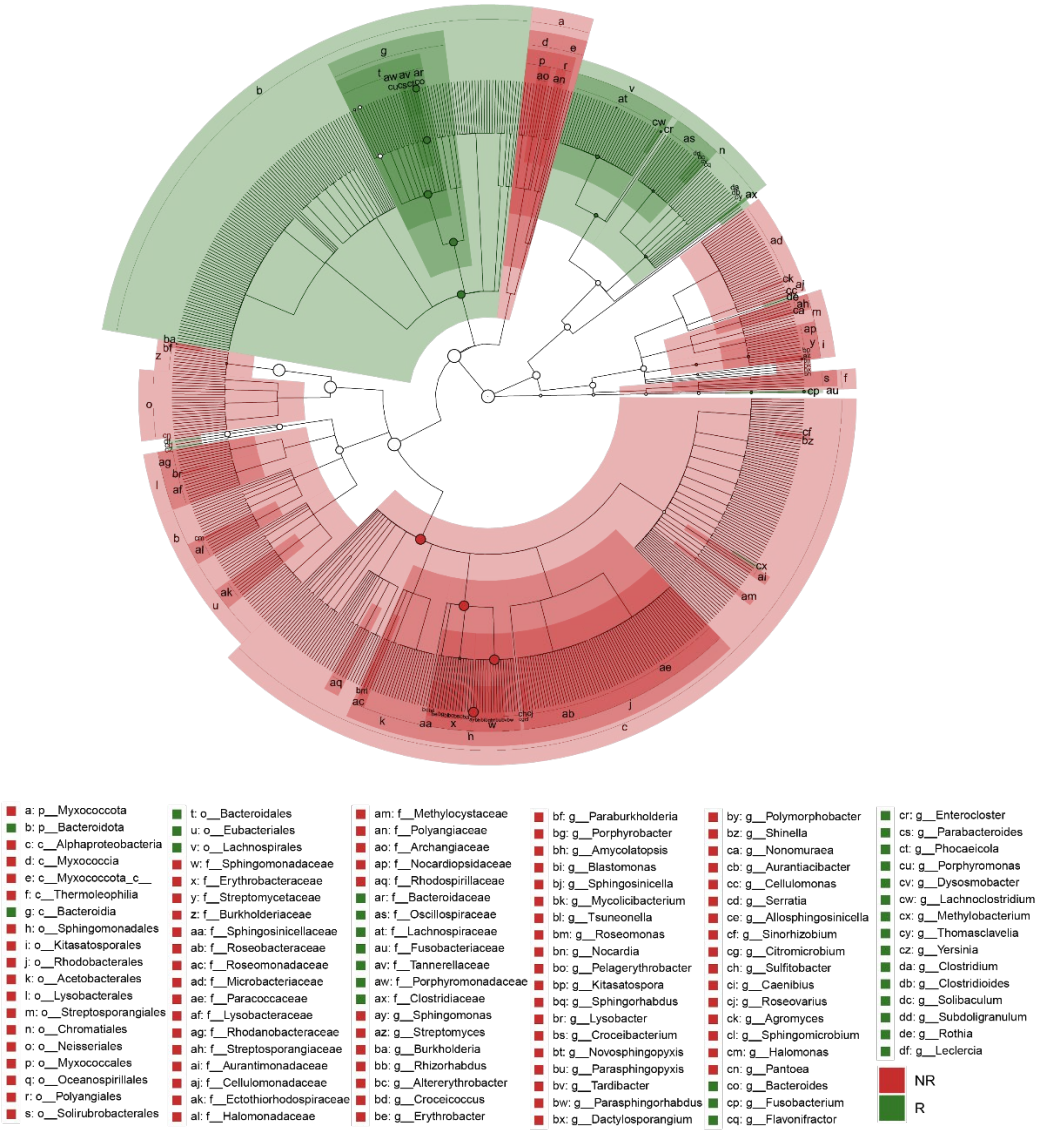

Figure 25. Cladogram of enriched microbiome in responders and non-responders. This cladogram illustrates the phylogenetic relationships among the microbiome species identified in the responder and non-responder groups.

연세대학교
YONSEI UNIVERSITY

We additionally compared the enriched microbiome between the two groups using log2 fold change (Figure 26). Specifically, we identified the top 10 genera that were increased in responders: *Bacteroides*, *Fusobacterium*, *Flavonifractor*, *Enterocloster*, *Parabacteroides*, *Phocaeicola*, *Porphyromonas*, *Dysosmobacter*, *Lachnoclostridium*, and *Methylobacterium*. Conversely, the top 10 genera that were decreased in responders included *Sphingomonas*, *Streptomyces*, *Burkholderia*, *Rhizorhabdus*, *Altererythrobacter*, *Croceicoccus*, *Erythrobacter*, *Paraburkholderia*, *Porphyrobacter*, and *Amycolatopsis*.

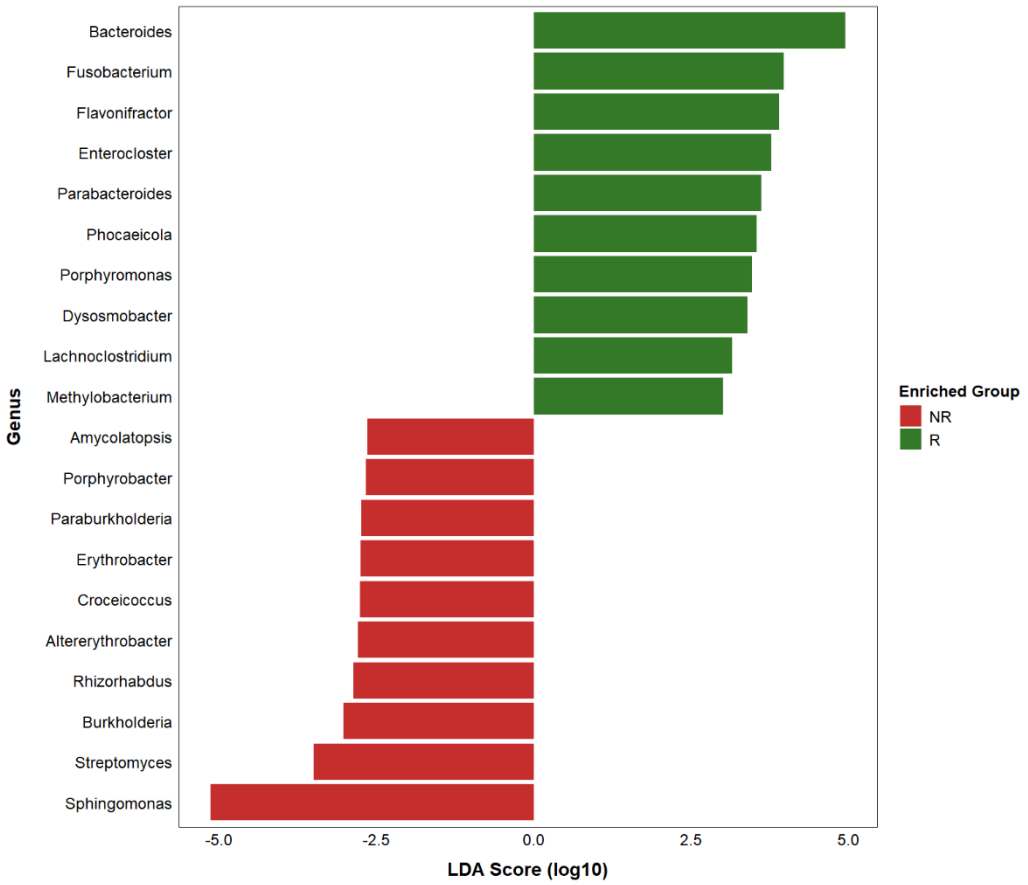

Figure 26. Enrichment of top 10 genus levels using log10 fold change. This figure presents the comparison of enriched microbiome between responders and non-responders based on log10 fold change.

Additionally, we examined the relative abundance at the phylum level (Figure 27). We compared the top five phyla with high relative abundance: *Pseudomonadota*, *Bacteroidota*, *Bacillota*,

연세대학교
YONSEI UNIVERSITY

Actinomycetota, and *Fusobacteriota* between the two groups. Among these, *Bacteroidota* was found to be significantly higher in responders ($p < 0.001$). Notably, we also observed an increase in the *Bacteroides* genus level in responders in accordance with previous results (Figure 24-26).

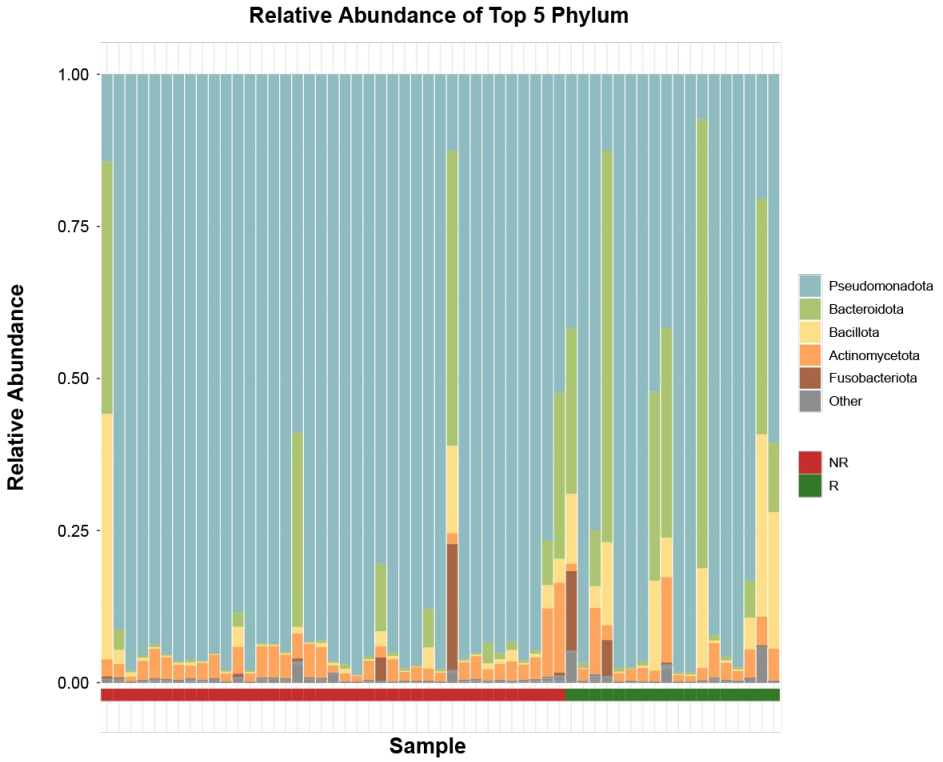

Figure 27. Comparison of relative abundance at the phylum level. This figure illustrates the relative abundance of the top five phyla: *Pseudomonadota*, *Bacteroidota*, *Bacillota*, *Actinomycetota*, and *Fusobacteriota* in responders and non-responders.

3.5.2. Gut microbial diversity indices analysis

We assessed the diversity indices of patients based on a genus-by-sample normalized matrix. When comparing these indices, no statistically significant differences were observed between the two groups (observed: 0.17, Shannon: 0.44, inverse Simpson: 0.32, Chao1: 0.17) (Figure 28).

연세대학교
YONSEI UNIVERSITY

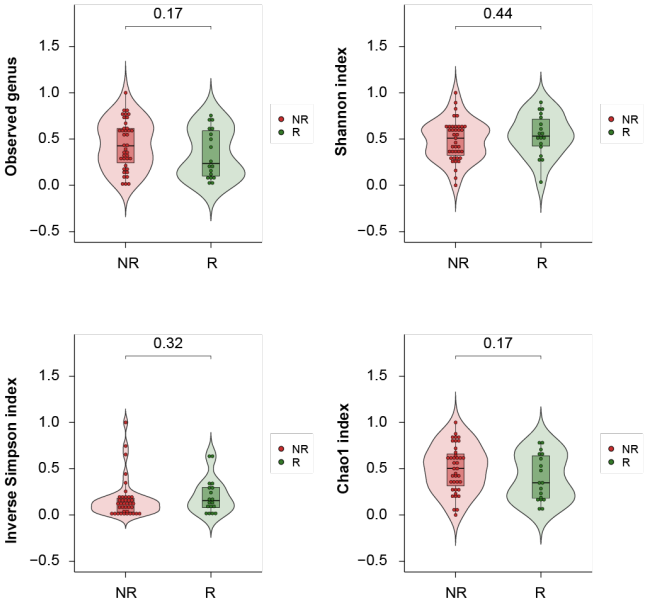

Figure 28. Comparison of microbial diversity indices between two groups. Each metric represents diversity calculated based on the genus-by-sample matrix, and it has the same meaning as the y-axis. The green color represents responders, while the red color represents non-responders.

3.6. Machine learning-based modeling for response prediction

3.6.1. Process of feature selection and importance

In this analysis, we performed feature selection for each modality: clinical information (C), DNA (D), RNA (R), metagenomic sequencing (M), pathology (P), and radiology (I).For the clinical information modality (C), we utilized four features: age, sidedness, stage, and sex, to determine feature importance based on a random forest model (Figure 29). Among these, age had the highest importance, followed by sidedness, stage, and sex, in contributing to the C modality-based model. Additionally, we assessed the contribution of each C modality feature through the odds ratios based on a ridge regression model. Among the features analyzed, sidedness exhibited an odds ratio greater than 1, while the other features demonstrated odds ratios less than 1.

연세대학교
YONSEI UNIVERSITY

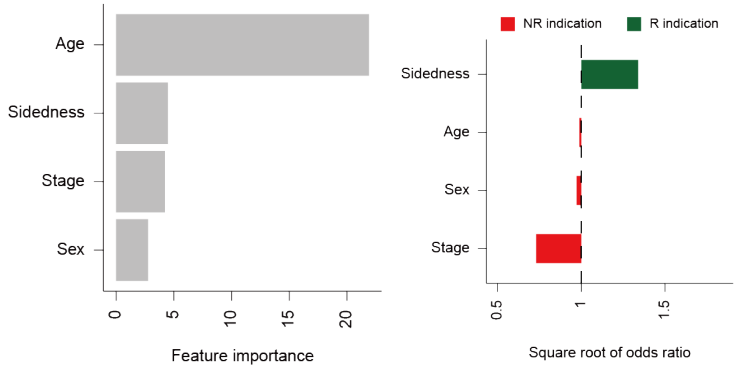

Figure 29. Feature importance and odds ratios for clinical information modality. This figure presents the feature importance ranking for the clinical information modality (C) based on a random forest model, highlighting the relative contributions of various features to the model's predictive performance. Additionally, it displays the odds ratios for these features as identified in the clinical information modality (C) through a ridge regression model. The x-axis has been scaled by applying the square root to the odds ratios, thereby facilitating a clearer visualization of the relationships between the features and clinical responses.

For the DNA modality (D), we utilized ten features: *FAT1*, *ABCA7*, *MYO10*, *KIF26A*, *HNRNPA0*, *RALGAPA2*, *GPR137B*, *PPP1R2B*, *ARNTL2*, and *PCDHGA6* (Figure 30). Among these, *FAT1* was identified as the most significant feature in contributing to the model based on the D modality. Furthermore, only *MYO10* exhibited an odds ratio of less than 1, while the remaining features showed odds ratios greater than 1.

연세대학교
YONSEI UNIVERSITY

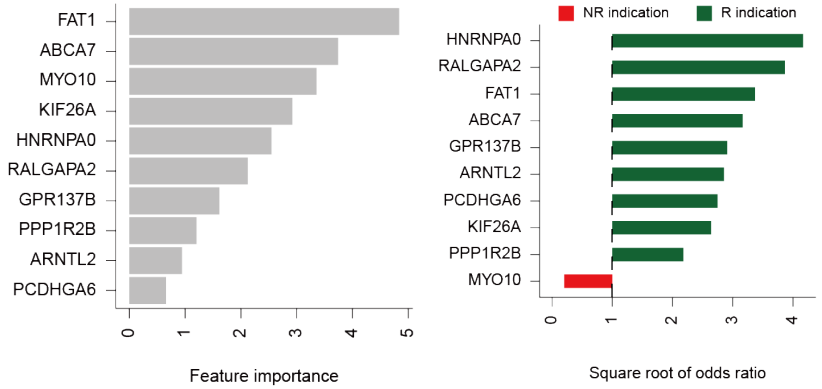

Figure 30. Feature importance and odds ratios for DNA modality. This figure presents the feature importance ranking for the DNA modality (D) based on a random forest model, highlighting the relative contributions of various features to the model's predictive performance. Additionally, it displays the odds ratios for these features as identified in the DNA modality (D) through a ridge regression model. The x-axis has been scaled by applying the square root to the odds ratios, thereby facilitating a clearer visualization of the relationships between the features and clinical responses.

In the RNA modality (R), we utilized ten features: CTL, CD8, IFNG, MSI score, Dysfunction score, MDSC, CD274, Exclusion score, TIDE, and M2-polarized macrophage (Figure 31). Among these, the cytotoxic lymphocyte score was identified as the most significant contributor to the R modality model. Furthermore, the exclusion score exhibited the highest odds ratio within the R modality, while MSI score, CD8, CTL, CD274, and M2-polarized macrophage all displayed odds ratios greater than 1. In contrast, the remaining features presented odds ratios below 1.

연세대학교
YONSEI UNIVERSITY

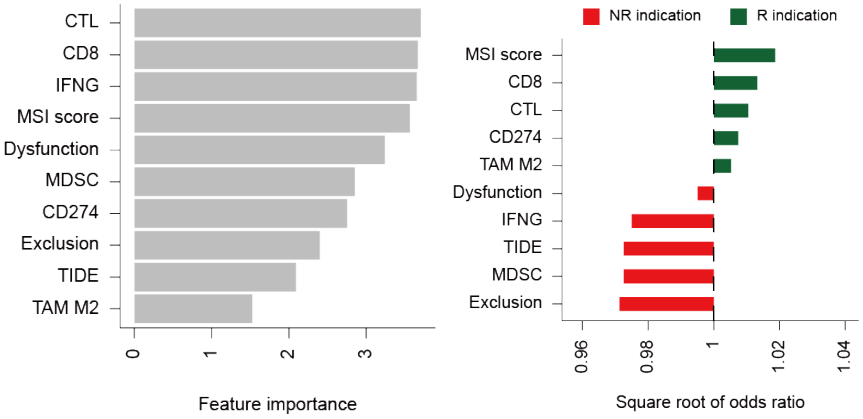

Figure 31. Feature importance and odds ratios for RNA modality. This figure presents the feature importance ranking for the RNA modality (R) based on a random forest model, highlighting the relative contributions of various features to the model's predictive performance. Additionally, it displays the odds ratios for these features as identified in the RNA modality (R) through a ridge regression model. The x-axis has been scaled by applying the square root to the odds ratios, thereby facilitating a clearer visualization of the relationships between the features and clinical responses.

In the context of the MGS modality (M), we employed ten features: *Umezawaea*, *Hydrogenimonas*, *Capillimicrobium*, *Shingomonas*, observed, *Citrobacter*, *Porphyromonas*, *Streptomyces*, *Pseudolysobacter*, and *Sporomusa* (Figure 32). Among these features, *Umezawaea* was recognized as the most influential factor in the M modality model. Additionally, *Porphyromonas* demonstrated the highest odds ratio, whereas *Pseudolysobacter*, *Citrobacter*, *Sporomusa*, and *Hydrogenimonas* showed odds ratios exceeding 1, while the other features had odds ratios below 1.

연세대학교
YONSEI UNIVERSITY

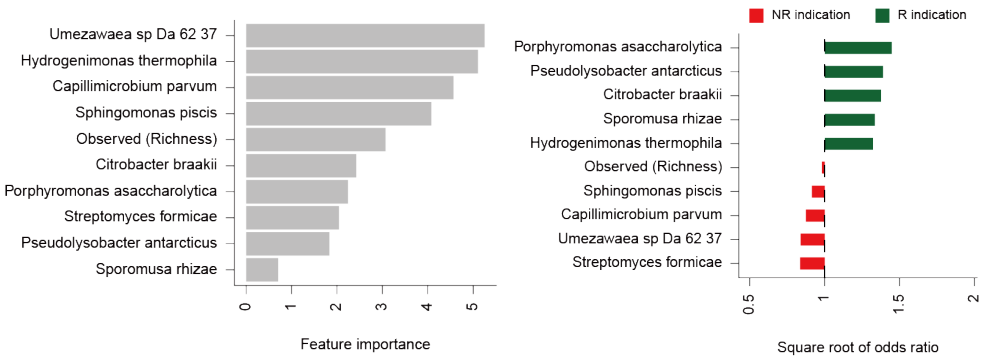

Figure 32. Feature importance and odds ratios for MGS modality. This figure presents the feature importance ranking for the MGS modality (M) based on a random forest model, highlighting the relative contributions of various features to the model's predictive performance. Additionally, it displays the odds ratios for these features as identified in the MGS modality (M) through a ridge regression model. The x-axis has been scaled by applying the square root to the odds ratios, thereby facilitating a clearer visualization of the relationships between the features and clinical responses.

In the pathology modality (P), we analyzed four key features: MSI status, KRAS, BRAF, and NRAS (Figure 33). Among these features, MSI status emerged as the most significant factor contributing to the P modality model. Additionally, MSI status exhibited the highest odds ratio, while BRAF and KRAS showed odds ratios exceeding 1. Conversely, the other features had odds ratios that fell below 1.

연세대학교
YONSEI UNIVERSITY

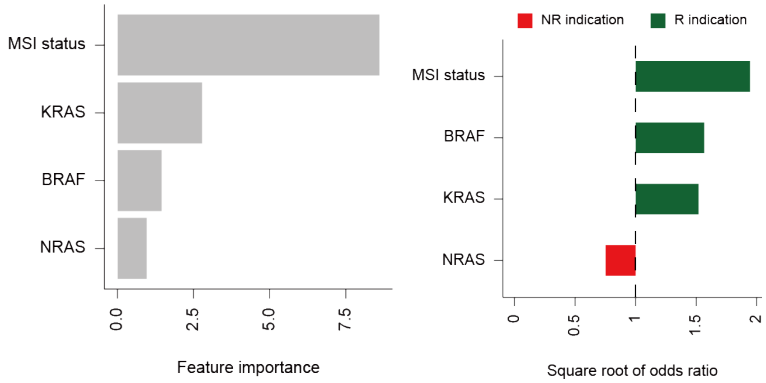

Figure 33. Feature importance and odds ratios for pathology modality. This figure presents the feature importance ranking for the pathology modality (P) based on a random forest model, highlighting the relative contributions of various features to the model's predictive performance. Additionally, it displays the odds ratios for these features as identified in the pathology modality (P) through a ridge regression model. The x-axis has been scaled by applying the square root to the odds ratios, thereby facilitating a clearer visualization of the relationships between the features and clinical responses.

In the radiology modality (I), we analyzed six key features: lung metastases, bone metastases, abdominal and pelvic metastases, increase of primary tumor, lymph node metastases, and head and neck metastases (Figure 34). Among these features, lung metastases emerged as the most significant factor influencing the I modality model. Furthermore, head and neck metastases exhibited the highest odds ratio, while all other features, except for the increase of primary tumor, demonstrated odds ratios below 1.

연세대학교
YONSEI UNIVERSITY

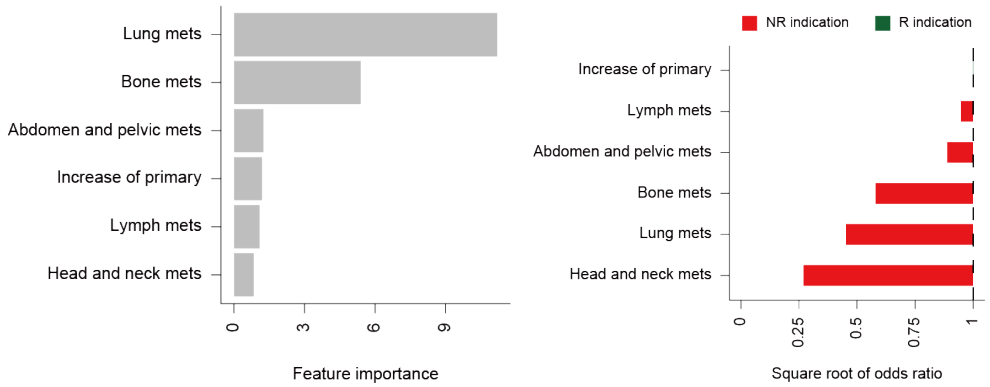

Figure 34. Feature importance and odds ratios for image modality. This figure presents the feature importance ranking for the image modality (I) based on a random forest model, highlighting the relative contributions of various features to the model's predictive performance. Additionally, it displays the odds ratios for these features as identified in the image modality (I) through a ridge regression model. The x-axis has been scaled by applying the square root to the odds ratios, thereby facilitating a clearer visualization of the relationships between the features and clinical responses.

3.6.2. Performance improvement by combining modalities and validation

We constructed models for each modality and then evaluated the performance of these models using an ensemble approach. We primarily focused on accuracy, sensitivity, positive predictive value (PPV), specificity, and negative predictive value (NPV), observing that these metrics increased with the combination of modalities. First, we confirmed that combining six modalities from the internal dataset resulted in increased accuracy, sensitivity, PPV, specificity, and NPV, demonstrating significantly better performance compared to a single-modality-based predictive tool for immune checkpoint inhibitor response (Figures 35-39). The accuracy values were as follows: for uni-modality, 0.7; for bi-modality, 0.69; for tri-modality, 0.7; for quadra-modality, 0.7; for penta-modality, 0.7; and for hexa-modality, 0.75 (Figure 35). Notably, even the uni-modality was higher than the RNA-based predictive tools IMPRES and TIDE. The sensitivity values were: uni-modality, 0.66; bi-modality, 0.63; tri-modality, 0.67; quadra-modality, 0.74; penta-modality, 0.8; and hexa-modality, 0.85 (Figure 36), which were also higher than those of IMPRES and TIDE. The PPV values were: uni-modality, 0.5; bi-modality, 0.59; tri-modality, 0.58; quadra-modality, 0.55; penta-modality, 0.55; and hexa-modality, 0.6 (Figure 37). Although relatively unstable compared to other metrics, these values remained higher than those of IMPRES and TIDE. The specificity values were:

연세대학교
YONSEI UNIVERSITY

uni-modality, 0.7; bi-modality, 0.7; tri-modality, 0.7; quadra-modality, 0.67; penta-modality, 0.65; and hexa-modality, 0.69 (Figure 38). Here, the uni-modality exhibited the highest performance, and in some datasets, it performed lower than IMPRES and TIDE. Finally, the NPV values were: uni-modality, 0.83; bi-modality, 0.81; tri-modality, 0.82; quadra-modality, 0.84; penta-modality, 0.86; and hexa-modality, 0.9 (Figure 39). These results indicate that the performance increased with the combination of modalities, surpassing that of IMPRES and TIDE.

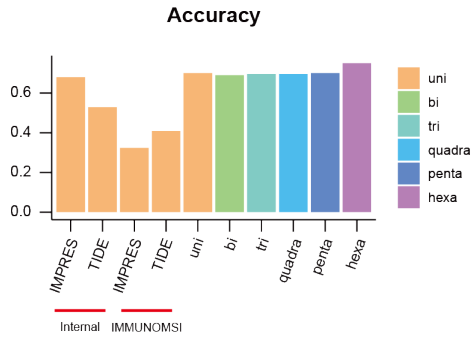

Figure 35. Comparison of Accuracy values by integrating modalities. The accuracy is presented for uni-modality (0.7), bi-modality (0.69), tri-modality (0.7), quadra-modality (0.7), penta-modality (0.7), and hexa-modality (0.75).

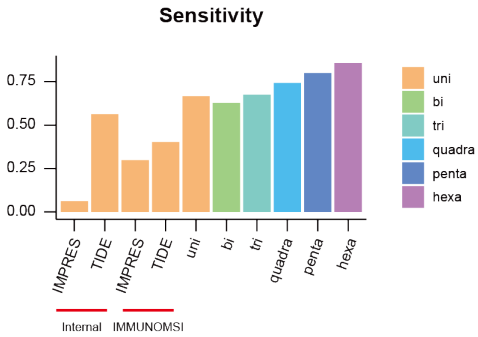

Figure 36. Comparison of Sensitivity values by integrating modalities. The accuracy is presented for uni-modality (0.66), bi-modality (0.63), tri-modality (0.67), quadra-modality (0.74), penta-modality (0.8), and hexa-modality (0.85).

연세대학교
YONSEI UNIVERSITY

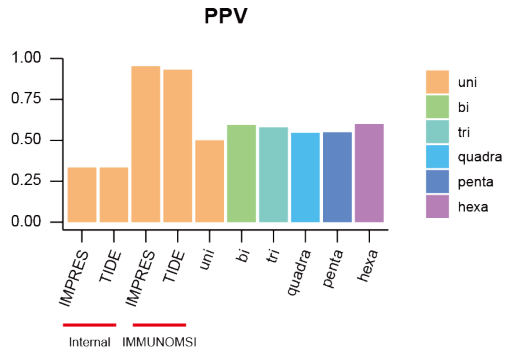

Figure 37. Comparison of PPV by integrating modalities. The accuracy is presented for uni-modality (0.5), bi-modality (0.59), tri-modality (0.58), quadra-modality (0.55), penta-modality (0.55), and hexa-modality (0.6).

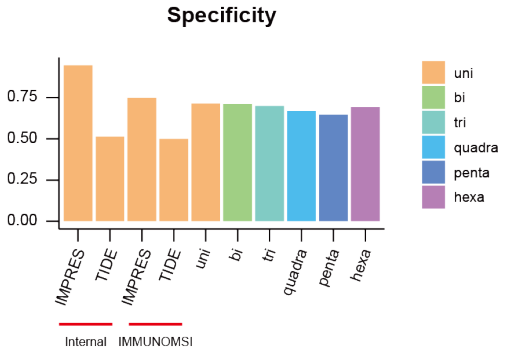

Figure 38. Comparison of Specificity values by integrating modalities. The accuracy is presented for uni-modality (0.7), bi-modality (0.7), tri-modality (0.7), quadra-modality (0.67), penta-modality (0.65), and hexa-modality (0.69).

연세대학교
YONSEI UNIVERSITY

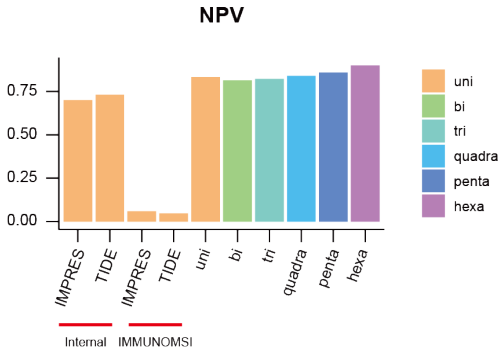

Figure 39. Comparison of NPV by integrating modalities. The accuracy is presented for uni-modality (0.83), bi-modality (0.81), tri-modality (0.82), quadra-modality (0.84), penta-modality (0.86), and hexa-modality (0.9).

3.6.3. Performance variations and validation based on modality combinations

We measured five metrics (accuracy, sensitivity, positive predictive value (PPV), specificity, and negative predictive value (NPV)) for all modalities based on combinations across the internal dataset, simulated dataset, and two external datasets (Figure 40-43). In the internal dataset, where imputation was performed, all metrics gradually improved as more modalities were combined. Specifically, in the hexa combination, we observed an accuracy of 0.6, sensitivity of 0.857, PPV of 0.6, specificity of 0.692, and NPV of 0.9 (Figure 40). Additionally, we assumed scenarios with missing features for each modality and utilized the internal dataset to evaluate model performance based on the simulated dataset without imputation (Figure 41). The results indicated that while the performance was best when all modalities were combined, accuracy, sensitivity, and PPV gradually increased with more modality combinations. In contrast, specificity and NPV remained stable.

연세대학교
YONSEI UNIVERSITY

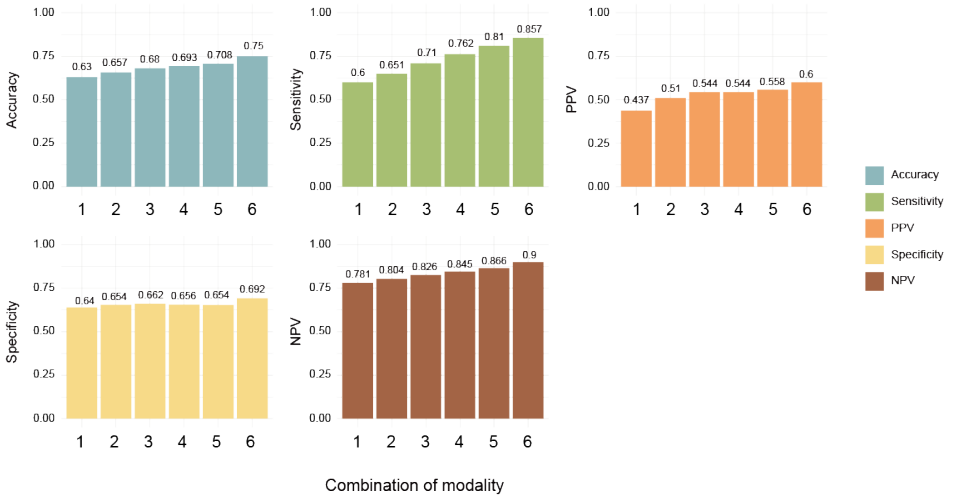

Figure 40. Comparison of multiple metrics by integrating modalities in the internal dataset with imputation. This figure visualizes five metrics, where the y-axis represents the value of each metric, and the x-axis denotes the number of modality combinations.

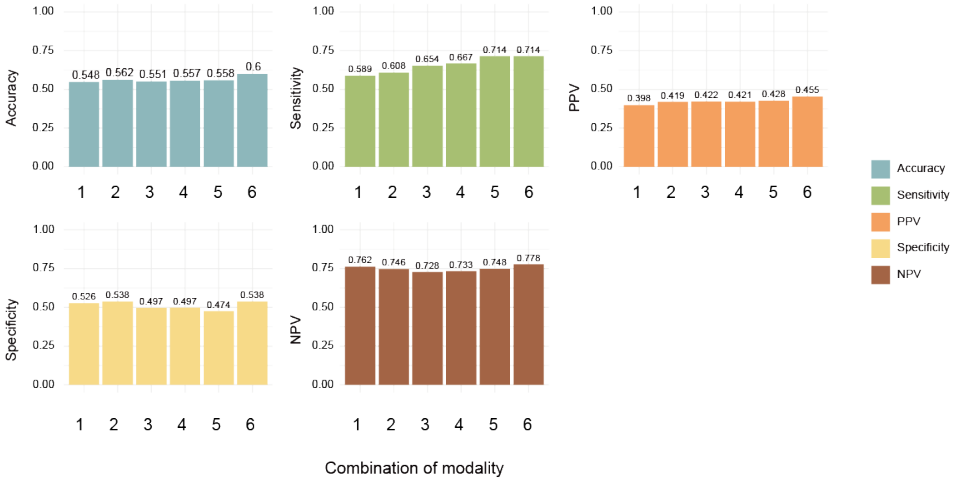

Figure 41. Comparison of multiple metrics by integrating modalities in the simulated dataset. This figure visualizes five metrics, where the y-axis represents the value of each metric, and the x-axis denotes the number of modality combinations.

연세대학교
YONSEI UNIVERSITY

We validated the model's performance using two external validation sets (Figures 42 and 43). Utilizing the C, D, R, and P modalities, we were able to evaluate the performance of the four-modality combinations in the IMMUNOMSI dataset. The performance metrics observed were an accuracy of 0.55, sensitivity of 0.71, positive predictive value (PPV) of 0.42, specificity of 0.46, and negative predictive value (NPV) of 0.75 (Figure 42). Although these metrics are relatively low, they replicate the performance observed in previous results (Figure 41) when utilizing the four-modality combination. Additionally, in the NIPICOL dataset, we assessed the performance of a three-modality combination using the C, D, and P modalities (Figure 43), yielding metrics of accuracy 0.65, sensitivity 0.71, PPV 0.5, specificity 0.62, and NPV 0.8. While both datasets showed somewhat low AUC values, our analysis confirmed that overall performance increases as modalities are integrated.

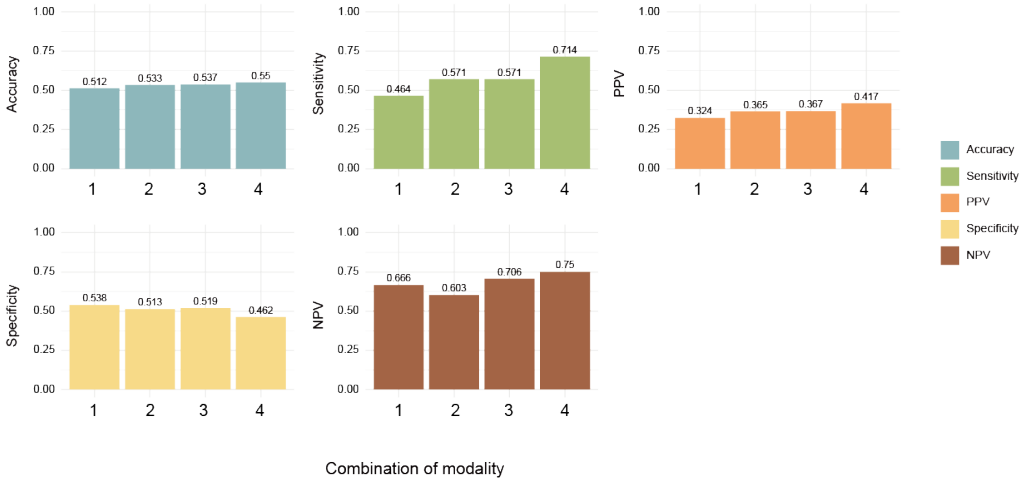

Figure 42. Comparison of multiple metrics by integrating modalities in the IMMUNOMSI dataset. This figure visualizes five metrics, where the y-axis represents the value of each metric, and the x-axis denotes the number of modality combinations.

연세대학교
YONSEI UNIVERSITY

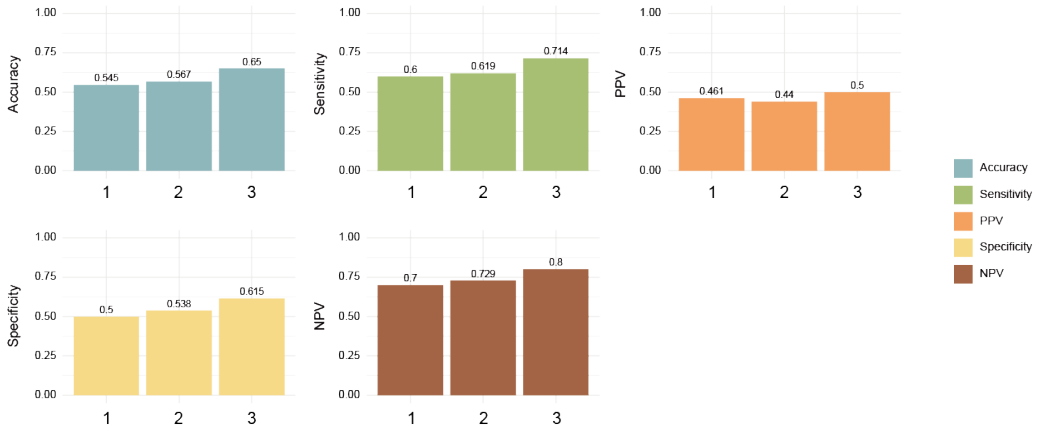

Figure 43. Comparison of multiple metrics by integrating modalities in the NIPICOL dataset. This figure visualizes five metrics, where the y-axis represents the value of each metric, and the x-axis denotes the number of modality combinations.

3.6.4. Machine learning based model performance on the internal dataset

In this analysis, we utilized an internal dataset consisting of 106 patient samples, which included partially combined data from clinical, DNA, RNA, metagenomic, pathology, and imaging modalities, to assess the performance of the predictive model when all modalities were integrated. We performed stratified 5-fold cross-validation, and the results showed a mean AUC of 0.8 in the test set, indicating that overfitting was adequately regulated (Figure 44). Furthermore, we created a simulation dataset by intentionally removing specific features from the internal dataset. Similarly, we performed stratified 5-fold cross-validation to assess the mean performance of the model that combined modalities without imputation (Figure 45). As a result, we found that the mean AUC was 0.655, indicating a decrease in model performance compared to the case with imputation.

연세대학교
YONSEI UNIVERSITY

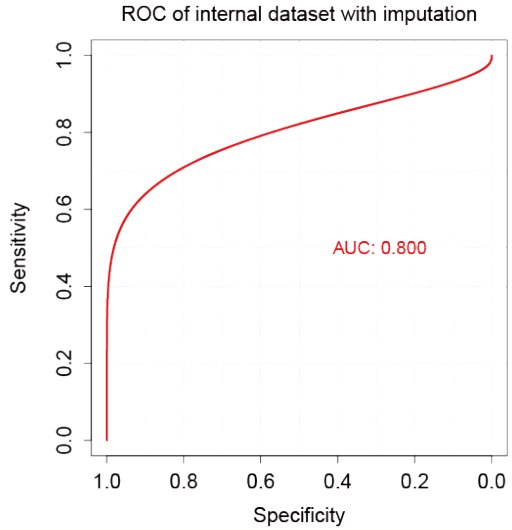

Figure 44. AUC of model performance with imputation. This figure represents the mean ROC of the test set after imputation, with the x-axis indicating specificity and the y-axis indicating sensitivity.

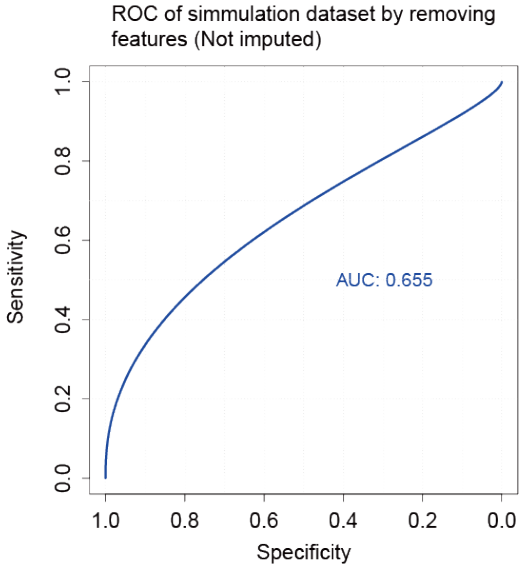

Figure 45. AUC of model performance without imputation. This figure represents the mean ROC of the simulation set without imputation, with the x-axis indicating specificity and the y-axis indicating sensitivity.

연세대학교
YONSEI UNIVERSITY

3.6.5. Machine learning based model performance on the external dataset

To evaluate the performance of the modality combination model, we utilized two independent external datasets: IMMUNOMSI and NIPICOL. The IMMUNOMSI dataset incorporated the C, D, R, and P modalities, whereas the NIPICOL dataset utilized the C, D, and P modalities to validate the model's performance (Figures 46 and 47). The IMMUNOMSI dataset showed an AUC of 0.697 (Figure 46), while the NIPICOL dataset demonstrated an AUC of 0.622 (Figure 47).

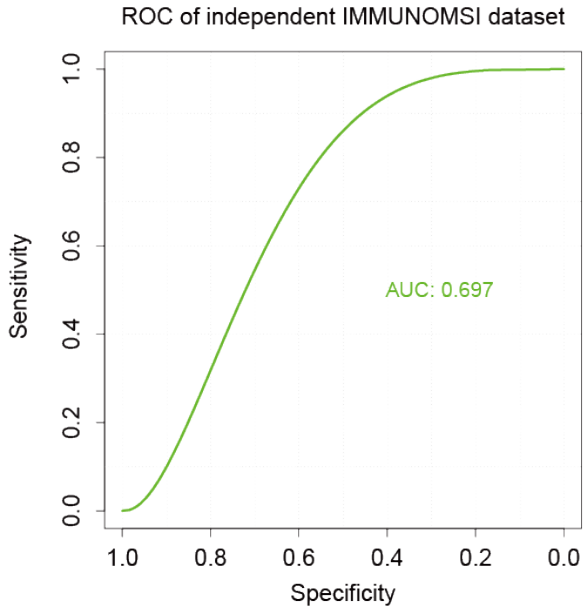

Figure 46. AUC of model performance on the external IMMUNOMSI dataset. This figure represents the mean ROC curve of the model evaluated on the IMMUNOMSI dataset, where the x-axis indicates specificity and the y-axis indicates sensitivity.

연세대학교
YONSEI UNIVERSITY

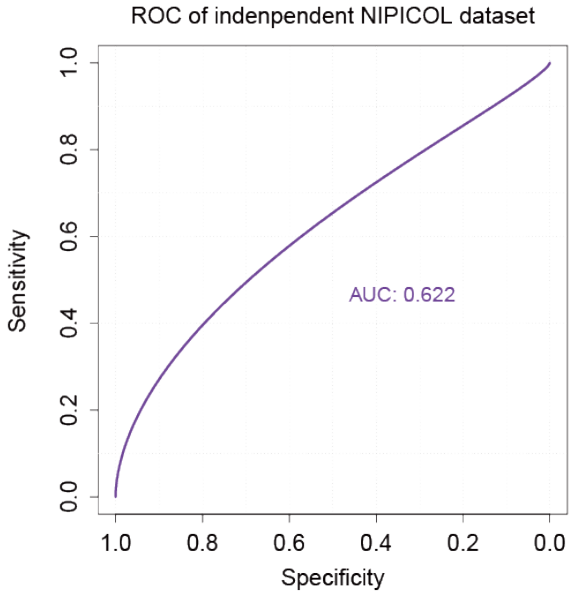

Figure 47. AUC of model performance on the external NIPICOL dataset. This figure represents the mean ROC curve of the model evaluated on the NIPICOL dataset, where the x-axis indicates specificity and the y-axis indicates sensitivity.

3.6.6. Deep learning model development

To address the challenges of limited sample size and the complexities of multimodal fusion, we developed a model utilizing deep learning techniques. Based on the HEALNet model, we fused all modalities for our development; however, we observed incomplete metrics (accuracy: 0.53, sensitivity: 0.6, specificity: 0.5, positive predictive value (PPV): 0.33, negative predictive value (NPV): 0.75) (Figure 48). We also examined the confusion matrix for the 51 samples that included all modalities (Figure 49). In the internal dataset, the HEALNet-based model achieved an AUC of 0.569, which was lower than the performance of the machine learning ensemble models (Figures 44-47 and 50).

연세대학교
YONSEI UNIVERSITY

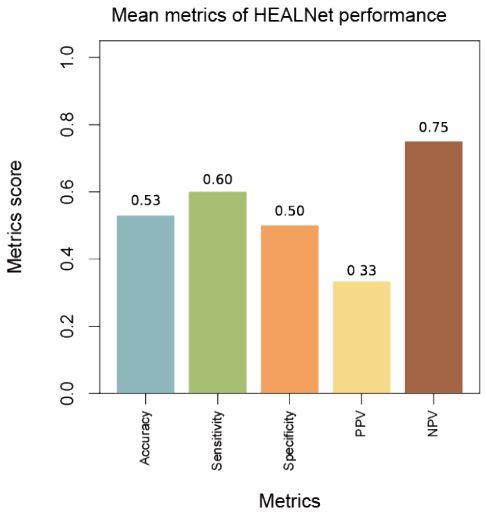

Figure 48. Performance metrics for the HEALNet Model with all modalities. This figure illustrates the performance metrics of the HEALNet model incorporating all modalities.

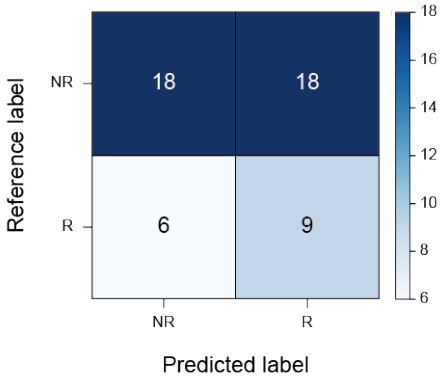

Figure 49. Confusion Matrix for the HEALNet Model. This figure presents the confusion matrix for the HEALNet model evaluated on 51 samples that included all modalities. The confusion matrix shows the distribution of true positives, true negatives, false positives, and false negatives, which provides insight into the model's classification performance.

연세대학교
YONSEI UNIVERSITY

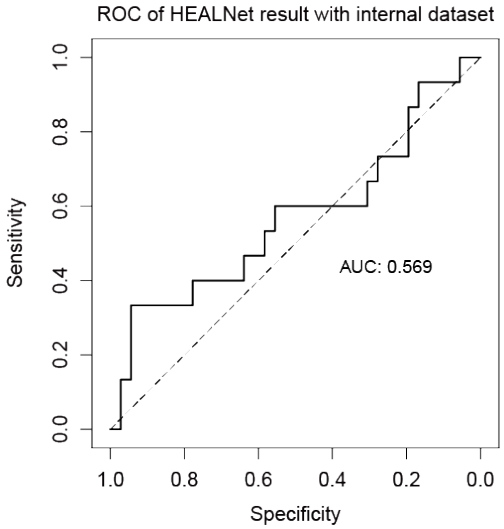

Figure 50. AUC of the model based on the HEALNet framework. This figure represents the ROC curve of the model evaluated on the internal dataset, where the x-axis indicates specificity and the y-axis indicates sensitivity.

연세대학교
YONSEI UNIVERSITY

4. DISCUSSION

Recent technological advancements have led to the emergence of multimodal approaches aimed at addressing the limitations of classifiers and predictors that rely solely on single modalities. While such initiatives have been undertaken in cancers such as breast, lung, and ovarian cancer, similar efforts in colorectal cancer remain limited. Furthermore, past analyses have often relied on a restricted scope of data, primarily focusing on DNA panels and specific blood biomarkers, resulting in significant gaps in the research. Additionally, studies utilizing single modalities have generally produced lower area under the curve (AUC) values, which suggests limited predictive power compared to multimodal approaches. These limitations highlight the need for comprehensive analyses that integrate various data types to enhance the robustness and reliability of predictive models for colorectal cancer. Moreover, tools based on specific modalities face a significant drawback as they become unusable when there are missing modalities or missing features. This unmet demand underscores the necessity for a multimodal-based immune checkpoint inhibitor response predictor in colorectal cancer that can provide response predictions even in the presence of missing modalities and features.

Our study addresses these challenges by collecting a comprehensive range of multimodal data from colorectal cancer patients undergoing immune checkpoint inhibitor therapies, leading to the development of a high-performance treatment response prediction model. We aimed to distinguish responders from non-responders with high accuracy by comparing these groups across each modality. Our analysis of the combined modalities revealed notable improvements across all five metrics used to evaluate classification performance. Our findings demonstrate significant novelty from several perspectives. First, we developed the highest-performing multimodal classifier for predicting responses to immune checkpoint inhibitors globally, addressing a critical gap in current research. Second, our results indicate that combining various modalities in colorectal cancer allows for more accurate classification of responder groups compared to predictions made using single-modal approaches. Third, the ability of our combined modality dataset—despite being minimal and containing missing modalities and features in real-world data—to provide useful prediction scores is particularly valuable. Fourth, we demonstrate that biomarkers identified within single modalities can still hold substantial utility.

연세대학교
YONSEI UNIVERSITY

While a study published in Nature Cancer in 2022 reported relatively high AUC values for unimodal approaches in non-small cell lung cancer, a global effort to develop a unimodal classifier for immune checkpoint inhibitor response in colorectal cancer has not yet been undertaken. Moreover, unimodal models often experience trade-offs between different metrics, as illustrated by significant gaps among AUC, recall, F1 score, specificity, and precision. In contrast, our integrated multimodal approach led to improved scores across accuracy, sensitivity, positive predictive value (PPV), specificity, and negative predictive value (NPV) metrics. This indicates that while unimodal models may demonstrate high sensitivity, they often lack precision; conversely, our final model effectively mitigated this trade-off.

Furthermore, as cancer arises from complex biological mechanisms, a diverse omics approach is essential for achieving more accurate classifications. This study represents the first global effort to develop a multimodal classifier for predicting responses to immune checkpoint inhibitors in colorectal cancer, showcasing commendable performance and enhancing the potential for personalized treatment strategies in this field.

However, we recognize certain limitations, including a relatively small sample size, data imbalances, algorithm simplicity, and validation requirements. Despite this, we addressed issues related to sample size and data imbalance by implementing data up-sampling within the training set. We also sought to regulate overfitting through stratified k-fold cross-validation. Although we could not use a full modality combined dataset for validation, we intend to validate our model's generalization performance using two independent external datasets that combine up to four modalities. Currently, our research applies ensemble techniques involving logistic regression, random forest, Naïve Bayes, and XGBoost algorithms, and we are also exploring the potential of attention models to enhance our predictions.

To address the limitations of data availability, we employed three validation methods to emphasize our model's robustness and generalizability. First, we validated the performance of individual modality features based on independent single-modality data. While this may not showcase the synergistic effects of all modalities, it confirmed the importance of the features we've utilized in our model. Second, we compared our model's performance against existing single-modality immune checkpoint inhibitor response predictors, such as TIDE and IMPRES, which provide predictions based on RNA signatures. This comparison illustrated how our model performs relative to existing tools. Lastly, to minimize the risk of overfitting, we conducted stratified k-fold cross-validation,

연세대학교
YONSEI UNIVERSITY

which is particularly valuable given the significant challenges posed by data imbalance.

By addressing these limitations, we aim to meet the unmet needs related to colorectal cancer immune checkpoint inhibitor treatment by developing a multimodal-based predictor that remains valuable even in scenarios involving missing modalities and features. Our results emphasize the study's importance on a global scale, highlighting its potential utility and contribution to addressing clinical unmet needs in this area.

연세대학교
YONSEI UNIVERSITY

5. CONCLUSION

In conclusion, our study presents the first multimodal classifier specifically designed to predict responses to immune checkpoint inhibitors in colorectal cancer. The integration of diverse data types has enhanced the robustness and reliability of our predictive model, which addresses the critical limitations observed in previous single-modal studies. By demonstrating improved performance across various metrics, our work significantly advances the field of colorectal cancer treatment, offering new insights into the potential of multimodal approaches.

Our findings underscore the importance of developing predictive models that can function effectively even in the presence of missing modalities and features, thereby meeting the clinical need for accurate treatment response predictions. The ability of our model to achieve high accuracy while addressing these common challenges emphasizes its utility in real-world settings, where data completeness cannot always be guaranteed.

Despite acknowledging certain limitations, such as a relatively small sample size and data imbalances, we have implemented strategies to mitigate these issues, including data up-sampling and comprehensive validation methods. This ensures that our model remains resilient and applicable to diverse patient populations.

Overall, this research not only fills a significant gap in the current understanding of colorectal cancer treatment but also lays the groundwork for future studies aiming to refine and expand multimodal predictive modeling in oncology. By leveraging this innovative approach, we hope to contribute to the advancement of personalized treatment strategies that can improve outcomes for patients with colorectal cancer.

연세대학교
YONSEI UNIVERSITY

References

1. Sawicki, T, Ruszkowska, M, Danielewicz, A, Niedźwiedzka, E, Arłukowicz, T, Przybyłowicz, K. E (2021). A review of colorectal cancer in terms of epidemiology, risk factors, development, symptoms and diagnosis. Cancers, 13(9), 2025.
2. Marcus, L, Lemery, S. J, Keegan, P, Pazdur, R (2019). FDA approval summary: pembrolizumab for the treatment of microsatellite instability-high solid tumors. Clinical Cancer Research, 25(13), 3753-3758.
3. Lin, K. X, Istl, A. C, Quan, D, Skaro, A, Tang, E, Zheng, X (2023). PD-1 and PD-L1 inhibitors in cold colorectal cancer: challenges and strategies. Cancer immunology, immunotherapy, 72(12), 3875-3893.
4. Mulet-Margalef, N, Linares, J, Badia-Ramentol, J, Jimeno, M, Sanz Monte, C, Manzano Mozo, J. L et al (2023). Challenges and therapeutic opportunities in the dMMR/MSI-H colorectal cancer landscape. Cancers, 15(4), 1022.
5. Adam, G, Rampášek, L, Safikhani, Z, Smirnov, P, Haibe-Kains, B, Goldenberg, A (2020). Machine learning approaches to drug response prediction: challenges and recent progress. NPJ precision oncology, 4(1), 19.
6. Vamathevan, J, Clark, D, Czodrowski, P, Dunham, I, Ferran, E, Lee, G et al (2019). Applications of machine learning in drug discovery and development. Nature reviews Drug discovery, 18(6), 463-477.
7. Ali, M, Aittokallio, T (2019). Machine learning and feature selection for drug response prediction in precision oncology applications. Biophysical reviews, 11(1), 31-39.
8. Dara, S, Dhamercherla, S, Jadav, S. S, Babu, C. M, Ahsan, M. J (2022). Machine learning in drug discovery: a review. Artificial intelligence review, 55(3), 1947-1999.
9. Vanguri, R. S, Luo, J, Aukerman, A. T, Egger, J. V, Fong, C. J, Horvat, N et al (2022). Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L) 1 blockade in patients with non-small cell lung cancer. Nature cancer, 3(10), 1151-1164.
10. Captier, N, Lerousseau, M, Orlhac, F, Hovhannisyan-Baghdasarian, N, Luporsi, M, Woff, E et al (2025). Integration of clinical, pathological, radiological, and transcriptomic data improves prediction for first-line immunotherapy outcome in metastatic non-small cell lung cancer. Nature communications, 16(1), 614.
11. Simmonds, M. C, Higgins, J. P (2016). A general framework for the use of logistic regression models in meta-analysis. Statistical methods in medical research, 25(6), 2858-2877.
12. Chen, S, Zhou, Y, Chen, Y, Gu, J (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics, 34(17), i884-i890.
13. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997.
14. Dobin, A, Davis, C. A, Schlesinger, F, Drenkow, J, Zaleski, C, Jha, S et al (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29(1), 15-21.
15. Van der Auwera, G. A, O'Connor, B. D (2020). Genomics in the cloud: using Docker, GATK, and WDL in Terra. O'Reilly Media.
16. Langmead, B, Salzberg, S. L (2012). Fast gapped-read alignment with Bowtie 2. Nature methods, 9(4), 357-359.
17. McLaren, W, Gil, L, Hunt, S. E, Riat, H. S, Ritchie, G. R, Thormann, A et al (2016). The

연세대학교
YONSEI UNIVERSITY

ensembl variant effect predictor. Genome biology, 17, 1-14.
18. Anders, S, Pyl, P. T, Huber, W (2015). HTSeq—a Python framework to work with high-throughput sequencing data. bioinformatics, 31(2), 166-169.
19. Zhang, Y, Parmigiani, G, Johnson, W. E (2020). ComBat-seq: batch effect adjustment for RNA-seq count data. NAR genomics and bioinformatics, 2(3), lqaa078.
20. Love, M. I, Huber, W, Anders, S (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology, 15, 1-21.
21. Guinney, J, Dienstmann, R, Wang, X, De Reynies, A, Schlicker, A, Soneson, C et al (2015). The consensus molecular subtypes of colorectal cancer. Nature medicine, 21(11), 1350-1356.
22. Wood, D. E, Lu, J, Langmead, B (2019). Improved metagenomic analysis with Kraken 2. Genome biology, 20, 1-13.
23. Lu, J, Breitwieser, F. P, Thielen, P, Salzberg, S. L (2017). Bracken: estimating species abundance in metagenomics data. PeerJ Computer Science, 3, e104.
24. Cao, Y, Dong, Q, Wang, D, Zhang, P, Liu, Y, Niu, C (2022). microbiomeMarker: an R/Bioconductor package for microbiome marker identification and visualization. Bioinformatics, 38(16), 4027-4029.

연세대학교
YONSEI UNIVERSITY

Abstract in Korean

임상, 유전체, 이미징 및 병리학적 다중 모달리티 데이터를 기반으로 한 대장암 면역항암요법 반응 예측기 개발

대장암(CRC)은 전 세계에서 세 번째로 흔한 암으로, 전체 암 발병률의 약 10%를 차지하며, 암 관련 사망의 두 번째 주요 원인이다. 이 중 미스매치 수리 결핍과 미세위성 불안정성이 높은 유형의 대장암은 면역항암요법, 특히 FDA에서 승인한 면역관문억제제에 긍정적인 반응을 보이지만, 미세위성 불안정성이 낮은 타입은 면역관문억제제에 저항성을 나타내며, 반응을 보이는 약제가 뚜렷하지 않아 여전히 약제 선택에 대한 미충족 수요가 남아 있다. 이러한 문제를 해결하기 위해 최근에는 임상 정보, 유전체 데이터, 영상 이미지 등 다양한 데이터 유형을 통합하여 예측 정확도를 향상시키고 최적의 의사결정을 돕기 위한 노력이 이루어지고 있다.

이번 연구는 세브란스 병원에서 면역항암제 투여 환자 106명(반응군 36명, 비반응군 70명)으로부터 샘플을 수집하여 차세대 유전자 시퀀싱, 메타게놈 시퀀싱, 임상 데이터, 병리학 및 영상 데이터를 활용했다. 이들을 통합하여 대장암을 위한 다중 모달리티 기반 면역항암요법 반응 예측기를 개발하는 것을 목표로 했다. 이를 통해 단일 모달리티 특징만으로는 반응자와 비반응자를 정확하게 구분하기에 충분하지 않다는 것을 확인하였으나, 다중 모달리티를 통합하면 시너지 효과를 생성하여 분류 정확성을 크게 향상시킬 수 있음을 입증하였다.

특히, 랜덤 포레스트, 로지스틱 회귀, 나이브 베이즈, XGBoost와 같은 기계 학습 기반의 앙상블 모델을 활용하여 정확도, 민감도, 특이도, 긍정적 예측 가치(Positive Predictive Value, PPV), 부정적 예측 가치(Negative Predictive Value, NPV)에서 개선된 성과를 확인했다. 개발한 예측기는 누락된 모달리티나 모달리티 내에서 누락된 특징이 있는 경우에도 반응 예측 점수를 제공하는 중요한 장점을 갖춘다. 이러한 견고함은 완전한 데이터가 항상 사용 가능하지 않은 임상 환경에서의 적용

연세대학교
YONSEI UNIVERSITY

가능성을 높일 수 있다. 본 모델의 일반화 성능은 계층화된 k-겹 교차 검증, 독립적인 외부 검증 세트, 단일 모달 면역 체크포인트 억제제 반응 예측기와의 비교를 통해 검증했다. 이번 연구는 대장암에서 면역 체크포인트 억제제 반응을 예측하기 위한 다중 모달리티 기반 모델 개발의 선구적인 노력을 나타내며, 모달리티 조합에 따라 예측 성능이 점진적으로 향상되는 것을 보여주어 치료 전략 및 환자 결과 개선의 잠재력을 제공한다.

핵심되는 말 : 대장암, 면역항암제 반응 예측 바이오마커, 멀티모달 결합, 기계학습