Machine learning-based meta-analysis of colorectal cancer and inflammatory bowel disease

by

A Project Report submitted to the

School of Graduate Studies

in partial fulfillment of the

requirements for the degree of

Master of Science

Supervisor: Dr. Hamid Usefi

Department of Computer Science

Memorial University of Newfoundland

Jun 2024

St. John's Newfoundland

Contents

List of Figures

List of Tables

List of Acronyms

CRC Colorectal Cancer

ML Machine Learning

SVFS Singular-Vectors Feature Selection

IBD Inflammatory Bowel Disease

eoCRC Early-Onset Colorectal Cancer

GEO Gene Expression Omnibus

STRING Search Tool for the Retrieval of Interacting Genes/Proteins

PLOS Public Library of Science

TCGA The Cancer Genome Atlas

WGCNA Weighted Gene Co-expression Network Analysis

DEG Differential Expression Analysis

RF Random Forest

SVM Support Vector Machine

DT Decision Tree

KNN K-Nearest Neighbors

PCA Principal Component Analysis

SMOTE Synthetic Minority Oversampling Technique

AUC Area Under the Curve

GO Gene Ontology

GSEA Gene Set Enrichment Analysis

COSMIC Catalogue Of Somatic Mutations In Cancer

Abstract

Colorectal cancer (CRC) is one of the leading causes of cancer-related death worldwide. Despite extensive research efforts, the mechanism of CRC remains poorly understood, and genetic biomarkers discovered thus far have not provided proper insight into the dynamics of CRC. One reason might be that most analysis methods perform univariate analyses and do not investigate the combination of genes that lead to disease. To fill this gap, we employ SVFS (Singular-Vectors Feature Selection), as well as several other machine learning algorithms, to identify genes associated with CRC. We developed an ensemble classifier model using identified genes to validate our findings and distinguish CRC tumour samples from adjacent normals. We validated our findings on 13 independent datasets and achieved significant results on all of them (correctly diagnosing 1755 cases out of 1807 and 115 controls out of 119). Several identified genes by our methodology have previously been reported to be associated with CRC, while other genes are novel and should be further researched. Furthermore, the same pipeline was applied to. Inflammatory Bowel Disease (IBD) since patients with IBD are at substantial risk of developing CRC. Following significant results on validation sets of IBD using identified genes (correctly 212 IBD cases out of 231 and 51 healthy controls out of 54), we examined IBD-related genes in conjunction with CRC-related genes to gain a better insight into suspected genes. A Python implementation of our pipeline can be accessed publicly at `https://github.com/AriaSar/CRCIBD-ML`.

Chapter 1

Introduction

Colorectal cancer (CRC) is one of the top three deadliest cancers worldwide, with an estimated 1.8 million cases and 881,000 fatalities in 2018 alone [1]. Timely detection of CRC can significantly improve prognosis and reduce mortality rates [2]. When CRC is diagnosed in individuals below the age of 50, it is referred to as early-onset CRC (eoCRC). Over the past few decades, the epidemiology of eoCRC has been subject to change, as reported by numerous studies. Starting from the 1990s, there has been a rise in the incidence of eoCRC across the world, including both high- and low-income countries [3, 4]. The rate of increase in eoCRC incidence is accelerating and is predicted to pose a significant public health challenge [3, 4]. Recently, the US Preventive Services Task Force recommended lowering the average-risk population screening age to 45 years [5, 6]. Possible justifications for the increasing incidence of eoCRC include a westernized diet, including red and processed meats; consumption

of monosodium glutamate, titanium dioxide, high-fructose corn syrup and synthetic dyes; obesity; stress; and widespread use of antibiotics [7].

Due to its heterogeneity, CRC is caused by many genes and environmental factors [8]. Epigenetics refers to alterations in gene expression or function without changes in DNA (the molecule that carries the genetic instructions used in the growth, development, functioning, and reproduction of all known living organisms) sequence. Primary epigenetic modifications include DNA methylation, post-transcriptional modifications of histone and non-coding RNA-mediated (RNA is a nucleic acid involved in protein synthesis and the transmission of genetic information from DNA to the rest of the cell) changes of gene expression [9]. Despite its significant recognition, the contribution of epigenetic events to cancer evolution needs further investigation [10, 11]. It is believed that the modifications in epigenetics and the changes in the expression of non-coding RNAs can be utilized as biomarkers for the diagnosis, prediction of treatment response and prognostication in the case of CRC [12]. The genetic and epigenetic modification of cancer-associated genes occurs independently but recurrently in CRCs, and that epigenome alterations probably control important tumour cell phenotypes, including escape from immune surveillance [13].

Recent studies have provided important insights into the molecular mechanisms that underlie the formation of CRC. Approximately 75% of CRC cases are sporadic, while the remaining cases are either linked to inflammatory bowel diseases (IBD) or have a familial origin [14]. It is estimated that the process of CRC tumorigenesis is

slow, taking almost two decades for a tumor to form [15]. Despite extensive research efforts and the elucidation of some pathways and genes, there exists a substantial lacuna in our understanding of these diseases. In particular, the dynamics and complex process of cancer cell invasion and metastasis is poorly understood [16, 17, 18].

Oncogenic transformation in CRC is known to be caused by the driver genes APC, KRAS, SMAD4, and TP53, which modulate global translational capacity in intestinal epithelial cells [19]. Given our present understanding of the intricate nature of cancer genomes, how cancer cells evolve over time under treatment, and how inhibiting targets affects the body, it is now advisable to move away from the one gene, one drug approach and embrace a 'multi-gene, multi-drug' model for making informed decisions regarding therapy [20]. In other words, the unidentified aspect of the disease may stem from the cumulative effects of multiple low-penetrance genes, which together pose a substantial risk [21]. To that end, there have been considerable interest in molecular subtype classification of CRC using gene expression data, hierarchical clustering, and machine learning [22, 20, 23, 24].

Machine learning (ML) techniques have demonstrated their efficacy in addressing biological queries [25, 26]. Owing to their notable accomplishments, the application of ML methods to biological data is expanding, revealing their considerable potential in tackling problems involving genomic datasets such as the imputation of missing SNPs and DNA methylation states, disease diagnosis [25, 27], antibody development [28], and numerous other areas. The use of ML has demonstrated great potential

in enhancing our comprehension of cancer dynamics, and it holds the possibility of substantially transforming our understanding of cancer dynamics by revealing fresh insights into the molecular mechanisms that drive cancer progression and impact treatment response [29, 30].

In this thesis, our primary goal is to investigate the genetic landscape that underlies the progression of both CRC and IBD, with a specific emphasis on additive gene interactions. Fig 1.1 provides a schematic representation of the research pipeline and the various tasks executed for both IBD and CRC.

We employed novel ML algorithms trained on case-control datasets from the GEO (Gene Expression Omnibus) database, consisting of 566 CRC cases and 262 controls. Through this process, we identified a subset of genes capable of cumulatively distinguishing between CRC and control samples. To demonstrate the efficacy of our selected genes, we conducted validation using the 40 most repeated genes selected by our pipeline on multiple external and independent datasets. Our model accurately classified 1755 CRC cases out of 1807 and 115 CRC controls out of 119, highlighting the strength of our approach. Regarding IBD, the model could diagnose 212 IBD cases out of 231 and 51 healthy controls out of 54. Interestingly, creating a model using only the top genes known to cause CRC did not generate a model as accurate as the one generated, including novel genes. This indicates that further research has to be done to understand the genes causing CRC fully.

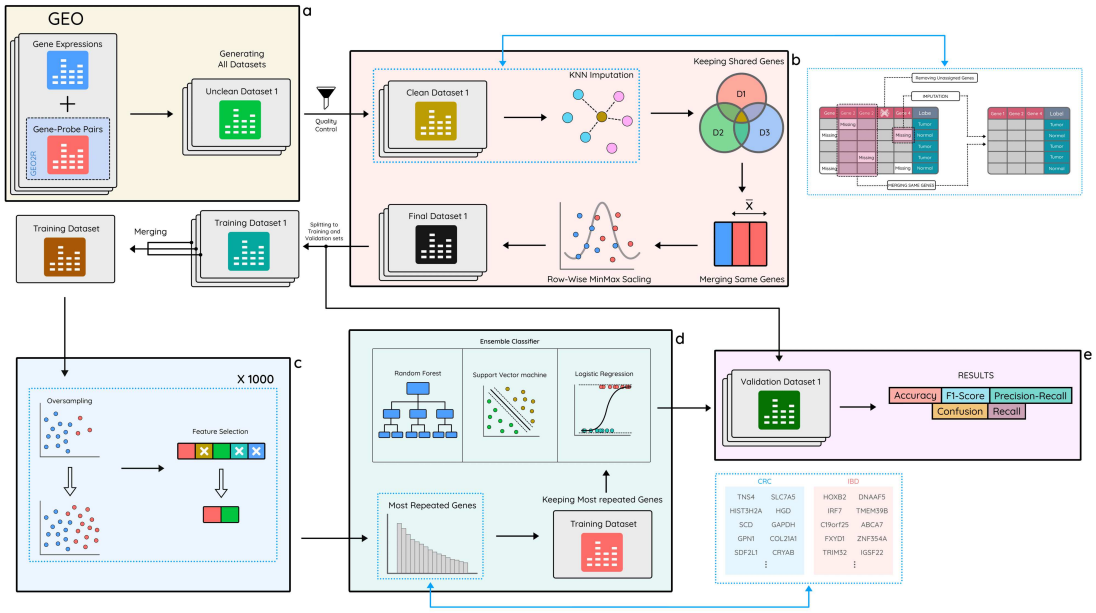

Figure 1.1: **Schematic representation of the research workflow.** **(a)** Raw datasets are retrieved from GEO, and tabular datasets are generated utilizing gene expression data and probe-gene mapping. **(b)** Data processing steps are performed, including discarding unassigned genes, imputing missing values, removing non-common genes, combining identical genes, and scaling each dataset. **(c)** After splitting datasets into training and validation sets and merging training sets to form a single set, a 1000-iteration oversampling/feature selection process is applied to identify the most prominent genes (genes with high contribution to CRC or IBD). **(d)** An ensemble classifier, comprising Random Forest, Support Vector Machine, and Logistic Regression, is trained on the training set. **(e)** The results are validated on the validation sets using the trained model, and four performance metrics – accuracy, F1-score, precision-recall, and confusion matrix – are employed for the evaluation of case-control sets and recall is employed for case-only sets.

Additionally, recognizing the heightened risk of CRC development in IBD patients, we also set out to identify a subset of genes capable of distinguishing between IBD cases and healthy controls. To accomplish this, we trained ML algorithms on GEO datasets comprising 288 IBD cases and 76 controls. Using the top 100 selected genes (most repeated genes after performing SVFS 1000 times), our validation on external IBD datasets led to the correct classification of 212 out of 231 IBD cases and 51 out of 54 healthy controls. We note that the misclassified samples included 9 inflamed IBD samples that were misclassified as healthy. To bridge CRC and IBD, we constructed a gene network using the STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) platform [31], revealing direct interactions between IBD and CRC genes, highlighting GAPDH's pivotal role. Our study recommends a closer examination of oncogenes TNS4, SLC7A5, and SCD within the context of the nuclear receptors meta-pathway. Furthermore, genes SLC7A5, SCD, GAPDH, and SDF2L1 are implicated in the mTOR signalling pathway (a pathway famous for its association with tumorigenesis [32]), underscoring the need for more investigation. These findings can potentially deepen our comprehension of the genetic mechanisms underlying CRC and IBD. The work done in this thesis has been published in the PLOS ONE journal [33].

Chapter 2

Background

Dorani et al. (2018) proposed a multi-variant method to find new risk variants by deploying two ensemble algorithms of random forest and the gradient boosting machine [34]. They utilized six different feature selection algorithms and eventually selected Tuned ReliefF (TuRF) [34]. After performing statistical interaction analysis, 17 pairwise and 16 three-way interactions were found [34]. Furthermore, two new genes have been identified as suspected CRC genes, in addition to four known identified genes ARRDC5, DCC, ALK, and ITGA1 [34]. Ding et al. (2019) used ML algorithms to identify CRC biomarkers that can be detected in blood, urine, or saliva and then created a classification model [35]. Researchers identified three genes ESM1, CTHRC1, and AZGP1 by creating three classifier models to determine which proteins result from gene expression in blood, urine, and saliva [35]. Then they made classifier models using six different algorithms [35]. Zhao et al. (2019) introduced

a three-module method for classifying CRC patients from controls [36]. In the first module, minimum redundancy maximum relevance (mRMR) was used to decrease the number of studied genes [36]. In the second module, they tried to tackle the unbalanced classes problem using the RUSBoost algorithm [36]. And in the third module, they implemented the mixed kernel function (MKF) based support vector machine (SVM) model for classifying patients and controls (MKF-SVM) [36]. The Whale Optimization Algorithm (WOA) was applied to find appropriate parameters for the model [18]. They managed to find 13 novel genes [36]. Finally, the classifier showed a geometric mean of 93.65%, which was better than other models proposed in previous similar studies [36].

Su et al. utilized gene expression data from The Cancer Genome Atlas (TCGA) and applied the Weighted Gene Co-expression Network Analysis (WGCNA) along with Differential Expression Analysis (DEG) [37]. They then performed feature selection to pinpoint the genes that are most closely correlated with the disease [37]. For classifying cases from controls and predicting the stages of cancer, they employed three classification techniques: Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT) [37]. Among these, RF demonstrated superior performance, achieving an accuracy of 99.88%, an F1 score of 0.9968, and a recall rate of 99.5% [37]. Furthermore, when diagnosing stages I through IV of colon cancer, it showed an accuracy of 91.5%, an F1 score of 0.7679, and precision of 86.94% [37]. However, a limitation of their study was the use of a single dataset and reliance on cross-

validation for evaluation rather than leveraging multiple datasets for training and validation [37]. Moreover, the sequence of using WGCNA and DEG prior to feature selection and classification might negatively impact the detection of additive genes [37]. The genes they identified as most significant in their study included GCNT2, GLDN, SULT1B1, UGT2B15, PTGDR2, GPR15, BMP5, and CPT2 [37].

Maurya et al. also employed DEG in conjunction with machine learning techniques to identify genes correlated with CRC [38]. They conducted DEG and machine learning processes in parallel, subsequently utilizing the intersecting genes identified by both methods for in-depth analysis [38]. Their research highlighted TMEM236 as a potential novel biomarker for CRC diagnosis [38]. However, similar to the approach taken by Su et al., there is an inherent risk in overlooking additive genes when relying solely on the intersection of DEG and genes pinpointed by feature selection [38]. Additionally, the use of distinct validation sets would have likely improved the study's robustness [38].

Chapter 3

Methodology

3.1 Data collection

3.1.1 Gene expression

Gene expression in a cell can be compared to how a library works. A cell has many genes, similar to how a library has many books. Each gene has instructions for making a specific protein, just like each book has information on a specific topic. In a cell, special proteins called transcription factors act like librarians. They choose which genes are turned on like librarians choose which books to show. When a gene is turned on, it's like reading a book. This is called transcription, where the gene's DNA is copied into RNA. Then, the cell uses some of these RNAs to make proteins, similar to using a recipe from a book to cook a dish. This is called translation. Also, different cells use different genes, just like different sections in a library have different kinds

of books. Bioinformatics data obtained from technologies such as microarrays and RNA-seq to study gene expression by measuring RNA levels [39]. In microarrays, DNA probes on a surface interact with RNA from a sample that's converted into cDNA (a form of DNA synthesized from a messenger RNA (mRNA) template in a process called reverse transcription, used especially in cloning or when studying gene expression) [39]. It's important to note that not all genes are transcribed into mRNA; there are various types of RNA, each with distinct roles in the cell. The level of fluorescence shows how much RNA is present [39]. RNA-seq gives more detail by sequencing all the cDNA, eliminating the need for specific probes [39]. Both methods produce large datasets showing gene expression in different conditions [39]. These datasets are arranged in gene expression matrices, comparing expression levels between samples from different conditions [39]. We can visualize these matrices as heatmaps, which help analyze gene relationships and understand biological processes [39]. In our research, we use biopsy samples instead of blood samples to study gene expression. Biopsy samples are better for looking at specific genes in a certain tissue, but getting these samples involves a small surgery. This can be hard depending on where the tumor is and how the patient is doing. Blood samples are easier to get and safer for the patient. But, they might not show all the different gene changes in the tumor like biopsies do. We chose biopsies because they give us a clearer picture of the genes in the tumor, which is important for our study.

3.1.2 Colorectal cancer

The datasets used in this study are derived from the GEO (Gene Expression Omnibus) database. Table 3.1 provides details for each colorectal cancer dataset. We combined 6 gene expression datasets from the GEO (Gene Expression Omnibus) database to form a training dataset, and an additional 13 different gene expression datasets were selected for validation. Some of the validation datasets contain only cases. All datasets in this study consist of biopsy samples, and no blood samples are included. We thoroughly examined all the validation datasets to ensure there was no data leakage. Data leakage occurs when information from outside the training dataset is used to create the model, leading to overly optimistic performance estimates that may not generalize well to new data. For instance, dataset GSE32323 was omitted due to a high probability of containing identical patients (with differing expressions) as those in dataset GSE21510. Additionally, to enhance reliability, gene expression samples were grouped based on geographical similarity (country/city) within either the training or validation sets as much as possible. It is important to note that datasets GSE68468, GSE103512 and GSE2109 encompass samples derived from various organs in addition to the colon and rectum; however, in our analysis, we only included samples derived from the colon and rectum. To be able to merge the training datasets together and then perform the validation, we only kept genes that are common between all training and validation datasets (10113 and 16413 genes for CRC and IBD, respectively).

3.1.3 Inflammatory bowel disease

We selected 5 gene expression IBD datasets from GEO for training and 6 different gene expression IBD datasets for validation. We included only those samples who had not undergone any specific treatment. Additionally, we excluded datasets consisting of blood samples. Dataset GSE16879 includes pre- and post-infliximab treatment samples, of which we selected only the pre-treatment ones. In GSE59071 and GSE48958, inactive samples were excluded. From GSE179285, we included only inflamed samples and from GSE4183, only IBD and normal samples were chosen. From GSE37283, patients diagnosed with ulcerative colitis with neoplasia were retained. Table 3.2 contains details of the datasets used for IBD. After discarding genes that are not common between all IBD datasets, all IBD datasets had uniformly 16,413 genes.

3.2 Data preprocessing

3.2.1 Structuring the raw data

The datasets used in this study are derived from the GEO (Gene Expression Omnibus) database. Datasets available on the GEO website are unsuitable for analyzing or machine learning purposes. To generate a suitably formatted dataset (.CSV, .pickle, feather, etc.), Series Matrix File (.TXT) was used. This file contains gene expression data for each probe (Each probe corresponds to a specific gene, and multiple probes can correspond to the same gene). Using Python language, a script was written for

converting Series Matrix File to the pickle format.

3.2.2 Data imputation

Many machine learning algorithms rely on complete data to work effectively, and missing values can pose problems for their function. Furthermore, incomplete data can lead to biased results or inaccurate predictions and decreased accuracy. Before applying most machine learning algorithms, such as some of those used in this research, we must either eliminate incomplete observations from the dataset or impute missing values. The datasets used in this research contain several missing values. By removing a column containing a missing value, we risk losing a critical disease-related gene. Also, removing a row with missing values weakens feature selection and the classifier model due to the low number of samples. In order to fill in these missing values, we employed the KNN (K-Nearest Neighbors) imputation algorithm [68] (number of neighbours=5). It is worth mentioning that columns with more than 5% missing values were discarded.

3.2.3 Mapping probes to genes

In order to convert each probe to its corresponding gene, we need an annotation table. GEO2R was used to retrieve this annotation table. As mentioned earlier, numerous probes are assigned to the same genes after matching probe IDs with corresponding genes from the GEO2R mapping file. To integrate these probes into a single gene,

the mean gene expression values were utilized. Then, in order to integrate training datasets and execute uniform validation on validation sets, we maintained common genes across all datasets. Consequently, 10,113 and 16,413 genes were retained for CRC and IBD analyses, respectively.

3.2.4 Scaling

There are significant differences between the gene expression ranges of the datasets. This results in poor performance of the classifier model. To determine the most appropriate scaler, several different scalers were applied to the data. Principal component analysis (PCA) was used to shrink the data dimension and plot each scaler's results. Figure 3.1 illustrates how row-wise MinMax normalization and quantile normalization result in better scattering of CRC datasets. This enables the model to capture the underlying data pattern of genes' contributions more effectively. As different genome datasets have different ranges, row-wise MinMax normalization has an advantage over a quantile transformer when a new dataset or a sample needs to be classified. In addition, it maintains gene expression correlation as well as transforms gene expression into a 0 to 1 range for each instance. Thus, row-wise MinMax scaling was performed after the feature selection of all datasets. IBD datasets were also transformed using row-wise MinMax scaling.

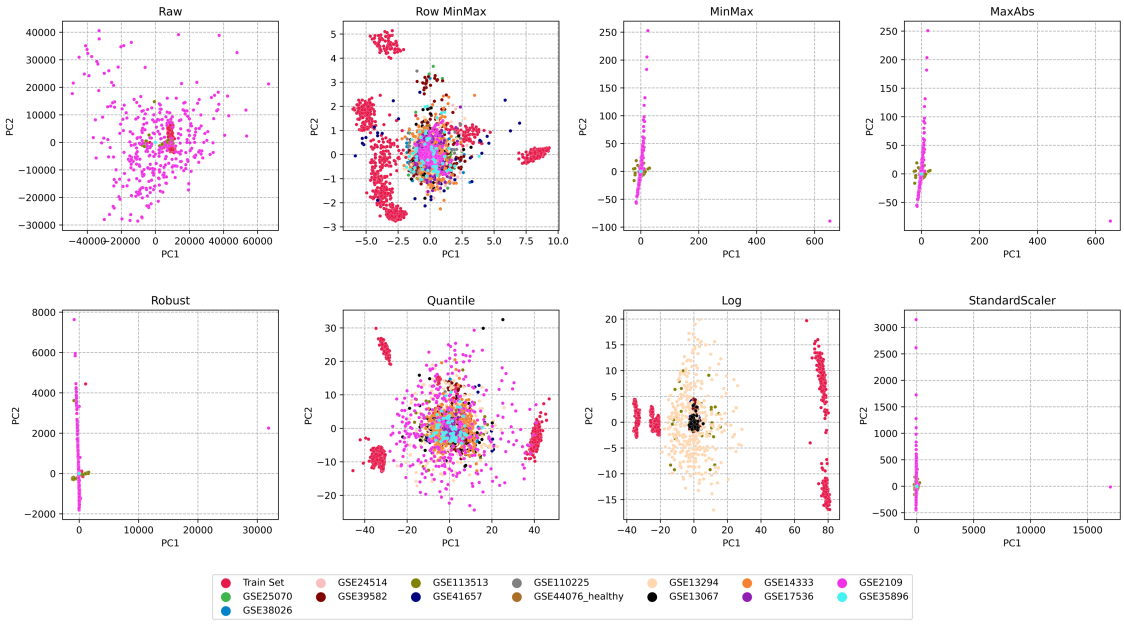

Figure 3.1: **Effect of different scalers on CRC training dataset and test datasets.**

3.2.5 Balancing the training set

Upon combining the training datasets to get a single dataset, the high ratio of cases to controls leads to poor performance for both feature selection and the classification model. To overcome this problem, we used SMOTE [69] (synthetic minority oversampling technique), an efficient oversampling approach, to balance the number of cases and controls in the training dataset.

3.3 Feature selection

Feature selection is probably the most critical part of this research. Finding the most disease-relevant genes among thousands of them is done by feature selection. In order to find additive genes associated with CRC and IBD, wrapper or hybrid feature selection techniques should be employed. However, wrapper methods cannot be implemented with high-dimensional datasets, like those used in this study, due to their computational complexity. In this study, we applied SVFS (Singular-Vectors Feature Selection), a hybrid feature selection method that demonstrated significant results recently compared to other feature selection methods on gene expression data [70, 71]. SVFS is a method designed for high-dimensional datasets. Given a matrix A with its Moore-Penrose pseudo-inverse $A^{\dagger}$, it is shown in [70, 71] that the projector $P_A = I - A^{\dagger}A$ partitions features into clusters based on their correlations. Initially, SVFS identifies and retains only those features that correlate with the class label, discarding others as irrelevant. In the subsequent step, it further clusters the remaining features and selects the most significant ones from each cluster. We used parameters suggested in the SVFS paper for biological data as parameters for our pipeline (Th_{irr}=3, Th_{red}=4, α=50, β=5, and k=100) [70, 71] (Th_{irr} used to filter out irrelevant features, Th_{red} eliminates redundant features, α sets the maximum size for feature clusters, β determines the number of top features selected from each sub-cluster, and k is the total number of features to select in the process).

When running the SVFS feature selection algorithm, it may return a different

subset of features each time. Hence, to obtain robust results and be sure that the genes found are related to the disease, we repeated the algorithm 1000 times for each disease and continued the research using the top 100 most frequently repeated genes. Oversampling is highly influential on dataset structure and, as a result, might drastically change the subset of genes selected by the feature selection algorithm. Hence, oversampling was performed before each round of feature selection to ensure a more robust selection.

After performing oversampling-feature selection 1000 times, only the selected genes were kept for further analysis, and all other genes were discarded from all datasets.

The noteworthy point is that common genes were chosen separately for each disease, so 10113 and 16413 genes were selected for CRC and IBD, respectively.

Figure 3.2 shows the 100 most frequently repeated genes and the number of repetitions of each for colorectal cancer and inflammatory bowel diseases.

3.4 Model evaluation

To ensure the classifier's performance is due to the genes identified rather than the model and to make the most accurate prediction, we used ensemble models of Random Forest, Logistic Regression, and Support Vector Machine. Additionally, ensemble models are more robust than single models and reduce the risk of overfitting. Based on Tables 3.1 and 3.2, some of the datasets are balanced, and some others are imbalanced.

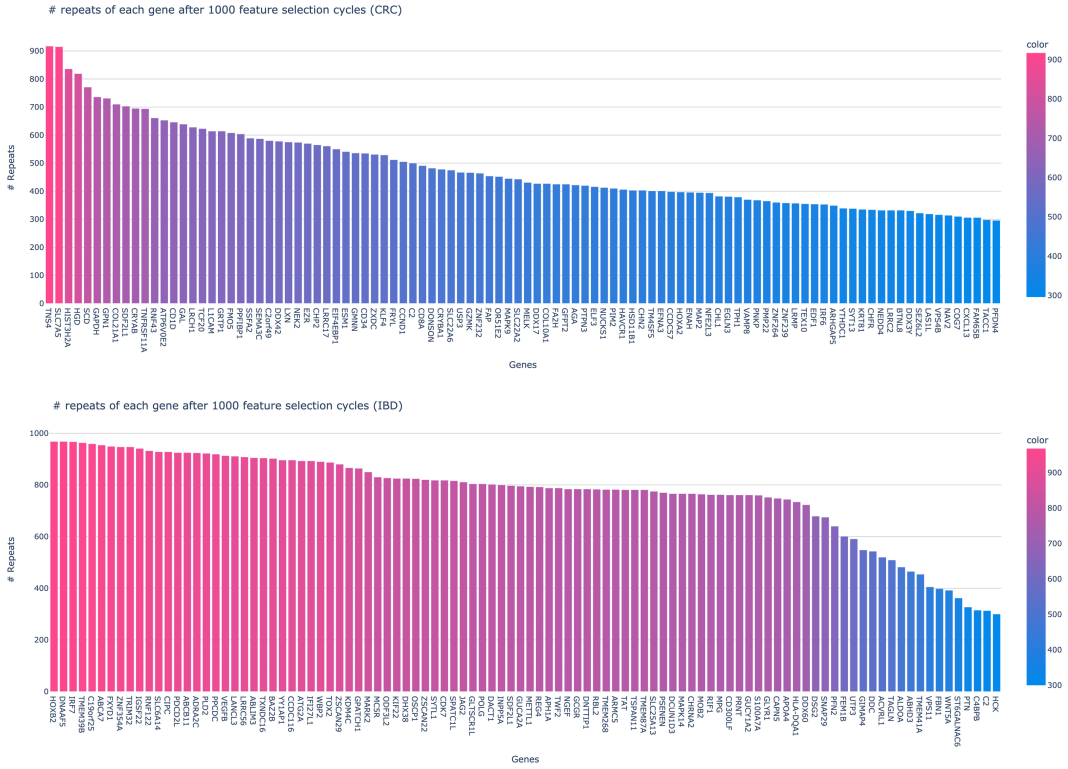

Figure 3.2: **The number of repetitions of each gene for the 100 most frequently repeated genes for CRC and IBD.**

The F1 score metric was used, as well as accuracy for imbalanced datasets. Also, confusion matrices and precision-recall curves were generated to provide better insight into model performance.

Dataset	# of Cases	# of Controls	Platform	Country/City or State	Usage	# of Probes
GSE21510[40]	123	25	GPL570	Japan/Tokyo	Training	54675
GSE44076[41]	98	98	GPL13667	Spain/Catalonia	Training	49386
GSE44861[42]	56	55	GPL3921	USA/MD	Training	22277
GSE68468[43]	186	55	GPL96	USA/MD	Training	22283
GSE89287[44]	46	17	GPL4133	Netherlands/Zuid-Holland	Training	45015
GSE103512[45]	57	12	GPL13158	USA/New York	Training	54715
GSE25070[46]	26	26	GPL6883	USA/CA	Validation	24526
GSE38026[47]	16	16	GPL11532	Germany/Kiel	Validation	33257
GSE24514[48]	34	15	GPL96	Finland/Helsinki	Validation	22283
GSE39582[49]	566	19	GPL570	France/Paris	Validation	54675
GSE113513[50]	14	14	GPL15207	China/Fujian	Validation	49395
GSE41657[51]	25	12	GPL6480	China/Beijing	Validation	41076
GSE110225[40]	17	17	GPL96	Greece/Athens	Validation	22283
GSE13294[52]	155	0	GPL570	Denmark/Aarhus N.	Validation	54675
GSE13067[52]	74	0	GPL570	Australia/Parkville	Validation	54675
GSE14333[53]	290	0	GPL570	Australia/Parkville	Validation	54665
GSE17536[54]	177	0	GPL570	USA/Nashville	Validation	54675
GSE2109[55]	351	0	GPL570	USA/Phoenix	Validation	54675
GSE35896[56]	62	0	GPL570	UK/Macclesfield	Validation	54675

Table 3.1: .

Detailed summary of CRC datasets used for training and validation.

'Cases' are tumour samples, and 'Controls' are adjacent normal samples from the same patients. All samples are taken using the biopsies. The '# of probes' column indicates the number of probe sets on the respective microarray platform.

Dataset	# of Cases	# of Controls	Platform	Country/City or State	Usage	# of Probes
GSE16879[57]	61	12	GPL570	Belgium/Leuven	Training	54666
GSE22619[58]	10	10	GPL570	Germany/Kiel	Training	54675
GSE59071[59]	82	11	GPL6244	Belgium/Leuven	Training	33252
GSE102133[60]	65	12	GPL6244	Belgium/Leuven	Training	33252
GSE179285[61]	70	31	GPL6480	USA/South San Francisco	Training	41000
GSE9452[62]	8	5	GPL570	Denmark/Copenhagen	Validation	54675
GSE36807[63]	28	7	GPL570	UK/London	Validation	54675
GSE37283[64]	11	5	GPL13158	USA/Chicago	Validation	54613
GSE4183[65]	15	8	GPL570	Hungary/Budapest	Validation	54675
GSE48958[66]	7	8	GPL6244	Belgium/Leuven	Validation	33252
GSE92415[67]	162	21	GPL13158	USA/Spring House	Validation	54613

Table 3.2: .

Detailed summary of inflammatory bowel disease datasets used for training and validation. 'Cases' are inflamed samples, and 'Controls' are samples from healthy patients. All samples are taken using the biopsies. The '# of probes' column indicates the number of probe sets on the respective microarray platform.

Chapter 4

Results

Colorectal cancer (CRC)

After running oversampling and feature selection 1000 times, we kept the first 100 most repeated genes in the training and testing datasets and removed all other genes. We performed a forward feature selection where most repeated genes were added iteratively; for instance, first, the model was trained using the most significant (frequent) gene, TNS4, and was evaluated on all test datasets. Following this, the second gene, SLC7A5, was added to the training list, and the model was trained and evaluated using these two genes. The process was repeated until the training set contained all 100 genes. We considered accuracy for balanced datasets and F1 score and accuracy for imbalanced datasets. Also, precision-recall curves were employed for a better insight into model performance. Figure 4.1 shows the testing results for each case-control

CRC dataset.

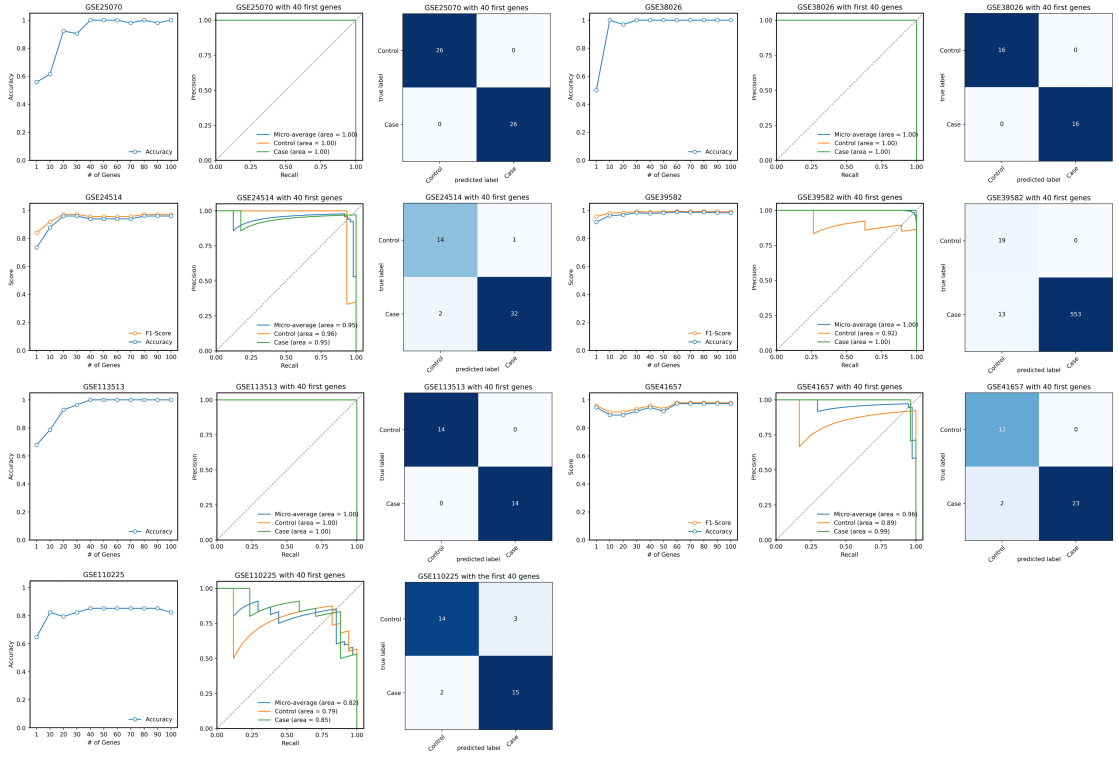

Figure 4.1: **Evaluation of identified CRC genes on independent validation sets.** Accuracy and F1-score are plotted for the different number of prominent genes utilized for training and validation. Confusion matrices and precision-recall curves (including AUC) are plotted using the first 40 prominent genes.

Several of our datasets consisted of cases only. So, we used the F1-score to report the validation results on these datasets. For case-only datasets, we adopted recall (also known as sensitivity or true positive rate), representing the proportion of correctly predicted case samples relative to the overall number of cases. Figure 4.2 illustrates the results for case-only datasets.

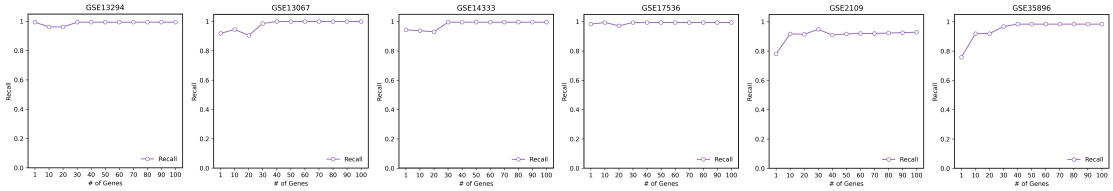

Figure 4.2: **Evaluation of the model trained on tumor and matched normal samples on case-only and control-only datasets.**

In order to achieve reliable results from ML algorithms, it should be noted that the validation datasets must not be utilized at any point during the training or model generation process. Given the validation results in Figures 4.1 and 4.2, we deduce that using 40 prominent identified genes, our model could diagnose 1755 cases out of 1807 and 115 controls out of 119. Some of the some of the most frequently-selected genes previously known to be involved in CRC are TP53 [72], APC [72], KRAS [72], MGMT [73], SMAD2 [74] and SMAD4 [74]. It is interesting to note that if we build a model just based on these well-known CRC genes, we do not get acceptable validation results. Indeed, we implemented a supplementary pipeline using only TP53, APC, KRAS, MGMT, SMAD2, and SMAD4, and it turned out that 100 controls out of 103 were misclassified, which is a poor performance (for this experiment, we had to exclude GSE38026 because KRAS gene does not exist in this dataset). The detailed results are presented in Figure 4.3.

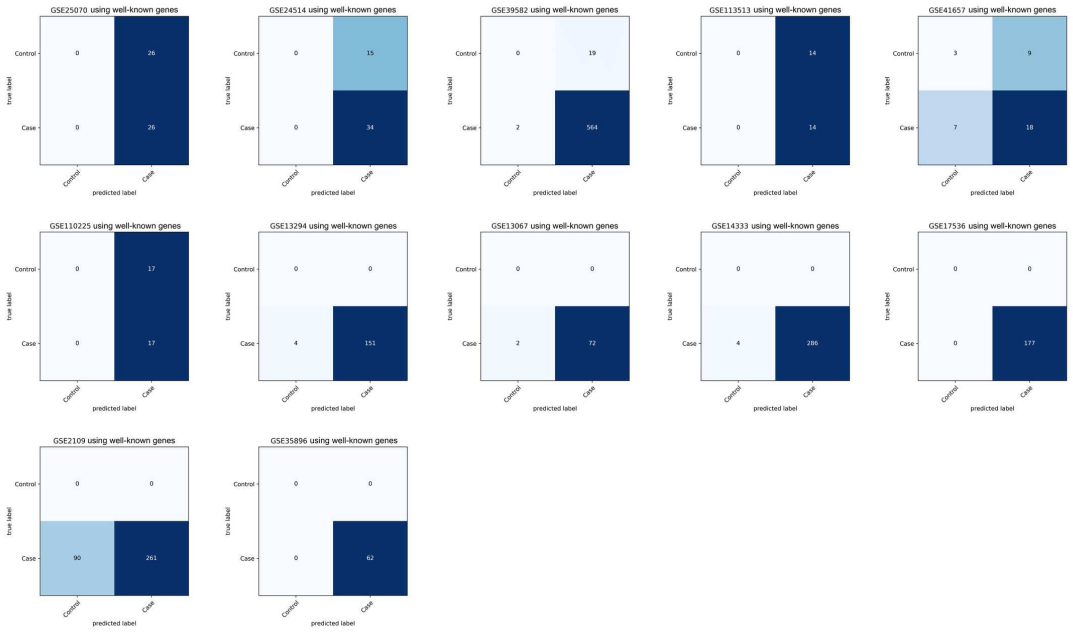

Figure 4.3: **Results of the supplementary pipeline using only TP53, APC, KRAS, MGMT, SMAD2, and SMAD4**

Inflammatory bowel disease (IBD)

The same methodology was employed for IBD, that is, SVFS was utilized on the training IBD dataset, and the most repeated 100 significant genes were selected, as shown in Table 5.3. As demonstrated in Figure 4.5, the ensemble classifier effectively distinguished inflamed samples from healthy samples in GSE9452, GSE37283, GSE4183 and GSE48958. In the case of GSE36807, the model accurately diagnosed all healthy samples, though nine inflamed samples were misclassified as healthy. Overall, the classifier exhibited strong performance, suggesting an acceptable identification of IBD-related genes by identifying 212 IBD cases out of 231 and 51 healthy controls

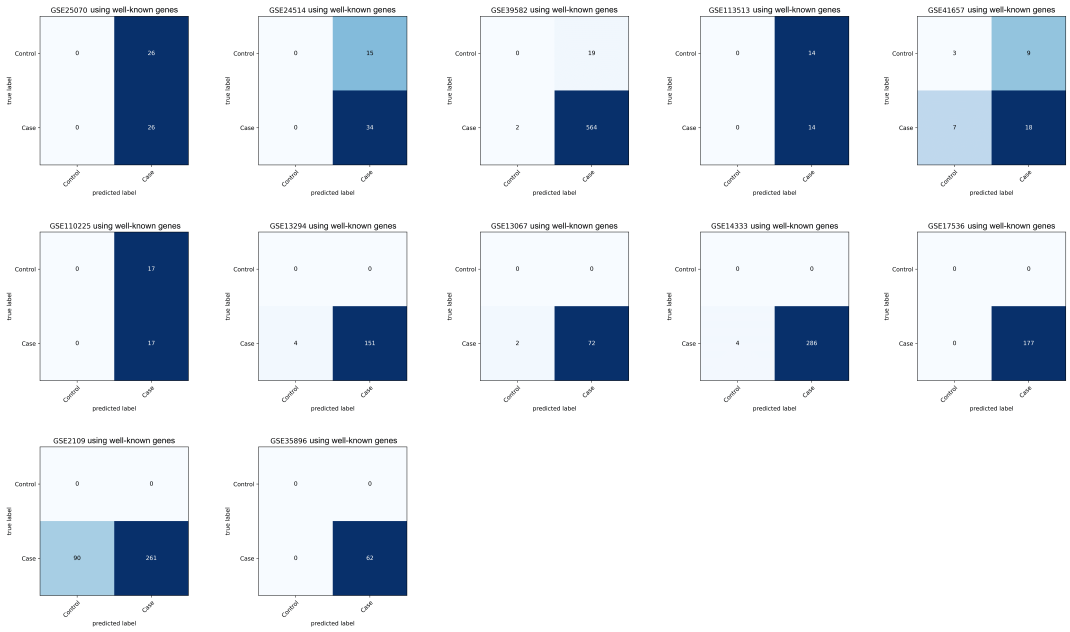

Figure 4.4: **Case-control classification results using well-known genes**

out of 54.

In this chapter, we presented the results of our study. For CRC, we used over-sampling and feature selection to determine the top 100 genes, incrementally training our model with these genes. Our results showed that by using 40 prominent genes, the model could accurately diagnose a majority of cases and controls. An experiment with well-known CRC genes revealed the importance of our feature selection process, as the model performed poorly with these genes alone. For IBD, we applied a similar approach and successfully identified the most repeated 100 significant genes. The ensemble classifier effectively distinguished inflamed samples from healthy samples in various datasets, demonstrating the successful identification of IBD-related genes.

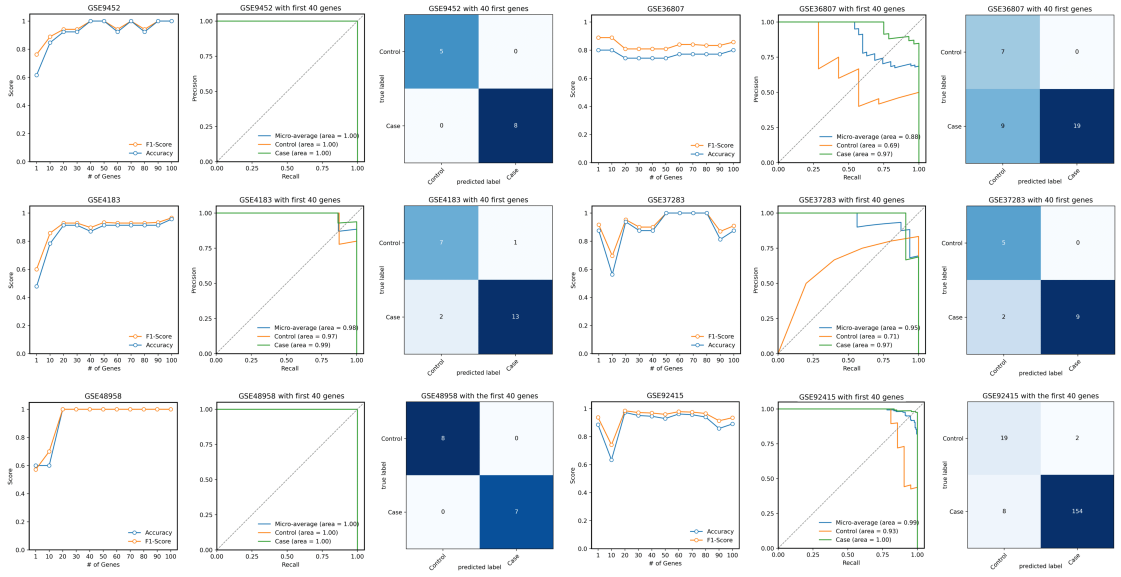

Figure 4.5: **Evaluation of identified IBD genes on independent validation sets.** Accuracy and F1-score are plotted for the different number of prominent genes utilized for training and validation. Confusion matrices and precision-recall curves (including AUC) are plotted using the first 40 prominent genes.

Overall, our study highlights the potential of machine learning in identifying key genes for CRC and IBD, contributing to improved diagnostics and understanding of these diseases.

Chapter 5

Genes and interactions

5.1 Interaction and network analysis

To find and understand the underlying interaction between identified genes, STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) [31] was employed. The STRING is an up-to-date database containing different information such as experiments, co-expressions, gene co-occurrence, gene fusion and neighbourhood. Its knowledge comes from several sources, including MINT [75], HPRD [76], BIND [77], DIP [78], BioGRID [79], KEGG [80], Reactome [81], IntAct [82], EcoCyc [83], NCI-Nature Pathway Interaction Database [84] and GO [85]. We constructed a gene network by integrating the most repeated 50 IBD genes with 50 CRC genes in the initial step, setting the interaction score to medium confidence. Figure 5.1 illustrates the resulting network.

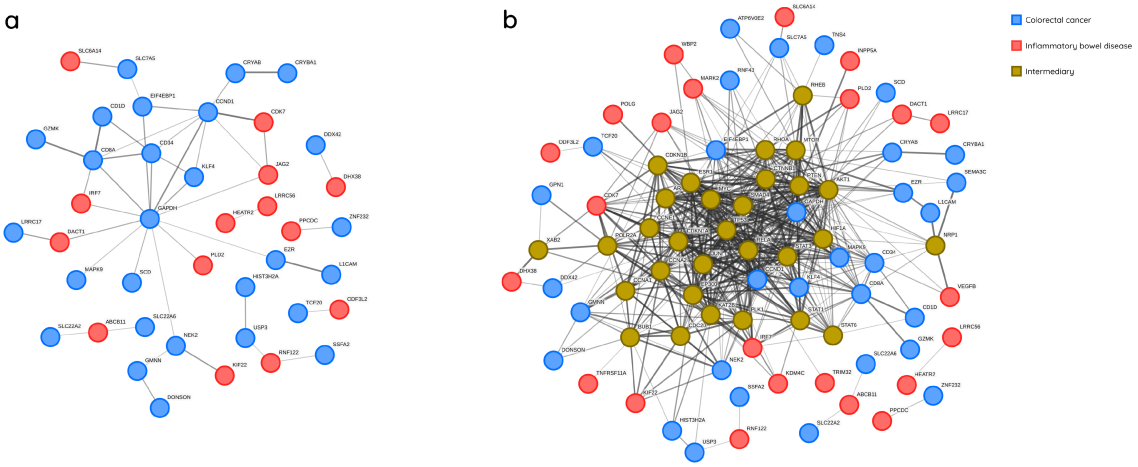

Figure 5.1: **IBD and CRC, gene interaction networks, generated by STRING for identified genes.** a Network generated based on CRC and IBD genes without the participation of intermediary genes. b Network generated based on CRC and IBD genes with the participation of intermediary genes.

As observed in Figure 5.1 (a), 13 IBD-associated genes and 27 CRC-associated genes have direct interaction (without intermediary genes). Intermediary genes are genes added by string but with the highest number of connections to identified genes by our pipeline. According to Figure 5.1, the GAPDH gene appears to play a pivotal role in linking the two gene subsets. To investigate potential interactions between CRC and IBD genes, we extended the network in Figure 5.1 (a) to include intermediary genes that may serve as a bridge between CRC and IBD genes. For this extended network in Figure 5.1 (b), we took into account only single intermediary genes (genes proposed by STRING as highly connected to our input genes), which is a drawback since the bridge could involve two genes, for example. Intermediary genes are selected

using string by counting genes with the highest number of connections to our identified genes. While single intermediary genes are more influential, other genes with minor additive effects are overlooked. We did not use two genes acting as a bridge because of the many combinations that those can create. Table 5.4 lists the full names of single intermediary genes. Owing to the network's complexity, we preserved genes in Figure 5.1 (b) with a higher number of connections in the network for illustrative purposes. For example, TP53 may be regarded as the most critical intermediary gene (This gene has the most number of connections in the network). As such, this gene and its adjacent genes might be fundamental to IBD and CRC progression. The intermediary genes are highly likely to contribute to the disease due to their strong connections to genes in our subset and, importantly, their bridging functions. We further explored the COSMIC (Catalogue Of Somatic Mutations In Cancer) database to determine if any of the genes in our subset had been previously reported as having a strong association with CRC. Tissue selection, Sub-tissue selection, Histology selection, and Sub-histology selection were set to Large intestine, Include all, Carcinoma, and Adenocarcinoma, respectively. TP53, SMAD4, RNF43, CTNNB1, and PTEN ranked among the top 20 most frequently mutated CRC-related genes listed in COSMIC. On the other hand, several of our reported significant genes were not on the COSMIC list and have not received adequate attention from researchers yet. We also performed GSEA (Gene Set Enrichment Analysis) using the most repeated 50 identified CRC genes. GSEA results (Table 5.1) showed that TNS4, SLC7A5, and SCD

are involved in the nuclear receptors meta-pathway. Genes SLC7A5, SCD, GAPDH and SDF2L1 were involved in the mTOR signalling pathway. Also, several other gene sets were associated with cell cycle regulation and transcription regulation. It must be emphasized that Figure 5.1 does not encompass all prominent genes and is derived from various experiments and prior research. Given the limited understanding of the underlying mechanisms of CRC and IBD, we propose to consider other important genes that are not part of the network in Figure 5.1. For instance, an in-depth investigation of TNS4, GAPDH, L1CAM, GAL, CRYAB, IRF7, GPN1, TMEM39B, EZR, and all other genes referred to in Tables 5.2 and 5.3 is needed.

Table 5.1: Gene Set Enrichment Analysis (GSEA) of most repeated 50 CRC-related genes

Term 1	ES	NES	NOM p-value	FDR q-value	FWER p-value	Lead Genes
mTORC1 signalling	0.896	1.733	0.001	0.006	0.005	SLC7A5, SCD, GAPDH, SDF2L1
Cell cycle	-0.605	-1.809	0.01	0.009	0.009	FAP, USP3, DONSON, CCND1, KLF4, GMNN, EIF4EBP1, EZR, NEK2
Nuclear receptors pathway	0.868	1.706	0.001	0.0111	0.012	TNS4, SLC7A5, SCD
Transcription regulator activity	-0.681	-1.623	0.040	0.027	0.031	ZNF232, CCND1, KLF4, ZXDC, GMNN
Cell cycle regulation	-0.578	-1.627	0.028	0.092	0.046	FAP, DONSON, CCND1, KLF4, GMNN, EIF4EBP1, NEK2

5.2 Review of the first ten CRC-related identified genes

In this section, we discuss the first ten most repeated CRC-related identified genes.

1. TNS4

According to Figure 3.2, TNS4 is the most repeated gene associated with colorectal cancer. In Figure 4.1, we can see that regarding the datasets GSE39582 and GSE41657, the model could make accurate decisions by just investigating this single gene. Several studies have previously reported this gene as a related gene to colorectal cancer. For example, Kim et al. proposed that TNS4 is crucial in CRC tumorigenesis, and TNS4 suppression might be a promising therapy in CRC [86]. Raposo et al. argued that the TNS4 role is critical in early-stage metastasis and, in addition, its knockdown improves sensitivity to Gefitinib [87] (medication used for different cancer types).

2. SLC7A5

SLC7A5 is the second most most repeated gene on our list. Najumudeen et al. conducted comprehensive research on SLC7A5's correlation with CRC [88]. They stated that SLC7A5 might be a potential target for treatment for KRAS-mutant colorectal cancer that is resistant to other therapies [88]. Huang et al. also included SLC7A5 as one of the five core genes contributing to the ferroptosis (a type of cell death) of colon cancer cells [89].

3. HIST3H2A

HIST3H2A is one of the novel potential biomarkers on our list. There is some evidence

that it is associated with lung cancer [90] and pancreatic cancer [91], but no studies have demonstrated its association with colorectal cancer. Therefore, further studies are needed to confirm its contribution to CRC.

4. HGD

HGD is another potential marker on our list. Yi et al. performed a study of rectal cancer tumour markers [92]. Their results demonstrated a substantial correlation between HGD and rectal cancer [92]. To our knowledge, no other studies have shown a significant correlation between HGD and colorectal cancer. As a result, this gene needs to be investigated further.

5. SCD

Cruz-Gil et al. identified SCD as a critical component of lipid metabolism in colorectal cancer (CRC) [93]. The relationship between SCD and ACSL increases the risk of relapse in CRC patients. [93]. Another study by Liao et al. suggests high SCD-1 expression is associated with advanced CRC [94].

6. GAPDH

Tang et al. also examined tumour versus non-tumour pairs among 195 cases and found considerable overexpression of GAPDH in CRCs [95]. In another study, Tarrado-Castellarnau et al. conducted research regarding GAPDH and found significant upregulation of this gene in colorectal cancer [96]. It was suggested that this gene might be helpful for early detection of CRC.

7. GPN1

To our knowledge, GPN1 is one of the novel genes identified in this study. As of yet, little is known about this gene, and more research needs to be done to understand it better.

8. COL21A1

COL21A1 is likely to be another novel biomarker. According to our research, Li et al. was the only study that considered COL21A1 as a potential diagnostic marker [97].

9. SDF2L1

Despite studies examining the association of this gene with other cancer types, such as Nasopharyngeal Carcinoma [98], this gene has not been recognized as a potential CRC marker.

10. CRYAB

CRYAB has been found to have a strong association with different types of cancer [99]. Several research studies have also examined the association between CRYAB and CRC. For instance, Deng et al. recognized CRYAB as a tumour-suppressor gene and a potential diagnostic marker [100]. Shi et al. also confirmed CRYAB as a prognostic CRC biomarker [99]. CRYAB has also been suggested as a valuable target for developing CRC treatments by Dai et al [101].

In this chapter, we explored the interactions and networks of genes associated with Colorectal Cancer (CRC) and Inflammatory Bowel Disease (IBD) using the STRING database. By integrating the most repeated 50 genes from each condition, we constructed a gene

network and analyzed the direct interactions between 13 IBD-associated and 27 CRC-associated genes. The GAPDH gene emerged as a pivotal link between the two gene subsets. We also extended the network to include intermediary genes, which may serve as bridges between CRC and IBD genes. Our analysis highlighted the importance of intermediary genes like TP53 in the progression of both diseases. Additionally, we performed Gene Set Enrichment Analysis (GSEA) on the identified CRC genes, revealing their involvement in various pathways, including the mTOR signalling pathway and cell cycle regulation.

Table 5.2: **100 first most repeated CRC genes in order. Pink cells are the first 40 genes used in classification.**

1) TNS4	2) SLC7A5	3) HIST3H2A	4) HGD	5) SCD
6) GAPDH	7) GPN1	8) COL21A1	9) SDF2L1	10) CRYAB
11) TNFRSF11A	12) RNF43	13) ATP6V0E2	14) CD1D	15) GAL
16) LRCH1	17) TCF20	18) L1CAM	19) GRTP1	20) FMO5
21) PPFIBP1	22) SSFA2	23) SEMA3C	24) C2orf49	25) DDX42
26) LXN	27) NEK2	28) EZR	29) CHP2	30) LRRC17
31) EIF4EBP1	32) ESM1	33) GMNN	34) CD34	35) ZXDC
36) KLF4	37) FRYL	38) CCND1	39) C2	40) CD8A
41) DONSON	42) CRYBA1	43) SLC22A6	44) USP3	45) GZMK
46) ZNF232	47) FAP	48) OR51E2	49) MAPK9	50) SLC22A2
51) MELK	52) DDX17	53) COL10A1	54) FA2H	55) GFPT2
56) AGA	57) PTPN3	58) ELF3	59) NUCKS1	60) PIM2
61) HAVCR1	62) HSD11B1	63) CHN2	64) TM4SF5	65) EFNA3
66) CCDC57	67) HOXA2	68) ENAH	69) MAP2	70) NFE2L3
71) CHL1	72) EGLN3	73) TPH1	74) VAMP8	75) PNKP
76) PMP22	77) ZNF264	78) ZNF239	79) LRMP	80) TEX10
81) EDF1	82) IRF6	83) ARHGAP5	84) YTHDC1	85) SYT13
86) KRT81	87) CHFR	88) NEDD4	89) LRRC2	90) BTNL8
91) DDX3Y	92) SEZ6L2	93) LAS1L	94) VPS4B	95) NAV2
96) COG7	97) CXCL13	98) FAM65B	99) TACC1	100) PFDN4

Table 5.3: **100 first most repeated IBD genes in order. Pink cells are the first 40 genes used in classification.**

1) HOXB2	2) DNAAF5	3) IRF7	4) TMEM39B	5) C19orf25
6) ABCA7	7) FXYD1	8) ZNF354A	9) TRIM32	10) IGSF22
11) RNF122	12) SLC6A14	13) CIPC	14) PDCD2L	15) ABCB11
16) ADRA2C	17) PLD2	18) PPCDC	19) VEGFB	20) LANCL3
21) LRRC56	22) ABLIM3	23) TXNDC16	24) BAZ2B	25) YY1AP1
26) CCDC116	27) ATG2A	28) IFI27L1	29) WBP2	30) TOX2
31) ZSCAN29	32) KDM4C	33) GPATCH1	34) MARK2	35) MC5R
36) ODF3L2	37) KIF22	38) DHX38	39) OSCP1	40) ZSCAN22
41) SYTL1	42) CDK7	43) SPATC1L	44) JAG2	45) GLTSCR1L
46) POLG	47) DACT1	48) INPP5A	49) SDF2L1	50) GUCA2A
51) METTL1	52) REG4	53) APH1A	54) TWF2	55) NGEF
56) GCGR	57) DNTTIP1	58) RBL2	59) TMEM268	60) ARMC5
61) TAT	62) TSPAN11	63) TMEM87A	64) SLC25A13	65) PSENEN
66) DCUN1D3	67) MAPK14	68) CHRNA2	69) MOB2	70) RIF1
71) MPG	72) CD300LF	73) PRNT	74) GUCY1A2	75) S100A7A
76) GLYR1	77) CAPN5	78) APOA4	79) HLA-DQA1	80) DDX60
81) DSG2	82) SNAP29	83) PFN2	84) FEM1B	85) UTP3
86) GIMAP4	87) DDC	88) ACVRL1	89) TAGLN	90) ALDOA
91) ABHD3	92) TMEM41A	93) VPS11	94) FBN1	95) WNT5A
96) ST6GALNAC6	97) PTN	98) C4BPB	99) C2	100) HCK

Table 5.4: **100 genes (identified using STRING) with the most number of connections to the genes identified by our pipeline**

STAT1	CCNE1	PLK1	STAT3	BUB1
SMAD4	MYC	ODF3L2	RHEB	STAT6
JUN	NRP1	CDKN1B	TP53	CCNA2
CDKN1A	CCNA1	AKT1	ESR1	POLR2A
RELA	MTOR	HIF1A	KAT2B	CDC20
RHOA	CTNNB1	XAB2	EP300	PTEN
AR	CCND3	CENPE	AURKB	SMAD2
HDAC2	HDAC1	CCNH	SKA1	SLC3A2
POLR2A	TRAF6	SP1	PCNA	EFTUD2
LGR5	CDKN2A	PRPF19	MX1	NOTCH1
EPRS	TCF3	SMAD3	MDM2	CCND2
STAT5A	CDC5L	CHUK	KIF11	SPC24
ISG15	IRF3	ALDOA	RPTOR	CDT1
CCNE2	HSP90AA1	GATA1	SLC9A1	CDK2
NDC80	RELASMAD4	CDK6	SLC9A3R1	CREBBP
SPC25	MNAT1	MYD88	CDK4	TPI1
GSK3B	CFTR	PPARA	NF2	BUB1B
RB1	COASY	YAP1	CDK9	ATF2
NUF2	SOX2	IRF1	MAML1	SF3B1
TTK	MAD2L1	CASC5	NR3C1	-

Chapter 6

Discussion

This study aimed to identify potential genes correlated with colorectal cancer (CRC) using multivariate machine learning methods. The study discovered several genes relevant to identifying CRC from tissue samples, some of which have not received enough attention in previous studies. Six independent validation sets demonstrated that a subset of 40 genes accurately diagnosed tumours and matched normal tissues. Additionally, the study investigated Inflammatory Bowel Disease (IBD), as patients with IBD are at high risk of developing CRC. We incorporated STRING by using identified CRC and IBD genes to identify potentially CRC-related genes. Several genes were identified, including previously reported causative genes and several novel ones. Machine learning methods provide a more comprehensive approach to identifying genes that may contribute to CRC development. However, this research has some limitations, including using single intermediary genes and ignoring others, missing and not examining a large number of genes that might be associated with cancer (we kept common genes across all datasets), and the risk of false positives.

Identifying novel genes correlated with CRC and IBD provides a foundation for future research into the underlying mechanisms of these diseases. Further studies focusing on genes' additive functions instead of single-variate analyses are necessary to confirm these genes' contributions. The potential significance of these findings for clinical practice includes the possibility of developing better prevention, detection, and treatment methods for CRC and IBD patients. In future studies, it is crucial to explore multi-gene interactions to understand the complex genetic interplay involved in colorectal cancer (CRC) and Inflammatory Bowel Disease (IBD). Incorporating more IBD datasets will enhance the robustness of the findings and may reveal additional genes associated with the disease. Additionally, expanding the analysis to include more genes, beyond those common across all datasets, could uncover further genetic factors contributing to CRC and IBD. Laboratory experiments are essential to validate the functional roles of these genes and their interactions, which could lead to the development of novel therapeutic targets and strategies for patient care.

Bibliography

[1] M. Araghi, I. Soerjomataram, M. Jenkins, J. Brierley, E. Morris, F. Bray, and M. Arnold. Global trends in colorectal cancer mortality: projections to the year 2035. *Int J Cancer*, 144(12):2992–3000, Jun 2019.

[2] M. Swiderska, B. ska, E. browska, E. Konarzewska-Duchnowska, K. ska, G. Szczurko, P. liwiec, J. Dadan, J. R. Ladny, and K. Zwierz. The diagnostics of colorectal cancer. *Contemp Oncol (Pozn)*, 18(1):1–6, 2014.

[3] Fanny ER Vuik, Stella AV Nieuwenburg, Marc Bardou, Iris Lansdorp-Vogelaar, Mário Dinis-Ribeiro, Maria J Bento, Vesna Zadnik, María Pellisé, Laura Esteban, Michal F Kaminski, et al. Increasing incidence of colorectal cancer in young adults in Europe over the last 25 years. *Gut*, 68(10):1820–1826, 2019.

[4] Rebecca L Siegel, Lindsey A Torre, Isabelle Soerjomataram, Richard B Hayes, Freddie Bray, Thomas K Weber, and Ahmedin Jemal. Global patterns and trends in colorectal cancer incidence in young adults. *Gut*, 68(12):2179–2185, 2019.

[5] Karina W Davidson, Michael J Barry, Carol M Mangione, Michael Cabana, Aaron B Caughey, Esa M Davis, Katrina E Donahue, Chyke A Doubeni, Alex H Krist, Martha

Kubik, et al. Screening for colorectal cancer: Us preventive services task force recommendation statement. *Jama*, 325(19):1965–1977, 2021.

[6] Giulia Martina Cavestro, Alessandro Mannucci, Francesc Balaguer, Heather Hampel, Sonia S Kupfer, Alessandro Repici, Andrea Sartore-Bianchi, Toni T Seppälä, Vincenzo Valentini, Clement Richard Boland, et al. Delphi initiative for early-onset colorectal cancer (direct) international management guidelines. *Clinical Gastroenterology and Hepatology*, 21(3):581–603, 2023.

[7] Lorne J Hofseth, James R Hebert, Anindya Chanda, Hexin Chen, Bryan L Love, Maria M Pena, E Angela Murphy, Mathew Sajish, Amit Sheth, Phillip J Buckhaults, et al. Early-onset colorectal cancer: initial clues and current views. *Nature reviews Gastroenterology & hepatology*, 17(6):352–364, 2020.

[8] M Toma, L Beluşică, M Stavarachi, P Apostol, S Spandole, I Radu, and D Cimponeriu. Rating the environmental and genetic risk factors for colorectal cancer. *J Med Life*, 5(Spec Issue):152–159, October 2012.

[9] Ajay Goel and C Richard Boland. Epigenetics of colorectal cancer. *Gastroenterology*, 143(6):1442–1460, 2012.

[10] James RM Black and Nicholas McGranahan. Genetic and non-genetic clonal diversity in cancer evolution. *Nature Reviews Cancer*, 21(6):379–392, 2021.

[11] Douglas Hanahan. Hallmarks of cancer: new dimensions. *Cancer discovery*, 12(1):31–46, 2022.

[12] Gerhard Jung, Eva Hernández-Illán, Leticia Moreira, Francesc Balaguer, and Ajay Goel. Epigenetics of colorectal cancer: biomarker and therapeutic potential. *Nature reviews Gastroenterology & hepatology*, 17(2):111–130, 2020.

[13] Timon Heide, Jacob Househam, George D Cresswell, Inmaculada Spiteri, Claire Lynn, Maximilian Mossner, Chris Kimberley, Javier Fernandez-Mateos, Bingjie Chen, Luis Zapata, et al. The co-evolution of the genome and epigenome in colorectal cancer. *Nature*, pages 1–11, 2022.

[14] Irfan M Hisamuddin and Vincent W Yang. Genetics of colorectal cancer. *Medscape General Medicine*, 6(3), 2004.

[15] Siân Jones, Wei-dong Chen, Giovanni Parmigiani, Frank Diehl, Niko Beerenwinkel, Tibor Antal, Arne Traulsen, Martin A Nowak, Christopher Siegel, Victor E Velculescu, et al. Comparative lesion sequencing provides insights into tumor evolution. *Proceedings of the National Academy of Sciences*, 105(11):4283–4288, 2008.

[16] Franziska Michor, Yoh Iwasa, Christoph Lengauer, and Martin A Nowak. Dynamics of colorectal cancer. In *Seminars in Cancer Biology*, volume 15, pages 484–493. Elsevier, 2005.

[17] William M Grady and John M Carethers. Genomic and epigenetic instability in colorectal cancer pathogenesis. *Gastroenterology*, 135(4):1079–1099, 2008.

[18] Axel Walther, Elaine Johnstone, Charles Swanton, Rachel Midgley, Ian Tomlinson, and David Kerr. Genetic prognostic and predictive markers in colorectal cancer. *Nature Reviews Cancer*, 9(7):489–499, 2009.

[19] Wouter Laurentius Smit, Claudia Nanette Spaan, Ruben Johannes de Boer, Prashanthi Ramesh, Tânia Martins Garcia, Bartolomeus Joannes Meijer, Jacqueline Ludovicus Maria Vermeulen, Marco Lezzerini, Alyson Winfried MacInnes, Jan Koster, et al. Driver mutations of the adenoma-carcinoma sequence govern the intestinal epithelial global translational capacity. *Proceedings of the National Academy of Sciences*, 117(41):25560–25570, 2020.

[20] Rodrigo Dienstmann, Louis Vermeulen, Justin Guinney, Scott Kopetz, Sabine Tejpar, and Josep Tabernero. Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nature Reviews Cancer*, 17(2):79–92, 2017.

[21] Mirjam M. de Jong, Ilja M. Nolte, Gerard J. te Meerman, Winette T. A. van der Graaf, Elisabeth G. E. de Vries, Rolf H. Sijmons, Robert M. W. Hofstra, and Jan H. Kleibeuker. Low-penetrance Genes and Their Involvement in Colorectal Cancer Susceptibility1. *Cancer Epidemiology, Biomarkers & Prevention*, 11(11):1332–1352, 11 2002.

[22] Justin Guinney, Rodrigo Dienstmann, Xin Wang, Aurélien De Reynies, Andreas Schlicker, Charlotte Soneson, Laetitia Marisa, Paul Roepman, Gift Nyamundanda, Paolo Angelino, et al. The consensus molecular subtypes of colorectal cancer. *Nature Medicine*, 21(11):1350–1356, 2015.

[23] Eva Budinska, Vlad Popovici, Sabine Tejpar, Giovanni D'Ario, Nicolas Lapique, Katarzyna Otylia Sikora, Antonio Fabio Di Narzo, Pu Yan, John Graeme Hodgson, Scott Weinrich, et al. Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *The Journal of Pathology*, 231(1):63–76, 2013.

[24] Laetitia Marisa, Aurélien de Reyniès, Alex Duval, Janick Selves, Marie Pierre Gaub, Laure Vescovo, Marie-Christine Etienne-Grimaldi, Renaud Schiappa, Dominique Guenot, Mira Ayadi, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Medicine*, 10(5):e1001453, 2013.

[25] Chunming Xu and Scott A. Jackson. Machine learning and complex biological data. *Genome Biology*, 20(1):76, Apr 2019.

[26] Zaoqu Liu, Long Liu, Siyuan Weng, Chunguang Guo, Qin Dang, Hui Xu, Libo Wang, Taoyuan Lu, Yuyuan Zhang, Zhenqiang Sun, et al. Machine learning-based integration develops an immune-derived lncRNA signature for improving outcomes in colorectal cancer. *Nature Communications*, 13(1):816, 2022.

[27] Qi Su, Qin Liu, Raphaela Iris Lau, Jingwan Zhang, Zhilu Xu, Yun Kit Yeoh, Thomas W. H. Leung, Whitney Tang, Lin Zhang, Jessie Q. Y. Liang, Yuk Kam Yau, Jiaying Zheng, Chengyu Liu, Mengjing Zhang, Chun Pan Cheung, Jessica Y. L. Ching, Hein M. Tun, Jun Yu, Francis K. L. Chan, and Siew C. Ng. Faecal microbiome-based machine learning for multi-class disease diagnosis. *Nature Communications*, 13(1):6818, Nov 2022.

[28] Emily K. Makowski, Patrick C. Kinnunen, Jie Huang, Lina Wu, Matthew D. Smith, Tiexin Wang, Alec A. Desai, Craig N. Streu, Yulei Zhang, Jennifer M. Zupancic, John S. Schardt, Jennifer J. Linderman, and Peter M. Tessier. Co-optimization of therapeutic antibody affinity and specificity using machine learning models that generalize to novel mutational space. *Nature Communications*, 13(1):3788, Jul 2022.

[29] Hannah L Nicholls, Christopher R John, David S Watson, Patricia B Munroe, Michael R Barnes, and Claudia P Cabrera. Reaching the end-game for GWAS: machine learning approaches for the prioritization of complex disease loci. *Frontiers in Genetics*, 11:350, 2020.

[30] Kyle Swanson, Eric Wu, Angela Zhang, Ash A Alizadeh, and James Zou. From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. *Cell*, 2023.

[31] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, Peer Bork, Lars J Jensen, and Christian von Mering. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res*, 51(D1):D638–D646, January 2023.

[32] Zhilin Zou, Tao Tao, Hongmei Li, and Xiao Zhu. mTOR signaling pathway and mTOR inhibitors in cancer: progress and challenges. *Cell & Bioscience*, 10(1):31, March 2020.

[33] Aria Sardari and Hamid Usefi. Machine learning-based meta-analysis of colorectal cancer and inflammatory bowel disease. *PLOS ONE*, 18(12):1–16, 12 2023.

[34] F. Dorani, T. Hu, M. O. Woods, and G. Zhai. Ensemble learning for detecting gene-gene interactions in colorectal cancer. *PeerJ*, 6:e5854, 2018.

[35] D. Ding, S. Han, H. Zhang, Y. He, and Y. Li. Predictive biomarkers of colorectal cancer. *Comput Biol Chem*, 83:107106, Dec 2019.

[36] D. Zhao, H. Liu, Y. Zheng, Y. He, D. Lu, and C. Lyu. Whale optimized mixed kernel function of support vector machine for colorectal cancer diagnosis. *J Biomed Inform*, 92:103124, Apr 2019.

[37] Ying Su, Xuecong Tian, Rui Gao, Wenjia Guo, Cheng Chen, Chen Chen, Dongfang Jia, Hongtao Li, and Xiaoyi Lv. Colon cancer diagnosis and staging classification based on machine learning and bioinformatics analysis. *Computers in Biology and Medicine*, 145:105409, 2022.

[38] Neha Shree Maurya, Sandeep Kushwaha, Aakash Chawade, and Ashutosh Mani. Transcriptome profiling by combined machine learning and statistical R analysis identifies tmem236 as a potential novel diagnostic biomarker for colorectal cancer. *Scientific Reports*, 11(1):14304, Jul 2021.

[39] Manolis Kellis et al. *Computational Biology - Genomes, Networks, and Evolution*. LibreTexts, 2023. Accessed: 2023.

[40] S. Tsukamoto, T. Ishikawa, S. Iida, M. Ishiguro, K. Mogushi, H. Mizushima, H. Uetake, H. Tanaka, and K. Sugihara. Clinical significance of osteoprotegerin expression in human colorectal cancer. *Clin Cancer Res*, 17(8):2444–2450, Apr 2011.

[41] X. Solé, M. Crous-Bou, D. Cordero, D. Olivares, E. ó, R. Sanz-Pamplona, F. Rodriguez-Moranta, X. Sanjuan, J. de Oca, R. Salazar, and V. Moreno. Dis-

covery and validation of new potential biomarkers for early detection of colon cancer. *PLoS One*, 9(9):e106748, 2014.

[42] B. M. Ryan, K. A. Zanetti, A. I. Robles, A. J. Schetter, J. Goodman, R. B. Hayes, W. Y. Huang, M. J. Gunter, M. Yeager, L. Burdette, S. I. Berndt, and C. C. Harris. Germline variation in NCF4, an innate immunity gene, is associated with an increased risk of colorectal cancer. *Int J Cancer*, 134(6):1399–1407, Mar 2014.

[43] G. Getz, H. Gal, I. Kela, D. A. Notterman, and E. Domany. Coupled two-way clustering analysis of breast cancer and colon cancer gene expression data. *Bioinformatics*, 19(9):1079–1089, Jun 2003.

[44] L. Zuurbier, A. Rahman, M. Cordes, J. Scheick, T. J. Wong, F. Rustenburg, J. C. Joseph, P. Dynoodt, R. Casey, P. Drillenburg, M. Gerhards, A. Barat, R. Klinger, B. Fender, D. P. O'Connor, J. Betge, M. P. Ebert, T. Gaiser, J. H. M. Prehn, A. W. Griffioen, N. C. T. van Grieken, B. Ylstra, A. T. Byrne, L. G. van der Flier, W. M. Gallagher, and R. Postel. Apelin: A putative novel predictive biomarker for bevacizumab response in colorectal cancer. *Oncotarget*, 8(26):42949–42961, Jun 2017.

[45] J. Brouwer-Visser, W. Y. Cheng, A. Bauer-Mehren, D. Maisel, K. Lechner, E. Andersson, J. T. Dudley, and F. Milletti. Regulatory T-cell Genes Drive Altered Immune Microenvironment in Adult Solid Cancers and Allow for Immune Contextual Patient Subtyping. *Cancer Epidemiol Biomarkers Prev*, 27(1):103–112, Jan 2018.

[46] T. Hinoue, D. J. Weisenberger, C. P. Lange, H. Shen, H. M. Byun, D. Van Den Berg, S. Malik, F. Pan, H. Noushmehr, C. M. van Dijk, R. A. Tollenaar, and P. W. Laird.

Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Res*, 22(2):271–282, Feb 2012.

[47] A. Wenke, H. Armbruster, K. Balschun, J. Starmann, H. Sültmann, and C. Röcken. Kras-genotype dependent gene expression pattern in colorectal cancer. `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE38026`, 2012. Accessed: May 18, 2012.

[48] P. Alhopuro, H. Sammalkorpi, I. ki, M. m, A. Raitila, J. Saharinen, K. Nousiainen, H. J. Lehtonen, E. vaara, J. Puhakka, S. Tuupanen, S. Sousa, R. Seruca, A. M. Ferreira, R. M. Hofstra, J. P. Mecklin, H. rvinen, A. ki, T. F. Orntoft, S. Hautaniemi, D. Arango, A. Karhu, and L. A. Aaltonen. Candidate driver genes in microsatellite-unstable colorectal cancer. *Int J Cancer*, 130(7):1558–1566, Apr 2012.

[49] L. Marisa, A. s, A. Duval, J. Selves, M. P. Gaub, L. Vescovo, M. C. Etienne-Grimaldi, R. Schiappa, D. Guenot, M. Ayadi, S. Kirzin, M. Chazal, J. F. jou, D. Benchimol, A. Berger, A. Lagarde, E. Pencreach, F. Piard, D. Elias, Y. Parc, S. Olschwang, G. Milano, P. Laurent-Puig, and V. Boige. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Med*, 10(5):e1001453, 2013.

[50] A. Shen, L. Liu, Y. Huang, Z. Shen, M. Wu, X. Chen, X. Wu, X. Lin, Y. Chen, L. Li, Y. Cheng, J. Chu, T. J. Sferra, L. Wei, Q. Zhuang, and J. Peng. Down-Regulating HAUS6 Suppresses Cell Proliferation by Activating the p53/p21 Pathway in Colorectal Cancer. *Front Cell Dev Biol*, 9:772077, 2021.

[51] X. Shi, Y. Zhang, B. Cao, N. Lu, L. Feng, X. Di, N. Han, C. Luo, G. Wang, S. Cheng, and K. Zhang. Gene expression profiling of colorectal normal mucosa, adenoma and adenocarcinoma tissues. `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE41657`, 2015. Accessed: Oct 17, 2012.

[52] R. N. Jorissen, L. Lipton, P. Gibbs, M. Chapman, J. Desai, I. T. Jones, T. J. Yeatman, P. East, I. P. Tomlinson, H. W. Verspaget, L. A. Aaltonen, M. ffer, T. F. Orntoft, C. L. Andersen, and O. M. Sieber. DNA copy-number alterations underlie gene expression differences between microsatellite stable and unstable colorectal cancers. *Clin Cancer Res*, 14(24):8061–8069, Dec 2008.

[53] R. N. Jorissen, P. Gibbs, M. Christie, S. Prakash, L. Lipton, J. Desai, D. Kerr, L. A. Aaltonen, D. Arango, M. ffer, T. F. Orntoft, C. L. Andersen, M. Gruidl, V. P. Kamath, S. Eschrich, T. J. Yeatman, and O. M. Sieber. Metastasis-Associated Gene Expression Changes Predict Poor Outcomes in Patients with Dukes Stage B and C Colorectal Cancer. *Clin Cancer Res*, 15(24):7642–7651, Dec 2009.

[54] J. J. Smith, N. G. Deane, F. Wu, N. B. Merchant, B. Zhang, A. Jiang, P. Lu, J. C. Johnson, C. Schmidt, C. E. Bailey, S. Eschrich, C. Kis, S. Levy, M. K. Washington, M. J. Heslin, R. J. Coffey, T. J. Yeatman, Y. Shyr, and R. D. Beauchamp. Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. *Gastroenterology*, 138(3):958–968, Mar 2010.

[55] Expression project for oncology (expo). `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi`, 2012. Accessed: Oct 68, 2015.

[56] A. Schlicker, G. Beran, C. M. Chresta, G. McWalter, A. Pritchard, S. Weston, S. Runswick, S. Davenport, K. Heathcote, D. A. Castro, G. Orphanides, T. French, and L. F. Wessels. Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. *BMC Med Genomics*, 5:66, Dec 2012.

[57] Van Steen K De Hertogh G Geboes K Schuit F Rutgeerts P Arijs I, Van Lommel L. Mucosal expression profiling in patients with inflammatory bowel disease before and after first infliximab treatment. `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16879`, 2009. Accessed: Jun 29, 2009.

[58] Patricia Lepage, Robert Häsler, Martina E Spehlmann, Ateequr Rehman, Aida Zvirbliene, Alexander Begun, Stephan Ott, Limas Kupcinskas, Joël Doré, Andreas Raedler, and Stefan Schreiber. Twin study indicates loss of interaction between microbiota and mucosa of patients with ulcerative colitis. *Gastroenterology*, 141(1):227–236, April 2011.

[59] W. Vanhove, P. M. Peeters, D. Staelens, A. Schraenen, J. Van der Goten, I. Cleynen, S. De Schepper, L. Van Lommel, N. L. Reynaert, F. Schuit, G. Van Assche, M. Ferrante, G. De Hertogh, E. F. Wouters, P. Rutgeerts, S. Vermeire, K. Nys, and I. Arijs. Strong Upregulation of AIM2 and IFI16 Inflammasomes in the Mucosa of Patients with Active Inflammatory Bowel Disease. *Inflamm Bowel Dis*, 21(11):2673–2682, Nov 2015.

[60] Sare Verstockt, Gert De Hertogh, Jan Van der Goten, Bram Verstockt, Maaike Vancamelbeke, Kathleen Machiels, Leentje Van Lommel, Frans Schuit, Gert Van Assche, Paul Rutgeerts, Marc Ferrante, Séverine Vermeire, Ingrid Arijs, and Isabelle Cley-

nen. Gene and mirna regulatory networks during different stages of Crohn's disease. *J Crohns Colitis*, 13(7):916–930, July 2019.

[61] M. E. Keir, F. Fuh, R. Ichikawa, M. Acres, J. A. Hackney, G. Hulme, C. D. Carey, J. Palmer, C. J. Jones, A. K. Long, J. Jiang, S. Klabunde, J. C. Mansfield, C. M. Looney, W. A. Faubion, A. Filby, J. A. Kirby, J. McBride, and C. A. Lamb. E Integrin and Gut Homing Integrins in Migration and Retention of Intestinal Lymphocytes during Inflammatory Bowel Disease. *J Immunol*, 207(9):2245–2254, Nov 2021.

[62] Jørgen Olsen, Thomas A Gerds, Jakob B Seidelin, Claudio Csillag, Jacob T Bjerrum, Jesper T Troelsen, and Ole Haagen Nielsen. Diagnosis of ulcerative colitis before onset of inflammation by multivariate modeling of genome-wide gene expression data. *Inflamm Bowel Dis*, 15(7):1032–1038, July 2009.

[63] Trinidad Montero-Meléndez, Xavier Llor, Esther García-Planella, Mauro Perretti, and Antonio Suárez. Identification of novel predictor classifiers for inflammatory bowel disease by gene expression profiling. *PLoS One*, 8(10):e76235, October 2013.

[64] J. Pekow, U. Dougherty, Y. Huang, E. Gometz, J. Nathanson, G. Cohen, S. Levy, M. Kocherginsky, N. Venu, M. Westerhoff, J. Hart, A. E. Noffsinger, S. B. Hanauer, R. D. Hurst, A. Fichera, L. J. Joseph, Q. Liu, and M. Bissonnette. Gene signature distinguishes patients with chronic ulcerative colitis harboring remote neoplastic lesions. *Inflamm Bowel Dis*, 19(3):461–470, Mar 2013.

[65] O. Galamb, B. rffy, F. Sipos, S. k, A. M. meth, P. Miheller, Z. Tulassay, E. Dinya, and B. r. Inflammation, adenoma and cancer: objective classification of colon biopsy

specimens with gene expression signature. *Dis Markers*, 25(1):1–16, 2008.

[66] J. Van der Goten, W. Vanhove, K. Lemaire, L. Van Lommel, K. Machiels, W. J. Wollants, V. De Preter, G. De Hertogh, M. Ferrante, G. Van Assche, P. Rutgeerts, F. Schuit, S. Vermeire, and I. Arijs. Integrated miRNA and mRNA expression profiling in inflamed colon of patients with ulcerative colitis. *PLoS One*, 9(12):e116117, 2014.

[67] W. J. Sandborn, B. G. Feagan, C. Marano, H. Zhang, R. Strauss, J. Johanns, O. J. Adedokun, C. Guzzo, J. F. Colombel, W. Reinisch, P. R. Gibson, J. Collins, G. rnerot, T. Hibi, and P. Rutgeerts. Subcutaneous golimumab induces clinical response and remission in patients with moderate-to-severe ulcerative colitis. *Gastroenterology*, 146(1):85–95, Jan 2014.

[68] Evelyn Fix and J. L. Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review / Revue Internationale de Statistique*, 57(3):238–247, 2024/04/04/ 1989. Full publication date: Dec., 1989.

[69] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, jun 2002.

[70] M. Afshar and H. Usefi. Dimensionality reduction using singular vectors. *Sci Rep*, 11(1):3832, Feb 2021.

[71] Hamid Usefi. Clustering, multicollinearity, and singular vectors. *Computational Statistics & Data Analysis*, 173:107523, 2022.

[72] Hidayati Husainy Hasbullah and Marahaini Musa. Gene therapy targeting p53 and KRAS for colorectal cancer treatment: A myth or the way forward? *International Journal of Molecular Sciences*, 22(21), 2021.

[73] L. Shen, Y. Kondo, G. L. Rosner, L. Xiao, N. S. Hernandez, J. Vilaythong, P. S. Houlihan, R. S. Krouse, A. R. Prasad, J. G. Einspahr, J. Buckmeier, D. S. Alberts, S. R. Hamilton, and J. P. Issa. MGMT promoter methylation and field defect in sporadic colorectal cancer. *J Natl Cancer Inst*, 97(18):1330–1338, Sep 2005.

[74] F. Arvelo, F. Sojo, and C. Cotte. Biology of colorectal cancer. *Ecancermedicalscience*, 9:520, 2015.

[75] Andrew Chatr-aryamontri, Arnaud Ceol, Luisa Montecchi Palazzi, Giuliano Nardelli, Maria Victoria Schneider, Luisa Castagnoli, and Gianni Cesareni. MINT: the molecular INTeraction database. *Nucleic Acids Res*, 35(Database issue):D572–4, November 2006.

[76] Suraj Peri, J Daniel Navarro, Troels Z Kristiansen, Ramars Amanchy, Vineeth Surendranath, Babylakshmi Muthusamy, T K B Gandhi, K N Chandrika, Nandan Deshpande, Shubha Suresh, B P Rashmi, K Shanker, N Padma, Vidya Niranjan, H C Harsha, Naveen Talreja, B M Vrushabendra, M A Ramya, A J Yatish, Mary Joy, H N Shivashankar, M P Kavitha, Minal Menezes, Dipanwita Roy Choudhury, Neelanjana Ghosh, R Saravana, Sreenath Chandran, Sujatha Mohan, Chandra Kiran Jonnalagadda, C K Prasad, Chandan Kumar-Sinha, Krishna S Deshpande, and Akhilesh Pandey. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res*, 32(Database issue):D497–501, January 2004.

[77] Gary D Bader, Doron Betel, and Christopher W V Hogue. BIND: the biomolecular interaction network database. *Nucleic Acids Res*, 31(1):248–250, January 2003.

[78] I Xenarios, D W Rice, L Salwinski, M K Baron, E M Marcotte, and D Eisenberg. DIP: the database of interacting proteins. *Nucleic Acids Res*, 28(1):289–291, January 2000.

[79] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*, 34(Database issue):D535–9, January 2006.

[80] M Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, January 2000.

[81] David Croft, Gavin O'Kelly, Guanming Wu, Robin Haw, Marc Gillespie, Lisa Matthews, Michael Caudy, Phani Garapati, Gopal Gopinath, Bijay Jassal, Steven Jupe, Irina Kalatskaya, Shahana Mahajan, Bruce May, Nelson Ndegwa, Esther Schmidt, Veronica Shamovsky, Christina Yung, Ewan Birney, Henning Hermjakob, Peter D'Eustachio, and Lincoln Stein. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res*, 39(Database issue):D691–7, November 2010.

[82] Henning Hermjakob, Luisa Montecchi-Palazzi, Chris Lewington, Sugath Mudali, Samuel Kerrien, Sandra Orchard, Martin Vingron, Bernd Roechert, Peter Roepstorff, Alfonso Valencia, Hanah Margalit, John Armstrong, Amos Bairoch, Gianni Cesareni,

David Sherman, and Rolf Apweiler. IntAct: an open source molecular interaction database. *Nucleic Acids Res*, 32(Database issue):D452–5, January 2004.

[83] Peter D Karp, Daniel Weaver, Suzanne Paley, Carol Fulcher, Aya Kubo, Anamika Kothari, Markus Krummenacker, Pallavi Subhraveti, Deepika Weerasinghe, Socorro Gama-Castro, Araceli M Huerta, Luis Muñiz-Rascado, César Bonavides-Martinez, Verena Weiss, Martin Peralta-Gil, Alberto Santos-Zavaleta, Imke Schröder, Amanda Mackie, Robert Gunsalus, Julio Collado-Vides, Ingrid M Keseler, and Ian Paulsen. The EcoCyc database. *EcoSal Plus*, 6(1), May 2014.

[84] Carl F Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H Buetow. PID: the pathway interaction database. *Nucleic Acids Res*, 37(Database issue):D674–9, October 2008.

[85] M A Harris, J Clark, A Ireland, J Lomax, M Ashburner, R Foulger, K Eilbeck, S Lewis, B Marshall, C Mungall, J Richter, G M Rubin, J A Blake, C Bult, M Dolan, H Drabkin, J T Eppig, D P Hill, L Ni, M Ringwald, R Balakrishnan, J M Cherry, K R Christie, M C Costanzo, S S Dwight, S Engel, D G Fisk, J E Hirschman, E L Hong, R S Nash, A Sethuraman, C L Theesfeld, D Botstein, K Dolinski, B Feierbach, T Berardini, S Mundodi, S Y Rhee, R Apweiler, D Barrell, E Camon, E Dimmer, V Lee, R Chisholm, P Gaudet, W Kibbe, R Kishore, E M Schwarz, P Sternberg, M Gwinn, L Hannick, J Wortman, M Berriman, V Wood, N de la Cruz, P Tonellato, P Jaiswal, T Seigfried, R White, and Gene Ontology Consortium. The gene ontology (GO) database and informatics resource. *Nucleic Acids Res*, 32(Database issue):D258–61, January 2004.

[86] S. Kim, N. Kim, K. Kang, W. Kim, J. Won, and J. Cho. Whole Transcriptome Analysis Identifies TNS4 as a Key Effector of Cetuximab and a Regulator of the Oncogenic Activity of KRAS Mutant Colorectal Cancer Cell Lines. *Cells*, 8(8), Aug 2019.

[87] T. P. Raposo, S. Susanti, and M. Ilyas. Investigating TNS4 in the Colorectal Tumor Microenvironment Using 3D Spheroid Models of Invasion. *Adv Biosyst*, 4(6):e2000031, Jun 2020.

[88] A. K. Najumudeen, F. Ceteci, S. K. Fey, G. Hamm, R. T. Steven, H. Hall, C. J. Nikula, A. Dexter, T. Murta, A. M. Race, D. Sumpton, N. Vlahov, D. M. Gay, J. R. P. Knight, R. Jackstadt, J. D. G. Leach, R. A. Ridgway, E. R. Johnson, C. Nixon, A. Hedley, K. Gilroy, W. Clark, S. B. Malla, P. D. Dunne, G. Rodriguez-Blanco, S. E. Critchlow, A. Mrowinska, G. Malviya, D. Solovyev, G. Brown, D. Y. Lewis, G. M. Mackay, D. Strathdee, S. Tardito, E. Gottlieb, Z. Takats, S. T. Barry, R. J. A. Goodwin, J. Bunch, M. Bushell, A. D. Campbell, O. J. Sansom, A. Campbell, A. Najumudeen, A. M. Race, I. Gilmore, G. McMahon, P. Grant, B. Yan, A. J. Taylor, E. Elia, S. Thomas, C. Munteanu, A. Al-Afeef, A. Burton, J. L. Vorng, X. Loizeau, W. Zhou, A. Nasif, A. Gonzalez, H. Koquna, M. Metodiev, M. Kyriazi, J. Zhang, L. Zeiger, J. Vande-Voorde, J. Morton, D. Soloviev, V. Wu, Y. Xiang, D. McGill, S. Maneta-Stravrakaki, J. Mistry, E. Kazanc, M. Yuneva, Y. Panina, C. S. Nanda, P. Kreuzaler, A. Ghanate, S. Ling, J. Richings, K. Brindle, A. Tsyben, G. Poulogiannis, A. Gupta, A. Tripp, E. Karali, N. Koundouros, T. Tsalikis, J. Marshall, M. Garrett, and H. Hall. The amino acid transporter SLC7A5 is required for efficient growth of KRAS-mutant

colorectal cancer. *Nat Genet*, 53(1):16–26, Jan 2021.

[89] H. Huang, Y. Dai, Y. Duan, Z. Yuan, Y. Li, M. Zhang, W. Zhu, H. Yu, W. Zhong, and S. Feng. Effective prediction of potential ferroptosis critical genes in clinical colorectal cancer. *Front Oncol*, 12:1033044, 2022.

[90] L. Zhang, Z. Zhang, L. Qin, X. Shi, Q. Su, and W. Mo. SDF2L1 Inhibits Cell Proliferation, Migration, and Invasion in Nasopharyngeal Carcinoma. *Biomed Res Int*, 2020:1970936, 2020.

[91] M. Zhao and R. Dai. HIST3H2A is a potential biomarker for pancreatic cancer: A study based on TCGA data. *Medicine (Baltimore)*, 100(46):e27598, Nov 2021.

[92] L. Yi, J. Qiang, P. Yichen, Y. Chunna, Z. Yi, K. Xun, Z. Jianwei, B. Rixing, Y. Wenmao, W. Xiaomin, L. Parker, and L. Wenbin. Identification of a 5-gene-based signature to predict prognosis and correlate immunomodulators for rectal cancer. *Transl Oncol*, 26:101529, Sep 2022.

[93] S. Cruz-Gil, R. Sanchez-Martinez, M. Gomez de Cedron, R. Martin-Hernandez, T. Vargas, S. Molina, J. Herranz, A. Davalos, G. Reglero, and A. Ramirez de Molina. in colorectal cancer progression by therapeutic miRNAs: miR-19b-1 role. *J Lipid Res*, 59(1):14–24, Jan 2018.

[94] C. Liao, M. Li, X. Li, N. Li, X. Zhao, X. Wang, Y. Song, J. Quan, C. Cheng, J. Liu, A. M. Bode, Y. Cao, and X. Luo. Trichothecin inhibits invasion and metastasis of colon carcinoma associating with SCD-1-mediated metabolite alteration. *Biochim Biophys Acta Mol Cell Biol Lipids*, 1865(2):158540, Feb 2020.

[95] Z. Tang, S. Yuan, Y. Hu, H. Zhang, W. Wu, Z. Zeng, J. Yang, J. Yun, R. Xu, and P. Huang. Over-expression of GAPDH in human colorectal carcinoma as a preferred target of 3-bromopyruvate propyl ester. *J Bioenerg Biomembr*, 44(1):117–125, Feb 2012.

[96] Míriam Tarrado-Castellarnau, Santiago Diaz-Moralli, Ibrahim H. Polat, Rebeca Sanz-Pamplona, Cristina Alenda, Víctor Moreno, Antoni Castells, and Marta Cascante. Glyceraldehyde-3-phosphate dehydrogenase is overexpressed in colorectal cancer onset. *Translational Medicine Communications*, 2(1):6, Jun 2017.

[97] M. Li, X. Sun, H. Yao, W. Chen, F. Zhang, S. Gao, X. Zou, J. Chen, S. Qiu, H. Wei, Z. Hu, and W. Chen. Genomic methylation variations predict the susceptibility of six chemotherapy related adverse effects and cancer development for Chinese colorectal cancer patients. *Toxicol Appl Pharmacol*, 427:115657, Sep 2021.

[98] L. Zhang, Z. Zhang, L. Qin, X. Shi, Q. Su, and W. Mo. SDF2L1 Inhibits Cell Proliferation, Migration, and Invasion in Nasopharyngeal Carcinoma. *Biomed Res Int*, 2020:1970936, 2020.

[99] C. Shi, Z. He, N. Hou, Y. Ni, L. Xiong, and P. Chen. Alpha B-crystallin correlates with poor survival in colorectal cancer. *Int J Clin Exp Pathol*, 7(9):6056–6063, 2014.

[100] J. Deng, X. Chen, T. Zhan, M. Chen, X. Yan, and X. Huang. CRYAB predicts clinical prognosis and is associated with immunocyte infiltration in colorectal cancer. *PeerJ*, 9:e12578, 2021.

[101] A. Dai, X. Guo, X. Yang, M. Li, Y. Fu, and Q. Sun. Effects of the CRYAB gene on stem cell-like properties of colorectal cancer and its mechanism. *J Cancer Res Ther*, 18(5):1328–1337, Sep 2022.