TECHNISCHE
UNIVERSITÄT
WIEN

DIPLOMA THESIS

Analysis of colorectal cancer and adenoma microbiome signatures and the application of machine learning classification as a potential screening tool

Executed at

Biome Diagnostics GmbH

and at

Institute of Chemical, Environmental and Bioscience Engineering
Faculty of Technical Chemistry, TU Wien

by

BSc Katarina Priselac
Registration Number 11724446

under the supervision of

Main advisor: Univ.Prof. Mag. Dr.rer.nat. MSc.Tox. Andreas Farnleitner
Co-advisor: Senior Scientist Dipl.-Ing. Dr.techn. Georg Reischer

Vienna, 10th July, 2023

Katarina Priselac

Abstract

Recent studies have shown an association between the development of colorectal cancer (CRC) and the composition of the patients' gut microbiome. The aims of this thesis were to identify microbial signatures in the gut microbiome associated with CRC and cancer precursor (adenoma), and to develop machine learning models for the screening of these diseases based on the composition of the stool microbiome.

Meta-analysis dataset containing 1786 samples from healthy individuals and adenoma and CRC patients was obtained from publicly available repositories. Differential abundance analysis (DAA) was performed to detect biomarkers using three methods: ALDEx2, ANCOM-BC, and MaAsLin2. Machine learning models for distinguishing healthy individuals and CRC or adenoma patients were trained on 80% of the dataset and tested on the remaining 20%, with several parameter options to optimise the performance.

DAA of CRC compared with healthy samples revealed a total of 39 differentially abundant taxa identified by all three methods. Comparison of adenoma and healthy samples resulted in 111 detected DA taxa by ALDEx2 and ANCOM-BC. The best machine learning performance for CRC-healthy classification was obtained using a support vector machine model with a radial kernel on a genus level with MaAsLin2 feature selection. This model yielded an area under the curve (AUC) of 0.84 for cross-validation and 0.80 for the test dataset. For the distinction between adenoma and healthy samples, the light gradient-boosting machine model using the 50 highest scoring species achieved an AUC of 0.85 for cross-validation and 0.72 for the test dataset.

Machine learning models performed comparably well in detecting CRC and better in detecting adenomas than the currently used fecal tests. For the first time, a large meta-analysis dataset was successfully used to demonstrate the suitability of machine learning algorithms for the identification of bacterial biomarkers and the development of microbiome-based diagnostic solutions for CRC and adenoma. The developed models were able to screen for these diseases non-invasively - based on stool samples, and with fairly high accuracy already. With further optimization, such tools could be used in the future to accompany colonoscopy in regular screening programs for colorectal cancer.

Kurzfassung

Jüngste Studien haben einen Zusammenhang zwischen der Entwicklung von Darmkrebs (CRC) und der Zusammensetzung des Darmmikrobioms der Patienten gezeigt. Ziel dieser Arbeit war es, mikrobielle Signaturen im Darmmikrobiom zu identifizieren, die mit Darmkrebs und Krebsvorstufen (Adenomen) assoziiert sind, und maschinelle Lernmodelle für das Screening dieser Krankheiten auf der Grundlage der Zusammensetzung des Stuhlmikrobioms zu entwickeln.

Ein Meta-Analyse-Datensatz mit 1786 Proben von gesunden Personen, Adenom- und Darmkrebs-Patienten wurde aus öffentlich zugänglichen Quellen beschafft. Die differenzielle Abundanzanalyse (DAA) wurde durchgeführt, um die Biomarker mit drei Methoden zu erkennen: ALDEx2, ANCOM-BC und MaAsLin2. Modelle für maschinelles Lernen zur Unterscheidung zwischen gesunden Personen und CRC- oder Adenom-Patienten wurden auf 80% des Datensatzes trainiert und auf den verbleibenden 20% getestet, wobei mehrere Parameteroptionen zur Optimierung der Leistung verwendet wurden.

Die DAA von Darmkrebs im Vergleich zu gesunden Proben ergab insgesamt 39 differenziell abundante Taxa, die mit allen drei Methoden identifiziert wurden. Beim Vergleich von Adenomen und gesunden Proben wurden 111 DA-Taxa von ALDEx2 und ANCOM-BC erkannt. Die beste maschinelle Lernleistung für die Klassifizierung von CRC-gesund wurde mit einem Support Vector Machine Modell mit einem radialen Kernel auf Gattungsebene mit MaAsLin2-Variablenauswahl erzielt. Dieses Modell ergab eine Fläche unter der Kurve (AUC) von 0.84 für die Kreuzvalidierung und 0.80 für den Testdatensatz. Für die Unterscheidung zwischen Adenomen und gesunden Proben erreichte das Light Gradient-Boosting Maschine Modell unter Verwendung der 50 ausgewählten Arten eine AUC von 0.85 bei der Kreuzvalidierung und 0.72 im Testdatensatz.

Die maschinellen Lernmodelle schnitten bei der Erkennung von kolorektalen Karzinomen vergleichbar gut und bei der Erkennung von Adenomen besser ab als die derzeit verwendeten Fäkaltests. Zum ersten Mal wurde ein großer Meta-Analyse-Datensatz erfolgreich genutzt, um die Eignung von Algorithmen des maschinellen Lernens für die Identifizierung bakterieller Biomarker und die Entwicklung mikrobiombasierter Diagnoselösungen für Darmkrebs und Adenome zu demonstrieren. Die entwickelten Modelle waren in der Lage, diese Krankheiten nicht-invasiv (auf der Grundlage von Stuhlproben) und bereits mit recht hoher Genauigkeit zu erkennen. Bei weiterer Optimierung könnten solche Tests in

Zukunft als Ergänzung zur Darmspiegelung in regelmäßigen Vorsorgeprogrammen für Darmkrebs eingesetzt werden.

Contents

CHAPTER 1

Introduction

1.1 Gut microbiome

The human gut microbiome comprises a collection of microorganisms that reside in the human gastrointestinal tract, including their genomes, genes, and gene products [1]. Organisms in the microbiome perform essential processes for host physiology and survival. Some functions of the gut microbiome include fermentation of indigestible food components into absorbable metabolites, synthesis of essential vitamins, elimination of toxic compounds, outcompetition of pathogens, strengthening of the intestinal barrier, and stimulation and regulation of the immune system [2].

Given the diverse range of functions of the gut microbiota and the fact that microbial genes are more abundant than genes in the human genome, it is not surprising that the intestinal microbiota plays a critical role in the human body [3]. Although they are involved in numerous vital and beneficial activities, some gut microbiome organisms are also associated with a number of diseases, including diseases both inside and outside the gut, such as rheumatoid arthritis, colorectal cancer, inflammatory bowel disease, obesity, diabetes, and cardiovascular diseases [4]. The shift in microbiome composition associated with diseased states is often referred to as dysbiosis [4].

Gut microbiome composition and function can be assessed from fecal samples (non-invasive) and tissue/biopsy samples taken during endoscopy. In both cases, samples are analyzed by respective methods from the fields of metagenomics, metatranscriptomics, metaproteomics, or metabolomics, depending on the analysis goal. Metagenomics enables the identification and quantification of organisms in a sample, whereas other omics deliver more insight into dynamics and functional processes [2].

1.2 Metagenomics and amplicon sequencing

The first microbial studies utilized the direct cultivation and isolation of microbes to identify and quantify them. This approach is limited because approximately 99% of microbes are currently uncultivable, and the growth conditions used may favor the selection of some species over others [5]. Metagenomics is a modern approach that is defined as the study of a collection of genetic material (genomes) present in a sample with a mixed community of organisms [6]. Metagenomic tools enable microbiome exploration without selection bias or constraints associated with cultivation methods [5]. Basic steps in metagenomic analysis include sample collection, DNA/RNA extraction, library preparation, genetic sequencing, and bioinformatic data analysis [5]. Because this thesis employed an amplicon sequencing dataset, the methodology for this approach is presented in more detail here.

The 16S rRNA gene is present in all bacteria, and consists of hypervariable regions spaced by ultra-conserved regions. The most important step in library preparation for 16S rRNA sequencing (also referred to as amplicon sequencing) is polymerase chain reaction (PCR) used to amplify the genetic material in a sample. Universal primers are used, as they can anneal to the conserved regions of bacteria, and therefore, the 16S rRNA hypervariable region can be amplified. Hypervariable regions are characteristic for each bacterium and are thus used to reliably infer taxonomy up to the genus level. [7, 5]

Amplified region of the 16S rRNA gene is sequenced in the next step. There are many sequencing technologies and platforms, however, next-generation sequencing (NGS) using Illumina has become the mainstream method of choice. The platform is based on sequencing by synthesis technology (SBS). During this type of sequencing, DNA is fragmented and adapters are added to both ends. Single-stranded DNA (ssDNA) is added to the flowcell, and fragments are attached to the surface due to complementary oligos on the adapters and the surface. In a so-called bridge-PCR, each fragment is amplified, forming clusters of identical sequences. SBS follows with the help of DNA polymerase and four dNTPs with specific fluorescence. The 3' end of these dNTPs contains an azide group, which blocks the incorporation of the next base. In this way, photos are scanned after each base incorporation, and based on the light signal, the base is identified. After scanning, the azide group is removed, and the process continues until the desired read length is obtained. After binding another end of ssDNA to the flowcell, forming the double stranded bridge and cleaving off of the original forward strand, the remaining reverse strand is sequenced in the same manner. [6, 8]

Bioinformatic analysis of acquired data starts with reads for each sample stored in FASTQ files (text files with nucleotide sequences, including quality scores for each base). If paired-end reads are available, the pairs are merged to obtain full sequences. The result of a bioinformatic pipeline is an amplicon sequence variant (ASV) table – a matrix with one dimension (rows) corresponding to samples and the second dimension (columns) corresponding to sequence variants (i.e., uniquely identified nucleotide sequences). This matrix contains information on how much of each ASV is present in each sample (the so-called ASV counts). Because ASVs correspond to DNA regions that are specific to each bacterium, taxonomy can be assigned to them. This is done by classifying the obtained ASVs with the help of reference sequences with known taxonomy. [9]

1.3 Machine learning

Machine learning is a part of artificial intelligence and is being used in numerous fields because of its ability to learn from the presented dataset (called training data) and make predictions about unknown data. There are three general types of algorithms: supervised learning, unsupervised learning and reinforcement learning [10]. This thesis employed different learners from the supervised learning category; therefore, a short explanation of the general concept and the applied algorithms is provided in this chapter.

Supervised learning algorithms use a set of variables as inputs (also predictors, features, independent variables) to predict the value of one or more outputs (responses, dependent variables). Outputs can have one of two general forms: quantitative or qualitative. Quantitative responses include measurements with a range of values, and measurements close in value are also close in nature. In contrast, qualitative responses are part of a finite set of values that can be considered as response classes, denoting them as categorical or discreet variables. Depending on the type of output variable, a supervised learning task is either a regression when predicting a quantitative response or classification when predicting a qualitative response. Since machine learning was used in this thesis to predict the disease status of a sample (healthy or CRC/adenoma), we are speaking of a classification task. [11]

The input data for ML algorithms are usually stored in a matrix form. In this case, we could directly use the ASV table (output of the bioinformatic pipeline). The table contains samples in rows and bacterial taxa (obtained by exchanging ASVs with corresponding taxonomic assignments) in columns (n samples $\times$ m taxa). Samples represent observations and bacterial taxa features (variables). Supervised learning algorithms also require a target vector to store the responses for each sample. This vector is therefore n-dimensional and has a value of 0 for samples that are healthy or a value of 1 for samples with CRC/adenoma (in two separate classification tasks, one for healthy vs. CRC and one for healthy vs. adenoma).

The difference between the ML algorithms arises from the different approaches for separating the classes. This results in different objective functions that ML models are attempting to optimize.

Four classifiers applied in this thesis were linear, LASSO, RIDGE, elastic net, and logistic regression. LASSO and RIDGE both optimise least squares with added penalty terms. Least squares is the basic regression model in which linear coefficients are chosen based on the minimization of the residual sum of squares - the squared difference between the actual response and the predicted/modelled response. LASSO additionally contains a so-called $L1$ norm in the objective function. $L1$ norm represents the sum of the absolute model coefficients. RIDGE contains the $L2$ norm, which is the sum of squared coefficients. The nature of the LASSO constraint results in some of the coefficients being exactly 0, which reduces the number of features because the features with zero coefficients do not contribute to the model. The elastic net uses a combination of $L1$ and $L2$ norms with an adjustable parameter to control the ratio between the penalties. In the logistic regression, the probabilities describing the possible outcomes of an observation are modeled using a logistic function. Using the Newton-Raphson algorithm to solve the problem of maximization of the log-likelihood function, we obtain a solution that can be considered as weighted least squares with an adjusted response. So, although the initial design of the logistic regression algorithm is completely different than the idea behind LASSO/RIDGE/elastic net, it also ends up being connected to the classic least squares problem. [11, 12]

Bayes' theorem generally describes the probability of an event based on prior conditional probability. Gaussian Naive Bayes algorithm for classification assumes that the likelihood of the features is Gaussian (normally distributed) and the parameters are estimated using maximum likelihood. [13, 14]

Random forests are tree-based models. The idea is to partition the feature space into regions and fit a simple model for each region. In each step, a split variable is chosen, and a split point (a value of this variable) defines the border of regions. This is referred to as tree building. Using a node impurity measure (e.g., misclassification error or Gini index), we can define an objective function that contains the chosen impurity measure and tree size (number of nodes, i.e., number of times a partition has been conducted) multiplied by a tuning parameter. This function should be minimized, and thus, the aim is to find a compromise between tree size and goodness of fit (large trees can lead to overfitting). Since single classification trees are very sensitive to small data changes, many trees are generated when using the random forest classifier and the decision is made by assigning an observation to the class which is predicted by the majority of the trees. [11]
Light Gradient Boosting Machine algorithm is also based on decision trees, but has some additional advantages, e.g. faster training and higher efficiency, lower memory usage and better accuracy. [15]

Support vector machines work by transforming the feature space into a higher-dimensional space to separate the data clouds of the groups that overlap. In this space, a hyperplane (in the linear case) is constructed and represents the border between data clouds. An objective function follows from a simple optimization problem, where the aim is to construct a separating hyperplane by finding the largest distance between the two classes of the training data. Hyperplanes are used in the linear case, however, sometimes, the classes can be separated much better with nonlinear decision boundaries. In this case, we can use other functions (also called kernels), e.g. radial basis function. [11]

1.4 Colorectal cancer and adenoma

Colorectal cancer (CRC) is one of the most prevalent cancers worldwide, ranking third among the most common cancers after breast and lung cancers, and is the second leading cause of cancer-related deaths in 2020 [16]. The adenoma-carcinoma sequence has been proposed as a process by which colorectal cancer arises, implying that most, if not all, colorectal carcinomas occur after the development of a malignant adenoma [17]. A colorectal adenoma is an unusual growth of cells formed in the lining of the colon [18]. Most of them are benign, but around 3-5% of affected individuals develop carcinoma during a subsequent period of 10 years [19]. The development of colorectal cancer is attributed to genetic mutations and factors such as diet, inflammation in the intestine and, as recently discovered, the gut microbiota. Dysbiosis of the gut microbiome has been described as a likely mutagenic mechanism by which genotoxic stress is generated in the gut environment, leading to colorectal cancer [20].

Currently, the most widely used screening tools for CRC include colonoscopy as a direct visualization method, and fecal immunochemical test (FIT) and fecal occult blood test (FOBT) as indirect and non-invasive tests [21]. Both non-invasive methods are based on testing for (hidden) blood in stool samples [22, 23]. Compared with symptom-recognized colorectal cancer, CRC detected by either invasive or non-invasive screening tools results in a higher overall survival rate and CRC-specific survival rate [24]. This finding highlights the importance of regular screening. To increase participation, screening tools should be as convenient and non-invasive as possible while maintaining high accuracy. Colonoscopy is the most reliable diagnostic method for CRC with high sensitivity (88.7%) and specificity (90.3%) [25]. However, it also has drawbacks such as the inability to detect cancers in the proximal colon as successfully as in the distal colon, which decreases the initially determined efficiency of this procedure [26]. The accuracy of diagnosis also depends on the experience of the physician who performs it [27]. One of the most accurate and widely used non-invasive methods for CRC screening, FIT, shows a sensitivity of 75% and a specificity of approximately 90% for CRC, however, the sensitivity for detecting advanced adenomas is much lower (20-40%) [28].

As already mentioned, bacteria in the intestinal microbiome can stimulate the development and progression of CRC through processes such as the induction of a chronic inflammatory state or immune response, altering stem cell dynamics, the biosynthesis of toxic and genotoxic metabolites, and affecting host metabolism [29]. A recent review summarized common biomarkers of gut dysbiosis in CRC patients with increased relative abundance of the organisms *Fusobacterium nucleatum*, *Parvimonas micra*, and *Peptostreptococcus anaerobes* [30]. Interestingly, *F. nucleatum* also showed potential for early diagnosis, since the performance of colorectal adenoma detection with fecal immunochemical test (FIT) was increased when combining FIT with the quantification of this biomarker [31]. It was also discovered that using the ratio of *F. nucleatum* to the probiotics *Faecalibacterium prausnitzii* and *Bifidobacterium* resulted in surprisingly good diagnostic performance [32]. A study on gut mucosal microbiome identified taxa with the highest abundances for cancer stages I-III, respectively [33]. Dominant phylotypes for cancer stages 0-III and for stage IV have also been identified in stool samples from young- and old-onset CRC patients [34].

Understanding changes in microbiome composition during CRC development offers a new strategy for the diagnosis of this disease [35]. Cancers diagnosed at earlier stages usually have higher survival rates [36]. This emphasizes the importance of identifying biomarkers specific to adenoma and early stages of cancer. Hence, the first aim of this thesis was to identify microbial signatures in the stool microbiome associated with colorectal adenoma and cancer using a large meta-analysis dataset of amplicon sequences. In addition, the goal was to develop a machine learning tool that can distinguish between health and colorectal cancer or adenoma and thus classify sequenced stool samples. These results were intended to motivate and serve as information for the development of a novel non-invasive screening tool that performs better than the currently available non-invasive methods (FIT and FOBT) and accompanies colonoscopy in the regular screening and prevention of colorectal cancer.

CHAPTER 2

Methods

2.1 Study inclusion

Google Scholar was used to search for studies that included microbiome amplicon sequence reads of stool samples from CRC and adenoma patients as well as healthy controls. The search was conducted in November 2021. Of the several studies initially found, only four met the conditions related to read quality (Q-score) and sequencing platform (Illumina): Baxter et al. [37], Zackular et al. [38], Zeller et al. [39], and Yang et al. [34]. Raw FASTQ files were downloaded from the NIH National Center for Biotechnology Information Sequence Read Archive (SRA) with accession numbers PRJNA290926 (Baxter et al.) and PRJNA763023 (Yang et al.), from the European Nucleotide Archive (ENA) with accession number ERP005534 (Zeller et al.), and from the Mothur Project website (Zackular et al.) [40]. All studies except Yang et al included sequencing information of the V4 region of the 16S rRNA gene, whereas the Yang study included the V3-V4 region. The number of participants for a certain disease status (healthy/adenoma/CRC) is given for each study and the entire meta-analysis dataset in Table 2.1.

2.2 Data preprocessing

Raw reads were processed using the bioinformatics workflow manager Nextflow. The workflow included **DADA2** pipeline (version 1.26) to generate a table of amplicon sequence variants (ASVs) with counts for each sample [9]. In addition to the reads in the FASTQ files, the workflow used user-defined parameters set after inspection of the reads, which were then used during the preprocessing steps (Table 2.2). The tool **Cutadapt** was used in the workflow to trim off any primer sequences and adapters [41]. However, primer removal was not performed in this case because primers were already removed in the published FASTQ files of the chosen studies. **FIGARO** was used to determine the

Table 2.1: Number of samples in the different studies and in the meta-analysis dataset. Because the meta-analysis dataset was created after preprocessing steps that excluded some samples, the number of samples from the studies do not fully sum to the number of samples in the final meta-analysis dataset.

Dataset	healthy	adenoma	CRC
Baxter	172	198	120
Zackular	30	30	30
Zeller	50	38	41
Yang	474	0	564
meta-analysis dataset	739	290	757

optimal trimming parameters (based on error rates) for paired-end reads for the DADA2 pipeline [42]. Flag `-a` was used to input the amplicon length (depending on the 16S rRNA region sequenced). To filter and trim reads in DADA2, the `filterAndTrim()` function was used, with the `truncLen` argument defined with forward and reverse trim position (FIGARO output) and the `maxEE` argument defined with forward and reverse expected error values (FIGARO output). Other arguments were the standard filtering parameters (set as in the DADA2 pipeline tutorial). For single-end reads, the truncation length parameter was set to 0 and `maxEE` was set to 2. After learning the error rates with the `learnError()` function, the core sample inference algorithm was applied with the `dada()` function. Finally, paired reads were merged with the `mergePairs()` function, and in the single-end case, the output of the `dada()` function was simply used to continue. The `makeSequenceTable()` function created the ASV table and chimeras in the table were removed with the `removeBimeraDenovo()` function. Samples with a read count lower than the user-defined sample depth (set to 5000) were removed from the ASV table due to low coverage. The **DECIPHER** (version 2.26.0) function `IdTaxa()` (similar to the DADA2 `assignTaxonomy()` function) was used to classify sequences and determine the confidence percentage for each assigned taxon [43]. Species were added using the `addSpecies()` function with the Silva reference fasta file. The `tryRC` argument was set to TRUE to use the reverse complement of sequences if it has a better match to the reference sequences [44].

The ASV tables from individual studies were merged into a single dataset containing 155146 different ASVs for 1786 samples. Since this dataset caused excessive memory consumption, it was necessary to reduce it before proceeding with further analysis. This was done using a taxonomy map table created with the `IdTaxa()` and `addSpecies()` functions in the pipeline. The dataset was therefore reduced to 1488 species. All subsequent data exploration analyses prior to machine learning (alpha and beta diversity, differential abundance analysis) were performed in R (version 4.2.1). The corresponding scripts with code can be found in the GitHub repository [45].

Table 2.2: User-defined parameters used as input for the preprocessing pipeline for each study dataset. The same (V4) region of the 16S rRNA gene was sequenced in Baxter, Zackular and Zeller studies, hence the same parameters were used for these datasets. Since the Yang study contained both paired-end (PE) reads and single-end (SE) reads, the parameters for each case were different.

	Baxter, Zackular, Zeller	Yang PE	Yang SE
amplicon length	250	460	460
forward length	251	250	427
reverse length	251	250	-

2.3 Alpha diversity and species evenness

Alpha diversity index and species evenness were calculated on a complete data set (ASV level). For alpha diversity, the `diversity()` function from the **vegan** package (version 2.6.2) was used, and for species evenness, the `diversityresult()` function from the **BiodiversityR** package (version 2.14.4), both with Shannon index. Counts for the respective diagnosis groups (healthy, adenoma, CRC) were first tested for normal distribution using the Shapiro-Wilk test. The Kruskal-Wallis test was used to test whether the difference in alpha diversity and evenness between the three groups was significant, followed by Dunn's test for multiple pairwise comparisons with Benjamini-Hochberg (BH) p-value correction to determine which groups were significantly different.

2.4 Beta diversity and data transformation

Beta diversity quantifies the (dis)similarity between samples. The differences between the diagnosis groups (healthy/adenoma/CRC) and the study dataset groups (Baxter/Zackular/Zeller/Yang) were of the greatest interest, so these two groupings were used for this analysis.

Beta diversity was visualized using principal component analysis (PCA) and principal coordinate analysis (PCoA). The `PcaHubert()` function of the **rrcov** package (version 1.7.0) was used to calculate the principal components. This function was chosen because it takes into account the high dimensionality and possible outliers. The `mcd` argument of the function was set to FALSE because only "tall" datasets are possible for the MCD estimator (i.e., the number of observations must be at least twice the number of variables, which is almost never the case for sequencing data). In this way, the ROBPCA algorithm was applied. The number of components k was not fixed, so the algorithm itself finds the optimal number of components to calculate [46].

The `vegdist()` function of the **vegan** package (version 2.6.2) was used to obtain the Aitchison distance object, with the `pseudocount` argument set to 0.000001 to handle zeros in the ASV table. The Aitchison distance is simply the Euclidean distance for centered log-ratio (clr) transformed data. Since sequencing data are compositional data, meaning that the values (counts) for each sample sum to a constant, the transformations and methods used to analyse the data should take into account the compositional nature of the dataset. This property is caused by the assay technology itself, since the number of counts for each sample is limited by an arbitrary total - the library size [47]. An obstacle when dealing with compositional data is that they do not exist in Euclidean space (and should not be treated as such). However, Aitchison found that compositional data can be mapped into real (Euclidean) space by using the log-ratio transformation [48]. This leads to the Aitchison distance being superior to the commonly used Bray-Curtis dissimilarity or Jensen-Shannon divergence, since these measures do not represent a true linear distance [49]. PCoA was therefore performed with the Aitchison distance object using the `pcoa()` function from the **ape** package (version 5.6.2). Logarithmic transformations have another advantage: they eliminate skewness and center the data. This is demonstrated by the example of the first sample in the meta-analysis dataset (Figure A.1). Centered data are better suited to meet the assumptions of linear models used later in machine learning. Clr-transformed data are also scale invariant, meaning that the same ratio is obtained for a sample with few read counts as for an identical sample with many read counts (only the precision of the clr estimate is affected) [49].

Permutational multivariate analysis of variance or PERMANOVA is a method for geometric partitioning of multivariate variation in the space of a chosen dissimilarity measure. The p-values are obtained by distribution-free permutation techniques (without assuming multivariate normality). The null hypothesis is that there are no differences in the positions of the group centroids in the space of the chosen dissimilarity measure [50]. This method was used to test the significance of the difference between the beta diversity of different diagnosis groups and study datasets. The `adonis2()` function of the **vegan** package (version 2.6.2) was used to test overall differences, and pairwise comparisons were tested using the `permanova_pairwise()` function of the **ecole** package (version 0.9.2021).

2.5 Differential abundance analysis

Differential abundance analysis (DAA) is used to determine differences in abundance of microorganisms between two or more groups [51]. Differentially abundant taxa between the groups with a disease (CRC and adenoma) and the healthy group were determined using three different methods. Recommended tools that take into account the compositional nature of microbiome datasets are ALDEx2 and ANCOM [49]. ALDEx2 implements the centered log-ratio transformation and ANCOM implements an alternative approach, additive log-ratio [52]. The `aldex()` function from the **ALDEx2** package (version 1.28.1) was used to analyse differential abundance based on the Wilcoxon rank sum test and Welch's t-test. The **ANCOMBC** package (version 2.1.1) allows bias correction. The `ancombc2()` function was used with study (and diagnosis) variables in the `fix_formula` argument to correct for bias caused by using different study datasets. p-values were corrected using the Benjamini-Hochberg procedure. Because only one taxon was differentially abundant in adenoma vs. healthy DAA using `ancombc2()`, the calculation was repeated without bias correction (without the study variable in the `fix_formula` argument) to obtain a comprehensive list of DA taxa for this comparison. Finally, a popular microbiome-specific method, the **Maaslin2** package (version 1.8.0) with the `Maaslin2()` function was used for differential abundance analysis [52]. The variable study was set as a random effect to account for this covariate.

2.6 Machine learning

Machine learning (ML) models have the ability to learn from the data we present to them (training data) and make predictions about unknown data. This thesis employed nine different learners from the supervised learning category of ML. Because the response was categorical in this case (0 - healthy, 1 - disease (CRC or adenoma); 2 classes), we refer to it as ML classification. [11]

The machine learning pipeline was written in JupyterLab (Python version 3.6.9). Two separate classifications were evaluated, one for CRC vs. healthy and one for adenoma vs. healthy. Multiclass classification (all 3 groups classified by the same model) was also investigated, but the separation between groups was less successful. Because more model optimization, analysis, and interpretation techniques are available for binary classification, this approach was chosen.

Several parameters were implemented in the pipeline to search for optimal settings for the classification tasks (overview in Table 2.3). All parameter combinations were tested using Katib, the Kubernetes-native project for automated machine learning (AutoML) [53], and output metrics were documented and inspected.

Table 2.3: Overview of the available parameter settings for the machine learning.

Taxonomy level	species, genus
Data transformation	compositional, subsampling
Feature selection	none, SelectKBest, Maaslin2
Model	Logistic regression (LR), LASSO, RIDGE, Elastic net (EN), Support vector machine with linear kernel (SVM linear), Support vector machine with radial kernel (SVM), Random forest (RF), Light gradient boosting machine (LGBM), Gaussian naive Bayes (GNB)

For each parameter setting, the feature dataset was split into 80% for training and validation and 20% for testing of the created model at the end. The former part was again divided into 80% for model training and 20% for model validation during 5-fold cross-validation (CV) with 5 repetitions. The output metrics of the created models (including AUC, sensitivity, specificity, accuracy, and F1-score) are the calculated mean of the metrics from the validation portions of the dataset (mean of the 25 values for each metric respectively during CV). All splits were stratified to maintain the class distribution. The workflow is shown in Figure 2.1 [54].

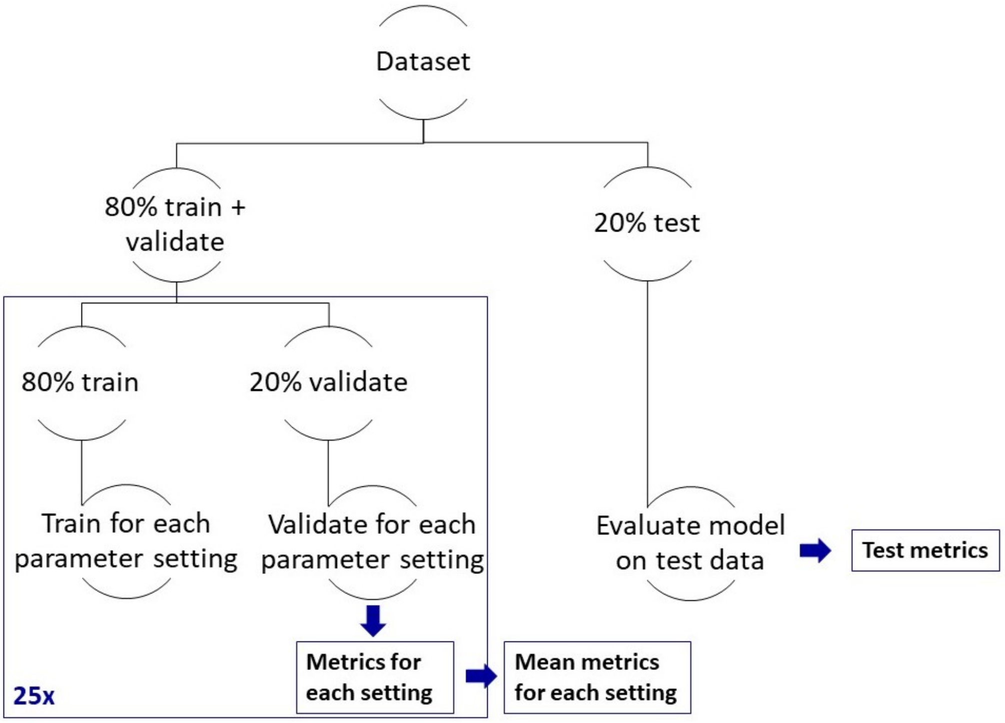

Dataset
80% train + validate
20% test
80% train
20% validate
Train for each parameter setting
Validate for each parameter setting
Metrics for each setting
Mean metrics for each setting
25x
Evaluate model on test data
Test metrics

Figure 2.1: Splitting of the meta-analysis dataset during ML workflow

The machine learning pipeline consisted of the following key steps, which are described in detail in the sections below:

- Importing the required libraries and functions, and setting the desired parameters
- Import of the datasets
- Data transformation
- Feature selection
- Model training and cross-validation
- Model testing
- Model interpretation

Installation of requirements, importing functions and setting hyperparameters

In this part, the required functions were imported from available libraries and self-created modules. A variety of parameters could be set to examine and compare the performance of ML algorithms (Table 2.3).

Data ingestion

All feature datasets (ASV tables) were imported into Python and stored as data frames in a count table dictionary. A taxonomy map dictionary was also created using the information from each taxonomy map. This dictionary was then used to rename the features (ASVs) to the specified taxonomic level so that the feature names contained the entire taxonomy up to the desired taxonomic resolution. Finally, the feature data frames from the count table dictionary were merged into a single feature dataset.

Data transformation

The merged dataset was transformed by either compositional or subsampling transformation. As discussed in Section 2.4, sequencing data are compositional because the counts for each sample sum to a constant, so compositional centered log-ratio transformation was one of the chosen methods. Prior to the transformation, the zeros in the dataset were replaced with a pseudocount (set to 0.000001) and the centred log-ratio transformation was calculated using the function `clr()` to remove the closure effects. In the subsampling transformation, random subsampling to an equal count is performed for each sample. The library **biom** (version 2.1.10) was used and the feature table was first converted to BIOM format with the function `Table()`. Subsampling was performed using the function `subsample()` with default parameters (except for the desired subsampling depth, which was one of the pipeline parameters, but was set constantly to 5000 because the compositional approach was preferred and subsampling was not explored with further subsampling depths).

Feature selection

Before selecting significant features, labelling was performed to create class labels, i.e., the response vector. Based on the metadata information, samples from the control group (healthy) were labelled with zeros and samples from the other diagnosis group (adenoma/CRC) were labelled with ones. The test dataset was also separated prior to feature selection, so that the test dataset remained completely uninvolved in the creation of the model.

If chosen, all features could have been used to train the models. Two other options were SelectKBest (SKB) and Maaslin2, a DAA tool introduced in Section 2.5. SKB is an algorithm based on univariate statistical tests to select k features with the highest scores (k was always set to 50 for the purposes of this analysis) [55]. It was implemented with

sklearn.feature_selection SelectKBest() function with mutual_info_classif scoring function (returns univariate scores of features). Maaslin2 was implemented in a module written in R (version 4.1.2) using the Maaslin2() function with the diagnosis group set as a fixed_effect and the study information as a random_effect to correct for the confounding effect of the study variable. The q-value significance threshold max_significance was left at the default value of 0.25. Before running Maaslin2, MD5 hashing of the feature names was performed to prevent renaming. Subsequently, the hashing was reversed to restore the feature names according to the taxonomy. The same feature selection method was then applied to the test dataset.

The pipeline also allowed removal of any of the datasets or the outliers detected by the robust PCA method (Section 2.4). However, these options were not explored in detail during AutoML as they did not seem to provide significantly better results and it was decided to preserve the entirety of the dataset.

Model training and cross-validation

Nine different classifiers were implemented in the pipeline: logistic regression (LR), LASSO, RIDGE, elastic net (EN), support vector machine with linear kernel (SVM linear), support vector machine with radial kernel (SVM), random forest (RF), light gradient boosting machine (LGBM), and Gaussian Naive Bayes (GNB). The package **sklearn** (version 0.24.2) was used for all models except for LGBM in the pipeline.

Four classifiers were defined using the LogisticRegression() function from the **sklearn.linear_model** module; LR, LASSO, RIDGE and EN. For LR the default settings of the function were used, only the parameter max_iter was adjusted (set globally to 10000 and always used with this function, as the default setting was not sufficient for this dataset). For LASSO, the penalty was set to "l1", C to 1 (1/ALPHA, ALPHA set to 1), and solver to "liblinear". For RIDGE, penalty was set to "l2", C was set to 1/2 (1/2*ALPHA), and solver to "liblinear". For EN, penalty was set to "elasticnet", solver to "saga", and l1_ratio to 0.5. Since the LogisticRegression() function was used for LASSO, RIDGE and EN models, these are not exactly the "real" LASSO, RIDGE and EN algorithms, but rather logistic regression with $L1$, $L2$, combination of $L1$ and $L2$ penalties respectively (see Section 1.3).

SVM models were implemented using the SVC() function from the **sklearn.svm** module. For SVM linear, kernel was set to "linear" and for SVM with radial kernel, the default kernel type "rbf" was used. In both cases, probability was set to TRUE to allow later probability estimates with the model.

For the random forest classifier, the function RandomForestClassifier() was used with the default settings from the module **sklearn_ensemble**.

LBGM was defined using the **lightgbm** package (version 3.3.2) and the LGBMClassifier() function with default settings.

GNB was implemented with the function `GaussianNB()` from the module **sklearn.naive_bayes** with default settings.

The selected model was fitted with training data using `fit()`, and the evaluation metrics were calculated during cross-validation (CV). `RepeatedStratifiedKFold()` with 5 folds (`n_splits`) and 5 repetitions (`n_repeats`) was used. A for loop was used to create a dictionary for each CV split of the training data, containing metrics including precision, recall, and thresholds (output of `precision_recall_curve()`) and predictions for validation datasets. The metrics and predictions from the dictionary were used to plot the receiver operating characteristic (ROC) and precision-recall (PR) curves, and the (non-normalised and normalised) confusion matrix (CM). Because changing the cut-off along the ROC curve allows adjustment of the sensitivity and specificity values, two sets of CM were created in the pipeline, one for the cut-off determined in the `predict()` method and one manually defined to classify the probabilities calculated by `predict_proba()` and thus modify the sensitivity and specificity.

Model testing

The model was then used to predict (`predict()`) the output of the test dataset. The predicted vector was compared to the actual response vector of the test data. ROC curve and other evaluation metrics were calculated based on the performance of the model on the test dataset. Confusion matrix for the test dataset was based on the output of `predict_proba()` and the calculated probabilities were classified using the adjusted cutoff value determined for the training dataset.

Model interpretation

SHAP library (version 0.41.0) was used for model interpretation. In particular, the global summary of the permutation explainer (`explainers.Permutation()`) on training data was used to visualize the n (set to 20) features that contribute to the model the most based on the Shapley values. According to the errors obtained while running the function, the argument `max_evals` should be at least 2*(number of features)+1. However, when the number of features was »249, the kernel died, so the argument `max_evals` was set to 500 to prevent kernel from crashing. This meant that it was impossible to interpret the models with all species. For the tree ensemble models, `TreeExplainer()` was implemented and a beeswarm plot (`plots.beeswarm()`) was created to see how the top features affected the output of the model.

CHAPTER 3

Results and discussion

3.1 Alpha diversity and species evenness

Alpha diversity quantifies the diversity of a community within a sample. The Shannon diversity index and additionally species evenness were chosen as measures for this analysis. Statistical tests were performed to determine if differences in alpha diversity measures were significant among the three diagnosis groups.

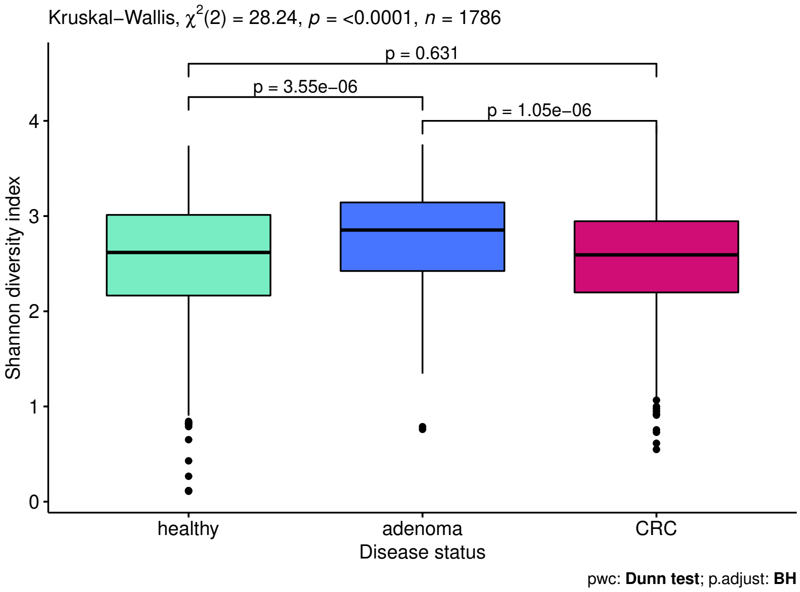


Figure 3.1: Comparison of alpha diversity between the healthy, adenoma, and CRC group. p-values of pairwise comparisons were calculated using Dunn's test with Benjamini-Hochberg correction.

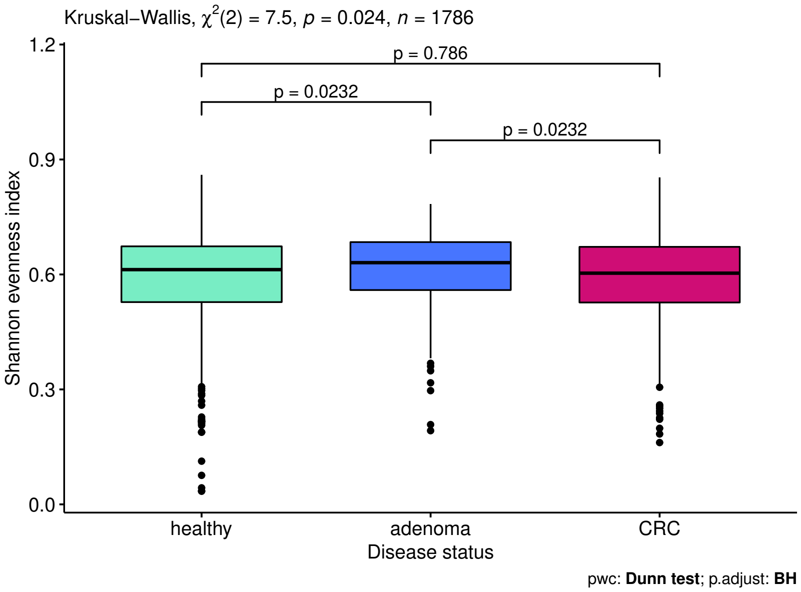


Figure 3.2: Comparison of Shannon evenness index between the healthy, adenoma, and CRC group. p-values of pairwise comparisons were calculated using Dunn's test with Benjamini-Hochberg correction.

Because the Shapiro-Wilk test showed that none of the three groups followed a normal distribution (Table 3.1), the Kruskal-Wallis test was performed to test the significance of differences for all three groups, followed by a Dunn's test for pairwise comparisons. This procedure was applied for all calculations in this chapter since the null-hypothesis of the Shapiro-Wilk test was rejected in all cases.

Table 3.1: p-values of the Shapiro-Wilk and Kruskal-Wallis tests for alpha diversity (Shannon diversity index) and species evenness (Shannon evenness index) to test for normal distribution and significance of differences between groups, respectively.

		healthy	adenoma	CRC
Shannon diversity index	Shapiro-Wilk	7.34e-12	3.33e-08	4.96e-06
	Kruskal-Wallis		7.37e-07	
Shannon evenness index	Shapiro-Wilk	< 2.20e-16	5.04e-10	2.86e-12
	Kruskal-Wallis		2.35e-02	

The Kruskal-Wallis test revealed a significant difference in alpha diversity among the three groups. Dunn's test showed that the difference is significant between the adenoma and CRC group, as well as between adenoma and healthy group, indicating that adenoma patients had significantly higher diversity compared with the other two groups. This is surprising, as it was expected that the healthy group would have the highest alpha diversity. The same results are obtained for the Shannon evenness index, which makes

Table 3.2: p-values of Dunn's test for pairwise comparisons of alpha diversity and species evenness between the three groups. Non-significant p-values are marked with "(ns)".

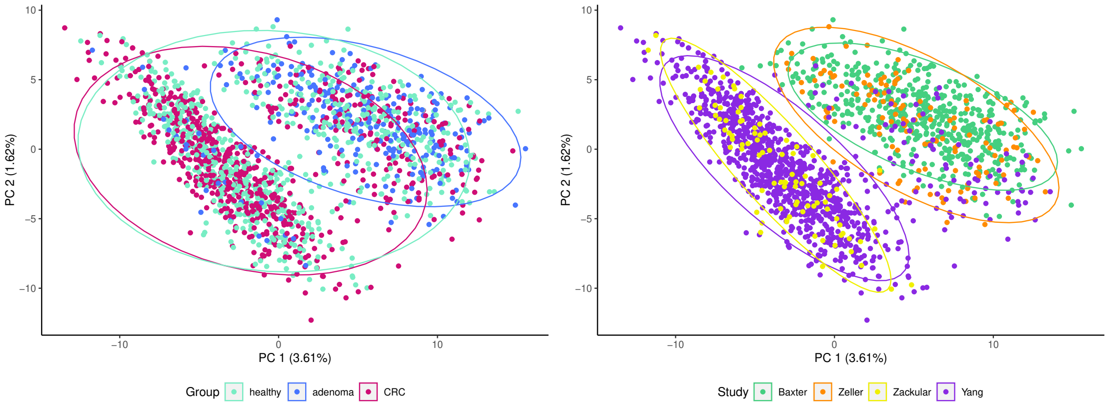
	adenoma - CRC	adenoma - healthy	CRC - healthy
Shannon diversity index	1.05e-06	3.55e-06	6.31e-01 (ns)
Shannon evenness index	2.32e-02	2.32e-02	7.86e-01 (ns)

sense since the Shannon evenness index is directly correlated with and calculated from the Shannon diversity index. However, the differences between the groups are less obvious here, as indicated by higher p-values.

Zeller et al. reported no significant changes in Shannon diversity or species and gene richness between the healthy, adenoma, and CRC group. The results of the Yang et al. paper showed a significant decrease in Shannon diversity index in old-onset CRC compared with an age-matched control group and a significant decrease in diversity of young-onset CRC patients compared with an age-matched control group. Interestingly, old-onset CRC patients also had a significantly lower diversity compared to young-onset CRC patients. In general, studies on healthy individuals have shown that the diversity of the microbiome increases with age or does not change significantly [56, 57]. As mentioned earlier, an intuitive hypothesis was that the healthy microbiome would be the most diverse, as low alpha diversity is often associated with a dysbiotic gut microbiome [57]. Inclusion of the samples from the Yang dataset could explain the diversity level observed in the CRC group, which is comparable to that of the healthy group. The alpha diversity values of 185 young-onset-CRC patients (age 36-46; making up 24.4% of the samples in the CRC group) could have a "balancing" effect on the overall alpha diversity of this group. Baxter et al. and Zackular et al. did not examine microbial diversity in their datasets. Other publications report differing results regarding diversity in fecal samples between groups, with some finding significant changes [58, 59, 60, 61] and some non-significant differences [62, 63, 64]. Sheng et al. reported a significant difference, however, with CRC patients having higher Shannon and Simpson diversity than healthy controls [59]. Overall, alpha diversity does not appear to be a meaningful metric for comparing the microbiomes of CRC and adenoma patients with healthy controls due to inconsistent results in the literature. In addition, ethnicity and geographic location are known to have a significant impact on gut microbiome composition [65, 56]. It is therefore difficult to analyse alpha diversity in a meta-analysis dataset consisting of samples from different countries and continents and to compare the groups, considering that each of them has a slightly different representation of different geographic locations. In this case, the adenoma group is the most distinct because it does not include samples from the Yang et al. dataset, which could explain the observed significant difference to other two groups.

3.2 Beta diversity

To analyse and visualise the (dis)similarity between sample groups, beta diversity was examined. Two ordination methods were used, principal component analysis (PCA) on a clr-transformed dataset (Figure 3.3) and principal coordinate analysis (PCoA) on an Aitchison distance object (Figure 3.4). In addition, (pairwise) PERMANOVA was used to quantify multivariate community-level differences between groups. The (dis)similarity was inspected between the three diagnosis groups as well as the four different study datasets present in the meta-analysis dataset.

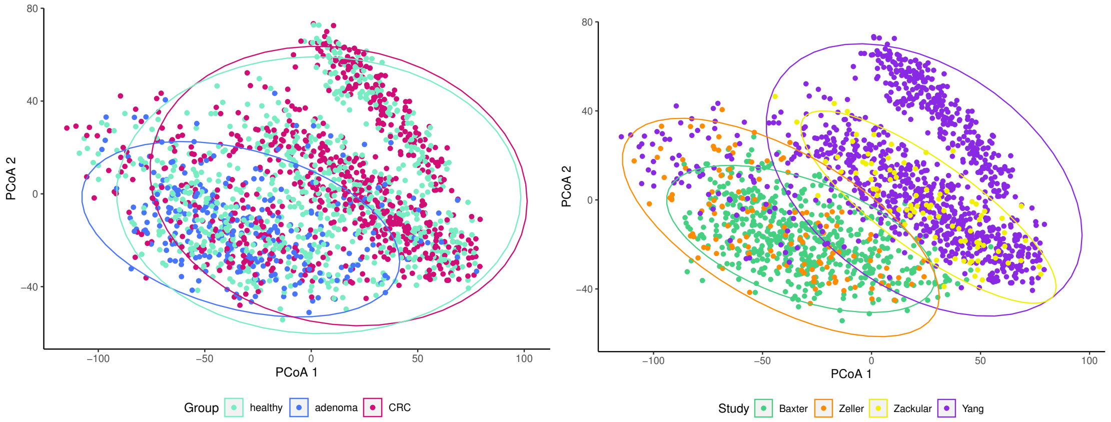


(a) PCA, distinction between diagnosis groups (b) PCA, distinction between study datasets

Figure 3.3: Principal component analysis (PCA) plot with the first two principal components (10 principal components calculated) and the distinction between different diagnosis groups (a) and study datasets (b).

As indicated in Methods (Section 2.4), the number of components k in the PCA function was determined by the algorithm's for finding the optimal k [46]. This resulted in 10 calculated principal components (PCs), which cumulatively explained only 12.6% of the variance. Therefore, different numbers of components were calculated and cumulative variance was checked to see if a satisfactory percentage of the cumulative explained variance (80-95%) could be achieved (Table A.1). In addition, screeplots (Figure A.2) were examined to possibly find the elbow point and determine the cut-off for the number of components, but this was not readily possible. Since the usual criteria (percentage of cumulative variance, elbow point in a screeplot) for determining the number of principal components could not be readily obtained, the 10 PCs calculated according to the algorithm's criterion were taken. Since PCA was used for two-dimensional visualisation of beta diversity, the calculation of 10 PCs was acceptable in this case because it saved some computational time and the information contained in the first two PCs did not change significantly when the number of components was varied.

Both ordination methods imply that the healthy and CRC groups are more widely dispersed, while the adenoma group is concentrated in a more limited area. This is



(a) PCoA, distinction between diagnosis groups (b) PCoA, distinction between study datasets

Figure 3.4: Principal coordinates analysis (PCoA) plot for Aitchison distance (1206 coordinates calculated) with the distinction between different diagnosis groups (a) and study datasets (b).

because of the Yang dataset, since it clusters "away" from the Baxter and Zeller datasets without containing any adenoma samples. The differences between the study datasets are more apparent on the plot. The Yang and Zackular datasets are closer together and separated from the Baxter and Zeller datasets, which also form an overlapping cluster. However, the clusters for both groupings cannot be clearly separated. The overlap of Yang and Zackular or Baxter and Zeller datasets is unknown. As discussed in Section 3.1, geography (and other attributes such as race and diet) is an important factor influencing the composition of the gut microbiome. However, the overlapping study datasets do not consist of samples from people from the same locations and with the same race. Although there is no explicit information about each sample in the metadata, Yang samples originate from China and Zackular samples from the United States (with 85.2% of the dataset coming from white individuals and only 9.1% from Asian individuals). The Baxter dataset is also from the USA, with 90.7% of the samples coming from white individuals. The Zeller samples are from European countries (Germany, France, Denmark, and Spain) and contain no information on race. The unique feature of the Yang dataset is also the targeted 16S rRNA region during sequencing (V3-V4 versus V4 in all other studies). This meta-analysis dataset has a very high dimensionality (approximately 1500 dimensions) and is obviously very complex, as PCA did not reveal a high proportion of explained information (variance) in the first components. This suggests that the observed separation in the 2D graphs in Figure 3.3 and Figure 3.4 does not tell much and represent the real situation, as it captures only a tiny part of the information contained in the dataset.

PERMANOVA results show a significant difference in microbial distance between all diagnosis groups and study datasets overall (p-value 0.001 for both distinctions, Table 3.3) and in pairwise comparisons (adjusted p-value 0.003 for all comparisons in diagnosis groups distinction (Table 3.4) and 0.006 when distinguishing study datasets (Table 3.5)). Zeller et al. reported a significant difference in taxonomic community composition between the healthy and CRC groups. Yang et al. also found significant differences for all comparisons (old-onset CRC vs. age-matched controls, old-onset CRC vs. young-onset CRC, young-onset CRC vs. age-matched controls, but also between old and young control groups). These frequently reported significant differences in taxonomic composition between diagnosis groups suggest that there is an apparent shift in the composition of the microbiome in patients with disease. As will be shown in the next chapters, this dysbiosis is consistent in the literature and in the results of the various methods used here and can therefore be used as an indicator for CRC. The significant difference between study datasets suggests that this information should be adequately accounted for in further analysis methods, as it could introduce bias. Therefore, it was included as a confounding variable (when possible) in the differential abundance analysis.

Table 3.3: Results of PERMANOVA for differences between diagnosis groups and study datasets.

	Df	SumOfSqs	R2	F	p-value
diagnosis groups	2	489352.5	0.017	15.852	0.001
Residual	1783	27521194.8	0.983		
Total	1785	28010547.3	1.000		
study datasets	3	1937838	0.069	44.149	0.001
Residual	1782	26072710	0.931		
Total	1785	28010547	1.000		

Table 3.4: Results of pairwise PERMANOVA for differences between diagnosis groups.p-values adjusted with Bonferroni correction.

pairs	SumOfSqs	F.Model	R2	p-value	p adjusted
healthy vs. adenoma	311598.15	20.315	0.019	0.001	0.003
healthy vs. CRC	44708.06	2.876	0.002	0.001	0.003
adenoma vs. CRC	464628.20	30.230	0.028	0.001	0.003

Table 3.5: Results of pairwise PERMANOVA for differences between study datasets. p-values adjusted with Bonferroni correction.

pairs	SumOfSqs	F.Model	R2	p-value	p adjusted
Baxter vs. Zeller	183519.94	12.644	0.019	0.001	0.006
Baxter vs. Yang	1438839.15	99.040	0.059	0.001	0.006
Baxter vs. Zackular	381315.12	27.868	0.043	0.001	0.006
Zeller vs. Yang	597343.71	39.452	0.033	0.001	0.006
Zeller vs. Zackular	353347.25	22.968	0.097	0.001	0.006
Yang vs. Zackular	64911.54	4.416	0.004	0.001	0.006

3.3 Differential abundance analysis

To investigate the differences in microbial composition between the two groups with disease (colorectal carcinoma/adenoma) and the healthy group, a differential abundance analysis was performed. The aim of this analysis was to find universal biomarkers for the intestinal dysbiosis that occurs in the development of colorectal cancer. Three tools were used, ALDEx2, ANCOM-BC and MaAsLin2, as it is recommended to use different approaches to obtain robust results. In a comparison of 14 different DAA methods, Nearing et al. concluded that ALDEx2 and ANCOM-II provided the most consistent results [52].

Organisms with a significant difference in abundance between groups are summarized in tables for each comparison and method. Only 10 taxa with the highest effect size values were listed for each sample group because analysis of all identified taxa would be too extensive. Evidence found in the literature regarding the function and effects of specific organisms on colorectal cancer development and their reported abundance in samples from other CRC related studies was discussed.

3.3.1 Colorectal cancer vs. healthy

ALDEx2

ALDEx2 identified 36 differentially abundant taxa enriched in the healthy group and 9 taxa enriched in the CRC group (supplementary files S1 and S2). Only taxa with significant corrected p-values (< 0.05) from both Welch's and Wilcoxon rank sum tests were considered differentially abundant. Bias correction was not performed because it was implemented for this method.

Enriched in healthy

Among the ten taxa with the highest effect size values in the healthy group (Table 3.6) are several bacteria from the families *Lachnospiraceae* (*Lachnospira pectinoschiza, Lachnospiraceae, Lachnospiraceae ND3007 group, [Eubacterium] eligens group, [Eubacterium] hallii group*) and *Ruminococcaceae* (*Faecalibacterium prausnitzii, Ruminococcus*). In a study comparing the composition of the microbiome between healthy controls and patients with some form of intestinal disease, the family *Lachnospiraceae* had a high relative abundance in healthy controls compared to CRC patients [66]. *Ruminococcaceae UCG-003* was also enriched in healthy controls in this comparison [66]. Most bacteria from these two families belong to the so-called SCFA (shortchain fatty acids) producers. These compounds are formed during the fermentation of some carbohydrates (including dietary fibre) that cannot be digested by humans. Degradation by the microbiota leads to the production of various SFCAs, e.g. acetate, propionate, butyrate, formate, succinate, which can be ingested by the host [67]. SFCAs are important for the control of inflammatory processes in the gut and interact directly with the host immune system. Studies have reported an association between higher SFCA levels and improved epigenetic state of host histones, as well as a reduction in inflammatory markers. Propionate and acetate

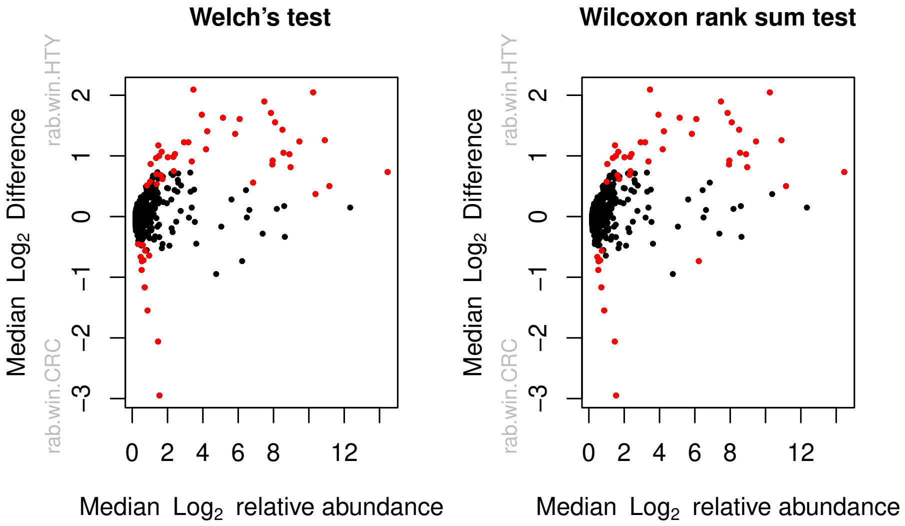


Figure 3.5: ALDEx2 result plots for DAA between CRC and healthy group. Red dots represent species identified by Welch's (left) or Wilcoxon rank sum test (right) (species with BH corrected p-values < 0.05). Association between the relative abundance and the magnitude of the difference per sample is shown in this plot. [51]

have been shown to promote the accumulation of Treg (regulatory T-cells important for self-antigen tolerance and autoimmune disease prevention) in the colon, while butyrate and propionate enhance Treg differentiation [68]. Dietary fibre intake has been shown to correlate with the abundance of SFCA-producing bacteria, including *Eubacterium* [69, 67]. Studies have also shown that the Western diet leads to a decrease of this and other desirable taxa, while the Mediterranean diet leads to an increase in *Eubacterium spp.* in the gut [67]. *Faecalibacterium prausnitzii* is an important butyrate producer with demonstrated anti-inflammatory and gut microbiota modulating properties. Cell-free supernatant of *F. prausnitzii* suppressed colorectal cancer cell growth *in-vitro* and researchers suggest that probiotic supplementation of *F. prausnitzii* may be beneficial for CRC prevention and management [70].

Monoglobus pectinilyticus was detected as a highly prevalent species in the healthy group compared to CRC patients based on whole-genome shotgun sequenced fecal samples [71].

Family *Peptostreptococcaceae* was found to be enriched in tissue samples from CRC patients compared with samples from healthy controls [63]. The well-known CRC enriched species *Peptostreptococcus stomatis*, identified both here and in the literature, also belongs to this family. It is therefore unexpected that this family was identified as enriched in the healthy group by all three methods.

Erysipelotrichaceae family was previously found to be increased in the lumen of CRC

Table 3.6: Differentially abundant taxa with the highest effect-size values identified by ALDEx2 in CRC vs. healthy comparison. Top 10 taxa enriched in the healthy group and all (9) identified taxa enriched in the CRC group. we.eBH- BH corrected p-value from Welch's test; wi.eBH- BH corrected p-value from Wilcoxon rank sum test; effect-effect size.

enriched in healthy	we.eBH	wi.eBH	effect
Lachnospira pectinoschiza	$1.58 \cdot 10^{-09}$	$2.23 \cdot 10^{-10}$	0.26
family *Lachnospiraceae*	$3.80 \cdot 10^{-12}$	$2.00 \cdot 10^{-13}$	0.25
genus *Monoglobus*	$1.01 \cdot 10^{-14}$	$1.66 \cdot 10^{-16}$	0.24
family *Peptostreptococcaceae*	$7.95 \cdot 10^{-11}$	$3.00 \cdot 10^{-12}$	0.24
Faecalibacterium prausnitzii	$4.79 \cdot 10^{-11}$	$1.00 \cdot 10^{-12}$	0.22
genus *Lachnospiraceae ND3007 group*	$4.92 \cdot 10^{-07}$	$1.13 \cdot 10^{-07}$	0.21
genus *[Eubacterium] eligens group*	$2.42 \cdot 10^{-06}$	$5.68 \cdot 10^{-07}$	0.19
genus *Ruminococcus*	$7.21 \cdot 10^{-08}$	$5.46 \cdot 10^{-08}$	0.19
Erysipelotrichaceae UCG-003 bacterium	$2.28 \cdot 10^{-07}$	$1.49 \cdot 10^{-07}$	0.18
genus *[Eubacterium] hallii group*	$1.08 \cdot 10^{-03}$	$3.42 \cdot 10^{-04}$	0.18
enriched in CRC			
genus *Parvimonas*	$1.50 \cdot 10^{-26}$	$1.15 \cdot 10^{-24}$	−0.41
genus *Fusobacterium*	$1.71 \cdot 10^{-13}$	$4.66 \cdot 10^{-11}$	−0.27
Peptostreptococcus stomatis	$2.87 \cdot 10^{-13}$	$4.34 \cdot 10^{-09}$	−0.24
genus *Porphyromonas*	$2.07 \cdot 10^{-09}$	$5.84 \cdot 10^{-06}$	−0.19
Dialister pneumosintes	$4.73 \cdot 10^{-05}$	$1.67 \cdot 10^{-03}$	−0.16
Fusobacterium nucleatum	$2.71 \cdot 10^{-04}$	$8.56 \cdot 10^{-03}$	−0.13
Gemella morbillorum	$1.25 \cdot 10^{-03}$	$1.91 \cdot 10^{-02}$	−0.11
genus *Peptostreptococcus*	$5.88 \cdot 10^{-05}$	$8.07 \cdot 10^{-03}$	−0.11
genus *Hungatella*	$7.07 \cdot 10^{-03}$	$4.43 \cdot 10^{-02}$	−0.09

patients [63]. There are no reports of it's abundance in stool samples.

Enriched in CRC

Genus *Parvimonas* and in particular species *Parvimonas micra* have been associated with colorectal cancer in the literature. This bacterium participates in the development of CRC by altering immune responses and promoting inflammation in the intestine [72]. Multivariate analysis also showed that *P. micra* is a risk factor for poor survival in CRC patients [73]. In addition, the same study analyzed $Apc^{Min/+}$ mice (mice with multiple intestinal neoplasia [74]) colonized with *P. micra* and found significantly higher tumor burden and tumor load in these mice. Colonization with *P. micra* also led to upregulation of genes involved in cell proliferation, stemness, angiogenesis, invasiveness, and metastasis. It enhanced the infiltration of Th17 cells and the expression of cytokines secreted by Th17 cells (Il-17, Il-22, and Il-23), which play a role in the development of

CRC, in the colon of $Apc^{Min/+}$, as well as conventional and germ-free mice [73].

Fusobacterium nucleatum is frequently enriched in the microbiome of CRC patients. *F. nucleatum* binds to CRC cells with virulence factors Fap2 and FadA and lipopolysaccharide (LPS). Signalling by CRC cells' receptors activates nuclear factor kappa B (NF-κB) (regulatory protein complex), which increases the expression of pro-inflammatory cytokines and oncogenes, leading to DNA damage and inflammatory responses and causing tumour progression [72]. Fap2 factor also binds to immune cells and causes immunosuppression [75]. Another mechanism of this bacterium is the recruitment of tumour-infiltrating immune cells, creating a pro-inflammatory microenvironment and promoting the progression of colorectal neoplasia [75]. *F. nucleatum* levels in stool were found to be higher in later stages of cancer, suggesting that this bacterium may have an impact on infiltration of CRC [76]. It may also be associated with more invasive cancer development, as one study showed that the frequency of patients with lymph node metastases was higher in the *Fusobacterium nucleatum* over-abundance group than in the under-abundance group [77].

Flynn et al. proposed a model of colonization and persistence of oral bacterial communities (present during periodontitis - an inflammatory disease in the mouth) in the colon that create a microenvironment for colon lesions ("oral-microbe-induced colorectal tumorigenesis model" [78]) [79]. Oral bacteria associated with CRC include *Peptostreptococcus stomatis* (but also, for example, the previously discussed *P. micra* and *F. nucleatum*). *P. stomatis* is a producer of saccharolytic and fermented products, which may explain its association with CRC, as it likely contributes to the acidic and hypoxic tumor microenvironment that supports bacterial colonization [80]. Another commonly reported species in CRC from *Peptostreptococcus* genus is *P. anaerobius*. Using an $Apc^{Min/+}$ mouse model, Long et al. reported that this species is associated with CRC via a signaling pathway involving NF-κB activation and induced cytokine and interleukin secretion [81]. Another study based on a mouse model showed that *P. anaerobius* is involved in the development of CRC by promoting cholesterol biosynthesis [82].

A commonly reported species associated with CRC from the identified differentially abundant *Porphyromonas* genus is *Porphyromonas gingivalis*. This bacterium is also found in periodontitis patients. Mouse models showed that *P. gingivalis* recruits tumour-infiltrating immune cells by regulating NLRP3 inflammasome (protein of the innate immune system [83]) activity, creating a proinflammatory microenvironment and promoting the progression of colorectal neoplasia [84].

Dialister pneumosintes is another species commonly found in periodontitis patients, and it has also been identified in association with CRC [71, 85]. *Gamella morbillorum* may have an immunosuppressive function in the development of CRC, as it has been shown to lower interleukin IL-12 levels and cleave IgA1 in oral infections in mice, allowing the bacteria to bypass the protective functions of the adaptive immune response [86, 87, 88]. CRC tumor cell development can be caused by epigenetic pathways such as silencing of tumor suppressor genes (TSGs) by e.g., promoter hypermethylation. Xia et al. reported CDX_2 (a TGS) promoter hypermethylation, upregulation of DNA methyltransferase, and promotion of colonic epithelial cell proliferation in germ-free and conventional mice by *Hungatella hathewayi* [89].

ANCOM-BC

ANCOM-BC with bias correction for the variable "study" (four different study datasets in the meta-analysis dataset) detected 63 differentially abundant taxa enriched in the healthy group (supplementary file S5) and 18 taxa enriched in the CRC group (supplementary file S6) (BH corrected p-value q < 0.05).

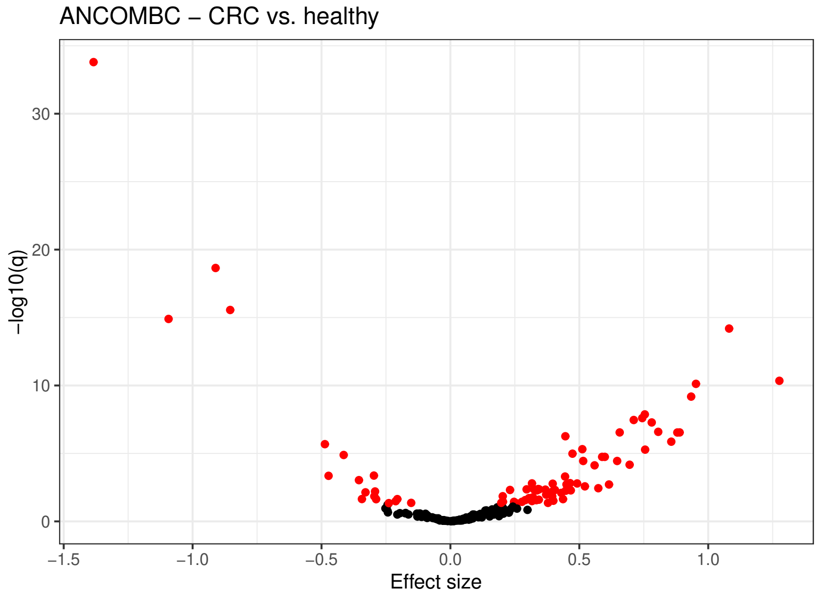


Figure 3.6: ANCOM-BC result plot for DAA between CRC and healthy group. Red dots represent differentially abundant species with BH corrected p-value q < 0.05. Taxa with effect size > 0 are enriched in the healthy group, whereas those with negative effect size are enriched in the CRC group.

Enriched in healthy

The genus *Subdoligranulum* was reported to be enriched in the non-cancer group compared to a colorectal cancer group in DAA from a study that analysed amplicon sequences from different sample types [90].

Megamonas hypermegale was reported as enriched in stool samples from healthy controls compared to CRC samples [91] and this organism was thought to produce SCFAs [92].

Enriched in CRC

The enrichment of *Ruminococcaceae UBA1819* in the CRC group is somewhat surprising, as the *Ruminococcaceae* family is involved in the production of gut-beneficial short-chain

Table 3.7: Differentially abundant taxa with the highest effect-size values identified by ANCOM-BC in CRC vs. healthy comparison. Top 10 taxa enriched in the healthy group and in the CRC group, respectively. q- BH corrected p-value; lfc- log-fold change, measure for effect size.

enriched in healthy	q	lfc
Faecalibacterium prausnitzii	$4.51 \cdot 10^{-11}$	1.28
genus *Monoglobus*	$6.45 \cdot 10^{-15}$	1.08
Lachnospira pectinoschiza	$7.55 \cdot 10^{-11}$	0.95
family *Peptostreptococcaceae*	$6.56 \cdot 10^{-10}$	0.93
Erysipelotrichaceae UCG-003 bacterium	$2.84 \cdot 10^{-07}$	0.89
genus *Faecalibacterium*	$2.88 \cdot 10^{-07}$	0.88
genus *Subdoligranulum*	$1.36 \cdot 10^{-06}$	0.86
genus *[Eubacterium] eligens group*	$2.57 \cdot 10^{-07}$	0.81
genus *Megamonas*	$5.25 \cdot 10^{-08}$	0.78
genus *Ruminococcus*	$5.26 \cdot 10^{-06}$	0.76
enriched in CRC		
genus *Parvimonas*	$1.60 \cdot 10^{-34}$	-1.38
genus *Fusobacterium*	$1.27 \cdot 10^{-15}$	-1.09
Peptostreptococcus stomatis	$2.25 \cdot 10^{-19}$	-0.91
genus *Porphyromonas*	$2.77 \cdot 10^{-16}$	-0.85
genus *Hungatella*	$2.08 \cdot 10^{-06}$	-0.49
family *Ruminococcaceae* genus *UBA1819*	$4.43 \cdot 10^{-04}$	-0.47
Gemella morbillorum	$1.29 \cdot 10^{-05}$	-0.41
Eisenbergiella tayi	$9.23 \cdot 10^{-04}$	-0.36
genus *Akkermansia*	$2.27 \cdot 10^{-02}$	-0.34
family *Prevotellaceae*	$7.26 \cdot 10^{-03}$	-0.33

fatty acids (SCFAs) and iso-butyrate, and this taxon is also commonly associated with a healthy gut [93, 66].

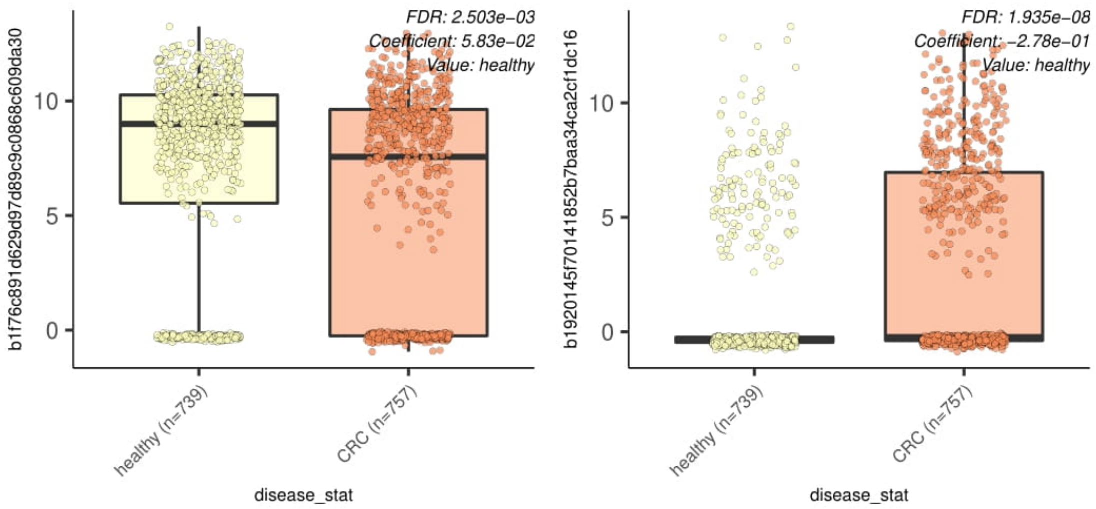
Eisenbergiella tayi was recently discovered in shotgun metagenomic sequencing in association with CRC [94, 95]. It was shown that *E. tayi* is highly enriched in the left colon cancer [95].

The genus *Akkermansia* was found to be significantly enriched in CRC samples in this dataset. A 2021 study found that *A. muciniphila* is significantly decreased in patients with colorectal cancer or adenoma compared to healthy controls. This study also showed that oral administration of *A. muciniphila* suppressed colon tumorigenesis in a mouse model [96]. However, a more recent study (2022) in mouse models showed that administration of this particular species resulted in more intestinal tumors and more colon damage. It also induced more Ki67+ proliferating cells (tumor proliferation), higher expression of proliferating cell nuclear antigen (PCNA), and increased gene expression of proliferation-associated molecules. This suggests that *A. muciniphila* may promote the formation of CRC by increasing early inflammatory levels and enhancing proliferation of intestinal epithelial cells [97].

Prevotellaceae family was enriched in the intestinal lumen of CRC patients [63]. *Prevotellaceae* has also been enriched in obese women [98] and epidemiological studies have found a strong association between obesity and colorectal cancer [63].

MaAsLin2

79 taxa were identified as enriched in the healthy group and 24 taxa were identified as enriched in the CRC group by MaAsLin2 with bias correction for using different study datasets (supplementary files S10 and S11).



(a) *Faecalibacterium prausnitzii*

(b) *Fusobacterium* genus

Figure 3.7: MaAsLin2 plots for two detected differentially abundant taxa in CRC vs. healthy DAA. *Faecalibacterium prausnitzii* (a) is enriched in healthy samples and *Fusobacterium* (b) is enriched in CRC samples.

Enriched in healthy

Parasutterella can ferment inulin and produce SCFAs [99]. This genus has been shown to support interspecies metabolic interactions in the intestine and may have a beneficial effect on intestinal mucosal homeostasis [100]. However, levels of this taxon have been reported to be elevated in CRC patients [101].

Enriched in CRC

Solobacterium moorei (and its higher taxonomic rank, family *Erysipelotrichaceae*) in the oral microbiome has been associated with reduced CRC risk [102]. However, in

Table 3.8: Differentially abundant taxa with the highest effect-size values identified by MaAsLin2 in CRC vs. healthy comparison. Top 10 taxa enriched in the healthy group and in the CRC group, respectively. qval- BH corrected p-value; coef- coefficient of the linear model, measure for effect size.

enriched in healthy	qval	coef
genus *Monoglobus*	$7.30 \cdot 10^{-17}$	3.57
Lachnospira pectinoschiza	$5.95 \cdot 10^{-13}$	3.46
Faecalibacterium prausnitzii	$8.89 \cdot 10^{-10}$	3.07
genus *[Eubacterium] ventriosum group*	$1.52 \cdot 10^{-10}$	3.06
family *Peptostreptococcaceae*	$2.10 \cdot 10^{-10}$	2.92
genus *Parasutterella*	$1.67 \cdot 10^{-10}$	2.85
Erysipelotrichaceae UCG-003 bacterium	$2.39 \cdot 10^{-08}$	2.84
genus *Lachnospiraceae ND3007 group*	$1.59 \cdot 10^{-09}$	2.76
genus *Megamonas*	$1.08 \cdot 10^{-10}$	2.70
genus *[Eubacterium] eligens group*	$8.93 \cdot 10^{-08}$	2.66
enriched in CRC		
genus *Parvimonas*	$2.20 \cdot 10^{-46}$	-5.76
Peptostreptococcus stomatis	$1.18 \cdot 10^{-33}$	-4.07
genus *Fusobacterium*	$3.38 \cdot 10^{-20}$	-4.02
genus *Porphyromonas*	$1.45 \cdot 10^{-26}$	-3.40
genus *Hungatella*	$3.17 \cdot 10^{-09}$	-2.00
Gemella morbillorum	$1.10 \cdot 10^{-09}$	-1.85
Eisenbergiella tayi	$9.97 \cdot 10^{-06}$	-1.52
Solobacterium moorei	$4.65 \cdot 10^{-08}$	-1.51
order *Oscillospirales* family *UCG-011*	$5.62 \cdot 10^{-04}$	-1.36
genus *Intestinimonas*	$4.73 \cdot 10^{-04}$	-1.29

a gut microbiome study, *S. moorei* was found to be enriched in CRC fecal samples [103]. A study examining the oral microbiota in relation to the gut microbiota found that *Solobacterium spp.* had significantly higher relative abundance in CRC patients compared with controls, and that *S. moorei* had significantly higher levels in the CRC advanced-stage group than in the early-stage group in both saliva and stool samples [104].

Oscillospirales is a common gut commensal anaerobe, but was identified here as enriched in CRC. Tran et al. indicated that there is potential competition between tumour-associated taxa and common gut anaerobes (such as *Oscillospirales* and *Lachnospiraceae*), as tumour-associated ASVs showed a strong negative correlation with gut commensals in their research [105]. However, this should suggest that *Oscillospirales* would be depleted in the CRC group, which is not the case according to the MaAsLin2 results.

A study analysing fecal metagenomic sequencing data from young CRC patients (20-49 years of age) compared with older CRC patients and healthy controls reported the enrichment of *Intestinimonas butyriciproducens* in young CRC samples [106]. Because the meta-analysis dataset analysed here consists of young-onset CRC patients (aged 36-46 years; representing 24.4% of the CRC group) from the Yang et al. study, it is not surprising that these results are consistent. *I. butyriciproducens* is a common butyrate producer, and although butyrate has generally been shown to be beneficial to the gut, some studies have reported enhancement of colonic neoplasia development in rats [107] and induction of colon cancer in mouse models [108], which has led to controversy about the effect of butyrate on gut health [109]. Several other SCFA producers mentioned above have also been identified as enriched in the CRC group.

3.3.2 Adenoma vs. healthy

ALDEx2

29 taxa were identified as enriched in the healthy group (supplementary file S3) and 88 as enriched in the adenoma group (supplementary file S4).

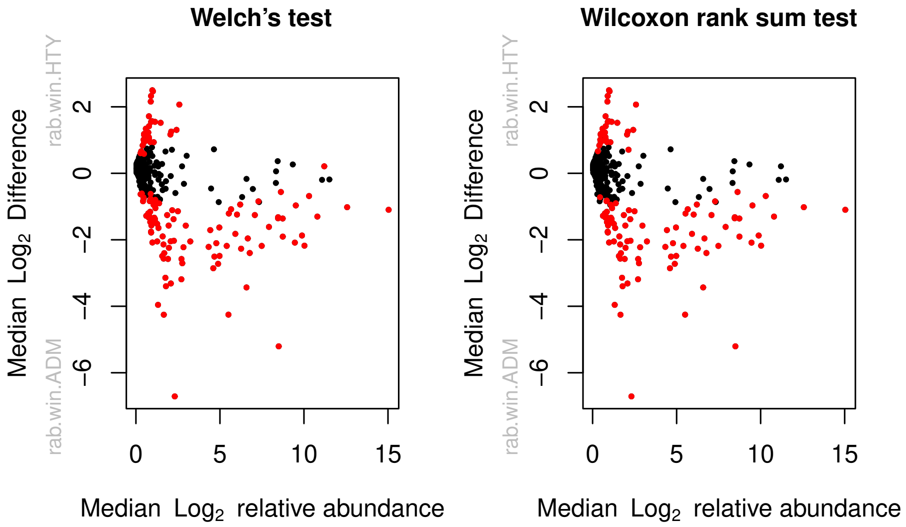


Figure 3.8: ALDEx2 result plots for DAA between adenoma and healthy group. Red dots represent species identified by Welch's (left) or Wilcoxon rank sum test (right) (species with BH corrected p-values < 0.05). Association between the relative abundance and the magnitude of the difference per sample is shown in this plot. [51]

Enriched in healthy

Anaerostipes hadrus is another butyrate producer that has been reported to be more abundant in the feces of healthy control subjects compared to samples from CRC patients [110, 60].

Bifidobaterium longum is a well-known probiotic bacterium that has been shown in studies to suppress mutagen-induced colon tumor development, have antigenotoxic effects, suppress colon tumor incidence, and reduce tumor volume when administered orally to rats [111].

Bacteroides thetaiotaomicron regulates the intestinal immune system and its colonisation in the intestine and interaction with the host leads to strengthening of the mucosal barrier against pathogens [112]. Administration of *B. thetaiotaomicron* alleviated the clinical signs of induced colitis in mice [113].

Shigella is generally responsible for infections in humans and causes the disease shigellosis with symptoms such as dysentery and fever [114]. The abundance of *Escherichia-Shigella*

Table 3.9: Differentially abundant taxa with the highest effect-size values identified by ALDEx2 in adenoma vs. healthy comparison. Top 10 taxa enriched in the healthy in the adenoma group. we.eBH- BH corrected p-value from Welch's test; wi.eBH- BH corrected p-value from Wilcoxon rank sum test; effect- effect size.

enriched in healthy	we.eBH	wi.eBH	effect
Anaerostipes hadrus	$5.14 \cdot 10^{-20}$	$2.13 \cdot 10^{-11}$	0.38
Bifidobacterium longum	$6.67 \cdot 10^{-22}$	$1.07 \cdot 10^{-11}$	0.35
Bacteroides thetaiotaomicron	$5.56 \cdot 10^{-17}$	$3.98 \cdot 10^{-09}$	0.35
Escherichia-Shigella	$4.94 \cdot 10^{-24}$	$7.32 \cdot 10^{-12}$	0.33
genus *Haemophilus*	$4.24 \cdot 10^{-12}$	$1.71 \cdot 10^{-07}$	0.27
genus *Parabacteroides*	$3.69 \cdot 10^{-08}$	$1.08 \cdot 10^{-07}$	0.26
genus *Collinsella*	$1.39 \cdot 10^{-09}$	$1.71 \cdot 10^{-05}$	0.22
family *Lachnospiraceae* genus *CAG-56*	$2.43 \cdot 10^{-09}$	$5.32 \cdot 10^{-05}$	0.22
Brevundimonas mediterranea	$2.51 \cdot 10^{-06}$	$1.38 \cdot 10^{-04}$	0.21
genus *Phascolarctobacterium*	$5.92 \cdot 10^{-09}$	$6.22 \cdot 10^{-05}$	0.21
enriched in adenoma			
Alistipes putredinis	$5.60 \cdot 10^{-25}$	$8.00 \cdot 10^{-27}$	-0.67
Bacteroides vulgatus	$0.00 \cdot 10^{+00}$	$8.99 \cdot 10^{-24}$	-0.57
Butyricicoccus faecihominis	$1.24 \cdot 10^{-13}$	$1.54 \cdot 10^{-14}$	-0.51
genus *[Ruminococcus] torques group*	$4.11 \cdot 10^{-25}$	$7.40 \cdot 10^{-26}$	-0.50
genus *Methanobrevibacter*	$4.91 \cdot 10^{-17}$	$2.00 \cdot 10^{-18}$	-0.48
class *Gammaproteobacteria*	$2.72 \cdot 10^{-12}$	$7.90 \cdot 10^{-17}$	-0.47
Akkermansia muciniphila	$1.60 \cdot 10^{-17}$	$1.07 \cdot 10^{-18}$	-0.47
family *Lachnospiraceae*	$3.95 \cdot 10^{-13}$	$2.53 \cdot 10^{-23}$	-0.47
order *Oscillospirales* family *UCG-011*	$1.40 \cdot 10^{-12}$	$5.64 \cdot 10^{-15}$	-0.44
family *Pasteurellaceae*	$3.25 \cdot 10^{-11}$	$9.16 \cdot 10^{-16}$	-0.43

in the gut correlated negatively with the intake of dietary fibre, fruits and vegetables, and the fecal butyrate concentration in the fecal microbiome of Crohn's disease patients [115]. However, there are reports of significantly lower abundance of *Escherichia-Shigella* in CRC patients compared with healthy controls [59]. Thus, it appears that this bacterium is not necessarily associated with a healthy microbiome composition, but discriminates between healthy and CRC gut microbial profiles.

A study using stool samples reported that *Haemophilus* is found in significantly higher proportion in colorectal cancer compared to controls [116]. A recent study of the intestinal mucosal microbiome found that high abundance of *Haemophilus* was associated with CRC recurrences and poorer rates of disease-free survival (DFS) or overall survival (OS) rates [117]. Therefore, the results of both ALDEx2 and ANCOM-BC in this dataset are not consistent with previous reports.

Although *Parabacteroides* has been shown to be both beneficial and pathogenic to human health, it is generally agreed that this genus, particularly *P. distasonis* plays a protective role in colorectal cancer [118]. Studies indicated an inverse correlation of *P. distasonis* levels with the presence of intestinal tumors, as well as anti-inflammatory and anti-tumor properties. A study examining the differences between sporadic colorectal adenomas and samples without lesions (controls) found that *P. distasonis* was detected only in histological samples from control subjects [119].

The genus *Collinsella* has been proposed as a fecal biomarker for early detection of CRC because it has been shown to be substantially elevated in CRC stage I compared with healthy controls [59].

Brevundimonas was reported to be reduced in CRC patients compared with healthy individuals [120] and also to have significantly lower relative abundance in tumor mucosa compared with the matched noncancerous mucosa in CRC patients [121]. Since exposure to aromatic hydrocarbons is known to be associated with CRC [122], *Brevundimonas* levels might have a positive influence, as this genus is able to degrade and detoxify aromatic compounds, thus reducing their toxic effect [123, 121].

The enrichment of *Lachnospiraceae* in the healthy group has already been discussed in CRC vs. healthy DAA (Section 3.3.1).

Phascolarctobacterium is a SCFA producer and has been reported to have beneficial effects on the host, including positive effects on mood in humans [124, 125].

Enriched in adenoma

It has been found that the levels of a member of the genus *Alistipes*, *A. finegoldii* are increased in an inflamed intestinal environment and contribute to the pathogenesis and formation of (right-sided) colorectal tumors [126, 127].

The role of the various species from the *Bacteroides* genus is complex. As discussed earlier, *B. thetaiotaomicron* is associated with beneficial effects on the gut. The impact of *B. vulgatus* is not as easy to describe, as the effect of this species appears to depend on the overall gut environment and also on the animal models used in the research. There are reports of *B. vulgatus* levels being elevated in patients with Chron's disease and triggering the expression of proinflammatory cytokines. This bacterium was also able to induce colitis and gastritis in transgenic rats, with similar results by other studies on humans. However, there are also reports showing alleviation of inflammation in mice, potential probiotic effects, and protection against induced colitis in mice. [128]
Wang et al. performed structural segregation of the gut microbiota between healthy volunteers and colorectal cancer patients and found that two OTUs closely related to *B. vulgatus* were enriched in the healthy controls [129].

A study examining biomarkers for early detection of CRC found that *Butyricoccus faecihominis* was more abundant in the gut microbiota of adenoma patients than in that of CRC patients (but not in that of adenoma patients compared with healthy individuals) [130]. This bacterium was isolated for the first time relatively recently (2016) from the stool of a healthy human and is known to function as a butyrate producer [131]. There is not much information about its association with the development of colorectal cancer.

Ruminococcus torques has been identified as one of the common species significantly associated with a high risk of colorectal cancer [132]. There are also reports of enrichment of circulating bacterial DNA of this taxon in stool samples from CRC patients [133].

Methanobrevibacter millerae was increased in adenoma samples compared with controls in amplicon-sequenced fecal samples [130]. Differential abundance analysis of CRC patients compared with healthy individuals revealed that genus *Methanobrevibacter* was enriched in CRC samples [134]. Many studies have reported elevation of methanogens in samples from patients with ulcerative colitis, intestinal polyps, and tumors. However, further research is needed to understand the mechanism by which they contribute to disease development [135].

The class *Gammaproteobacteria* was detected as enriched in adenoma stool samples compared with controls, especially the order *Enterobacteriales*, family *Enterobacteriaceae* [136]. *E. coli*, a bacterium belonging to this taxonomy, is suspected of promoting colorectal cancer, although it is generally considered a commensal bacterium. There are two mechanisms described by Allen et al. by which *E. coli* may contribute to the development of CRC. One is through the production of genotoxins such as colibactin, which can damage double-stranded DNA, leading to neoplastic transformation. Second is by causing inflammatory response through various pattern recognition receptors (PRRs), a process in which neoplastic progression is enhanced. [137]

Akkermansia muciniphila's inconclusive role in CRC was discussed in Section 3.3.1. According to the results based on this dataset, the high abundance of this species is associated with cancer/adenoma dysbiosis rather than a healthy colon.

The enrichment of *Lachnospiraceae* in the adenoma group was not expected because this taxon is usually associated with healthy gut flora (Section 3.3.1). However, this family was detected among the top 10 species for distinguishing adenoma patients from CRC patients using the random forest algorithm [61]. Nevertheless, there are no other reports of *Lachnospiraceae* enrichment in the adenoma group compared with a healthy control group.

Oscillospirales order is detected as enriched in the adenoma group, although this was also unexpected, just as its enrichment in the CRC group (Section 3.3.1).

There are no reports on the abundance of *Pasteurellaceae* in stool samples from CRC/adenoma patients. The only finding in the literature is that this taxon had a significantly higher relative abundance in cancerous tissue than in the intestinal lumen of CRC patients [63].

ANCOM-BC

ANCOM-BC with bias correction for using different study datasets detected only one differentially abundant species *Ruminococcus champanellensis* enriched in the healthy group (supplementary file S9). With the intention of identifying more taxa, the bias correction was omitted. This resulted in the discovery of 136 DA taxa, 48 of which were enriched in the healthy group and 88 in the adenoma group (supplementary files S7 and S8).

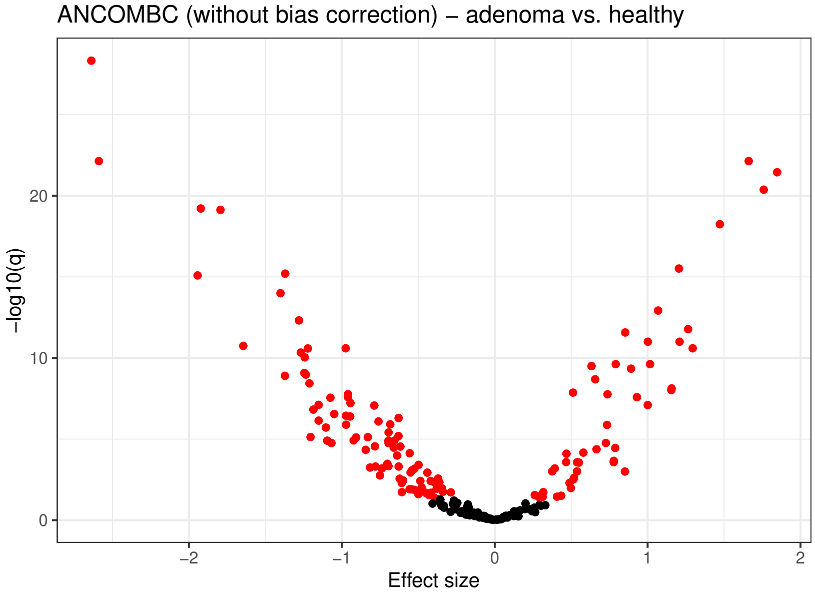


Figure 3.9: ANCOM-BC result plot for DAA between adenoma and healthy group. Red dots represent differentially abundant species with BH corrected p-value q < 0.05. Taxa with positive effect size are enriched in the healthy group, whereas those with negative effect size are enriched in the adenoma group.

Enriched in healthy

The genus *Dialister* was found to be enriched in healthy in adenoma vs. healthy comparison, whereas *Dialister pneumosintes* was identified as differentially enriched in CRC (Table 3.6). However, the literature indicates that this taxon is actually more abundant in healthy stool samples than in samples from CRC patients [91]. As mentioned in Section 3.3.1, *Megamonas* was previously found to be enriched in healthy samples compared to CRC samples [91].

Enriched in adenoma

Bacteroides uniformis can adapt to different intestinal environments and is considered a potential probiotic with multiple effects on host health [128]. It was also reported that an OTU related to *B. uniformis* was enriched in healthy controls compared to CRC patients [129]. Park et al. found that the relative abundance of the genus "*Family*

Table 3.10: Differentially abundant taxa with the highest effect-size values identified by ANCOM-BC in adenoma vs. healthy comparison. Top 10 taxa enriched in the healthy group and in the adenoma group, respectively. q- BH corrected p-value; lfc- log-fold change, measure for effect size.

enriched in healthy	q	lfc
Escherichia-Shigella	$3.54 \cdot 10^{-22}$	1.85
Bifidobacterium longum	$4.21 \cdot 10^{-21}$	1.76
Anaerostipes hadrus	$7.22 \cdot 10^{-23}$	1.66
Bacteroides thetaiotaomicron	$5.67 \cdot 10^{-19}$	1.47
genus *Parabacteroides*	$2.52 \cdot 10^{-11}$	1.30
genus *Collinsella*	$1.67 \cdot 10^{-12}$	1.27
genus *Phascolarctobacterium*	$9.95 \cdot 10^{-12}$	1.21
genus *Haemophilus*	$3.06 \cdot 10^{-16}$	1.21
genus *Dialister*	$7.49 \cdot 10^{-09}$	1.16
genus *Megamonas*	$9.50 \cdot 10^{-09}$	1.16
enriched in adenoma		
Alistipes putredinis	$4.70 \cdot 10^{-29}$	-2.64
Bacteroides vulgatus	$7.22 \cdot 10^{-23}$	-2.59
Akkermansia muciniphila	$8.13 \cdot 10^{-16}$	-1.94
genus *Methanobrevibacter*	$6.08 \cdot 10^{-20}$	-1.92
family *[Ruminococcus] torques group*	$7.41 \cdot 10^{-20}$	-1.79
Bacteroides uniformis	$1.80 \cdot 10^{-11}$	-1.65
genus *Family XIII AD3011 group*	$1.01 \cdot 10^{-14}$	-1.40
genus *Ruminococcus*	$1.26 \cdot 10^{-09}$	-1.37
Butyricicoccus faecihominis	$6.32 \cdot 10^{-16}$	-1.37
family *Pasteurellaceae*	$4.82 \cdot 10^{-13}$	-1.28

XIII AD3011 group" was lower in the CRC group than in the control group based on amplicon sequencing of stool samples [138]. There are no reports of the abundance of these two taxa in comparison between adenoma patients and healthy individuals, but in the analysis of CRC, the results of previous studies do not appear to be consistent with the ANCOM-BC results here.

Ruminococcus genus was reported as enriched in colorectal cancer patients compared to control subjects [59, 139]. The abundance of this genus was particularly elevated in patients with stage I of CRC, suggesting that *Ruminococcus* may be a biomarker for early detection of CRC [139], and the results of this dataset confirm this suggestion. The only species identified by ANCOM-BC with bias correction, *Ruminococcus champanellensis*, belongs to this genus. However, according to those results, it is enriched in the healthy group. There are no reports of its association with CRC. Although it is already known that *Ruminococci* serve as degraders of complex polysaccharides in the intestine [140], this particular species is even able to degrade insoluble dietary fibre, which makes it unique [141].

MaAsLin2

MaAsLin2 with bias correction for the study datasets identified only one DA taxon enriched in the adenoma group. This tool was not explicitly used during data exploration in R, as were the other two DAA methods, but was used as part of the machine learning pipeline for feature selection. Therefore, the settings for this method were kept constant for the CRC vs. healthy and adenoma vs. healthy classification. Omitting the bias correction as in ANCOM-BC would likely lead to more DA taxa and more comparable results with the other two methods.

Table 3.11: Differentially abundant taxon identified by MaAsLin2 in adenoma vs. healthy comparison, enriched in the adenoma group. qval- BH corrected p-value; coef- coefficient of the linear model, measure for effect size.

enriched in adenoma	qval	coef
family *Erysipelatoclostridiaceae*	$2.00 \cdot 10^{-03}$	-1.80

The family *Erysipelatoclostridiaceae* is a newly discovered taxon (in 2019) [142]. There are no reports of it being associated with colorectal cancer or adenoma.

3.3.3 Results comparison of differential abundance analysis methods

As can be seen in the Venn diagram of differentially abundant taxa for the CRC vs. healthy analysis, the three methods produce relatively consistent results (Figure 3.10). MaAsLin2 yields the most identified taxa, of which 22 taxa are found only with this method. The largest overlap is between MaAsLin2 and ANCOM-BC with 80 shared identified organisms. ANCOM-BC completely overlaps with either one of the other two methods and there are no taxa identified exclusively by this tool. 39 taxa are detected by all three methods. The list of these bacteria can be found in supplementary file S13.

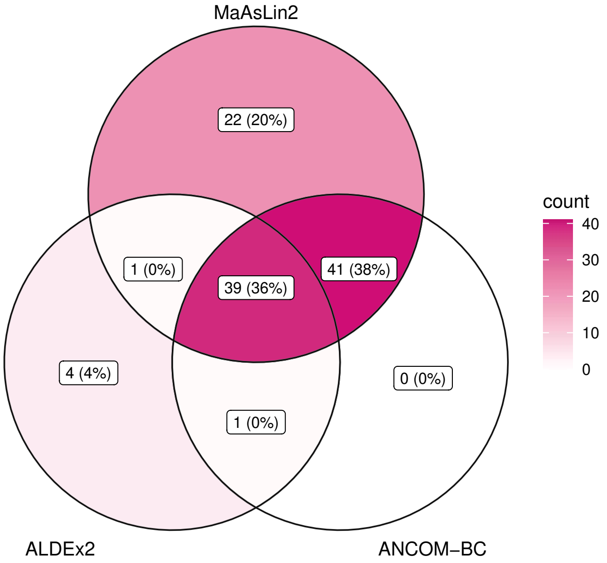


Figure 3.10: Venn diagram of the differentially abundant taxa found between healthy and CRC group by three different methods.

MaAsLin2 yielded only one identified taxon for adenoma vs. healthy DAA, which is not recognised by the other two methods (Figure 3.11). However, as mentioned earlier, this was the only method that applied a bias correction for adenoma vs. healthy analysis, as ALDEx2 does not have this option and it was omitted in ANCOM-BC to obtain more DA organisms. This resulted in a rather high number of identified taxa by ANCOM-BC, but still comparable to the number (and identity) of taxa detected by ALDEx2. 78% of all taxa are shared between the two methods and are listed in S14 (MaAsLin2 was not taken into account here).

Most of the identified taxa in DAA for both comparisons are consistent with findings in the literature. Some bacteria were not expected to be enriched in a particular group based on previous results from other studies. This is not so surprising because not all studies cited here used stool samples and amplicon sequencing to assess microbiome composition. Many of them analyzed biopsy samples, which have a significantly different composition than stool samples [143, 144]. The site from which the biopsy sample was

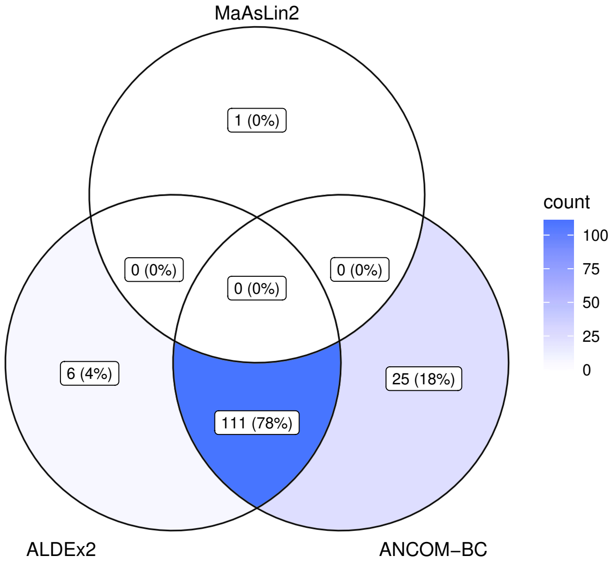


Figure 3.11: Venn diagram of the differentially abundant taxa found between healthy and adenoma group by three different methods.

taken (cancerous/noncancerous tissue, intestinal region) is also of great importance [63]. Therefore, it is not always meaningful to compare these results with fecal sequencing results. In addition, many studies have used animal models, which is not always directly translatable to the condition in the human body. However, the field of microbiome composition and function and its use for the diagnosis of (colorectal) cancer is still in its infancy. Due to the lack of high quality studies using fecal amplicon sequencing, studies with different methodologies were included in the discussion.

The bacteria identified here, especially the consensus organisms of the three methods (S13 and S14), are quite reliable biomarkers for colorectal cancer/adenoma dysbiosis - or at least the best possible with the available methods and metadata information. The dataset used for analysis here is robust and the sample size is large enough for each group. Another argument for its robustness is a diverse origin of the samples (European, American, and Asian). However, this dataset does not represent all races equally, as Whites and Asians are most commonly represented here. The results may therefore be missing some unique microbiome signatures of other races. The problem was also the missing metadata, as correcting for variables (other than study) such as cancer stage, age, race, diet type etc. would yield more reliable results. Nonetheless, bias correction was a problem in the adenoma vs. healthy DAA with only one confounding variable (study), since a single taxon was identified by ANCOM-BC and MaAsLin2, respectively. It is generally possible that there are few, or even none differentially abundant organisms between groups. However, in the case of colorectal adenoma, this is highly unlikely, as the literature results indicate a microbiome dysbiosis in adenoma patients. The beta diversity results on this dataset also indicate a significantly different taxonomic composition. For this reason, it was justified to omit the bias correction to allow the algorithm to identify more species (which was done in ANCOM-BC). However, it should also have been omitted in MaAsLin2 to ensure consistency in the methodology.

The researchers emphasise that no single OTU is increased in all individuals with CRC and that microbial community structure is more informative for the gut dysbiosis that occurs in colorectal cancer development than are abundance differences of individual taxa [144, 145]. This is because of the heterogeneity of CRC occurrence- not all individuals have the same type of CRC dysbiosis. Flemer et al. found that the most common taxa associated with CRC (*Fusobacterium, Peptostreptococcus, Parvimonas*) were significantly enriched in only 20-30% of CRC patients. However, they succeeded in defining four microbial clusters of the CRC-associated microbiota, at least one of which was more than twofold increased (compared to the mean in all control samples) in all but one of the individuals with CRC (a study with 70 CRC patients and 56 healthy controls) [144]. Identifying microbial clusters in larger data sets and examining the composition of each cluster would be a wise next step in future research in this area.

3.4 Machine learning models

Nine different algorithms, three feature selection options, two data transformation methods, and two taxonomy levels were explored in detail during automated runs of the machine learning pipeline, where all parameter combinations were tested (Table 2.3). Two separate classifications were modelled: CRC vs. healthy and adenoma vs. healthy. The best performing models (based on area under the curve (AUC)) per algorithm are presented in the following subsections for both classifications. Results are based on 80% of the meta-analysis dataset used for training. Output metrics (AUC, sensitivity and specificity) are the mean of metrics resulting from a 5-fold cross-validation with 5 repeats (25 values). The best performing models were used to predict the classes of samples from a test dataset, consisting of the remaining 20% of the meta-analysis dataset.

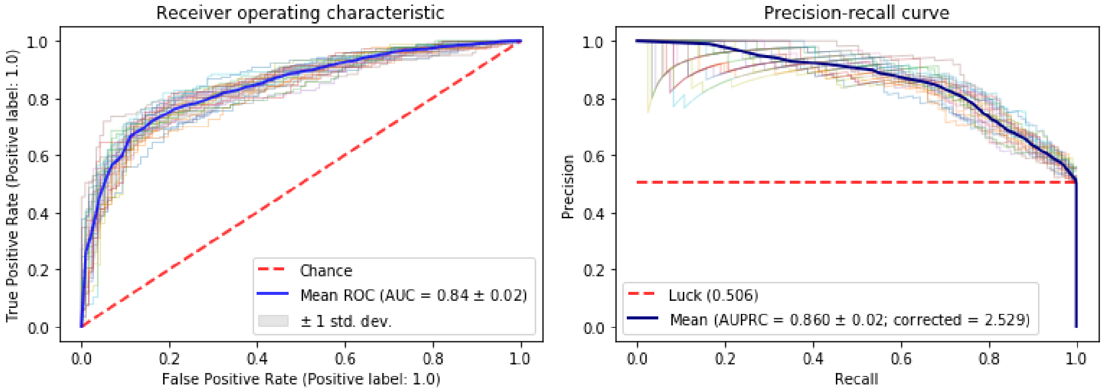
3.4.1 Colorectal cancer-healthy classification

Table 3.12: CRC vs. healthy classification: the best performing model (based on AUC) for each algorithm. FS- feature selection, AUC- area under the curve, SKB- SelectKBest; ML algorithm abbreviations can be found in Table 2.3. Data transformation for the best performing models was always compositional (centered log ratio).

Model	Taxonomy	FS	AUC	Sensitivity	Specificity	Test AUC
LR	genus	SKB	0.826	0.718	0.770	0.750
LASSO	genus	SKB	0.826	0.719	0.772	0.753
RIDGE	genus	SKB	0.823	0.775	0.715	0.757
EN	genus	SKB	0.825	0.718	0.774	0.754
SVM linear	genus	SKB	0.824	0.711	0.779	0.764
SVM	genus	Maaslin2	0.843	0.724	0.824	0.794
RF	genus	SKB	0.834	0.712	0.789	0.800
LGBM	species	None	0.829	0.707	0.791	0.787
GNB	species	SKB	0.803	0.585	0.863	0.746

Overall, we can conclude that the genus taxonomy level and the SelectKBest algorithm for feature selection seem to work best for most of the algorithms. The genus level is probably the better option because many taxa do not contain species-level information and because this also reduces the dimension of the dataset (from 1488 species to 940 genera). Dimensionality reduction is an important step before machine learning because models that contain many variables tend to overfit to the training data at hand, resulting in high evaluation metrics (AUC) for the training dataset, but perform poorly on unknown data. This is a common problem with sequencing datasets because of the large number of identified taxa. It is therefore important to preprocess the data and select the features (taxa) that are most important for classification. The SelectKBest algorithm with the mutual information scoring function finds k features that have the highest mutual information with the target variable. In other words, these k variables contain the most information needed to obtain the target variable (i.e., the class membership of a sample). The number k was set constant to 50. This is another hyperparameter that should be varied in future optimization of these models to check which values give optimal results.

Data for all models presented in Table 3.12 were preprocessed using compositional data transformation, as this method gave better results than subsampling. Using CRC-healthy classification on a genus level as an example, the mean AUC of the best-performing models was 0.825 ± 0.011 for compositional transformation and 0.740 ± 0.051 for subsampling (with a subsampling depth set to 5000). In addition to the overall higher AUC mean, all individual algorithms performed better with the compositional data transformation in this example. For this reason, and because of the explanation of the advantages of using the compositional data transformation with count data in Section 2.4, the centered log-ratio transformation was the preferred data transformation method, and subsampling was omitted from AutoML for adenoma-healthy classification to reduce the number of calculated experiments and thus the computation time.

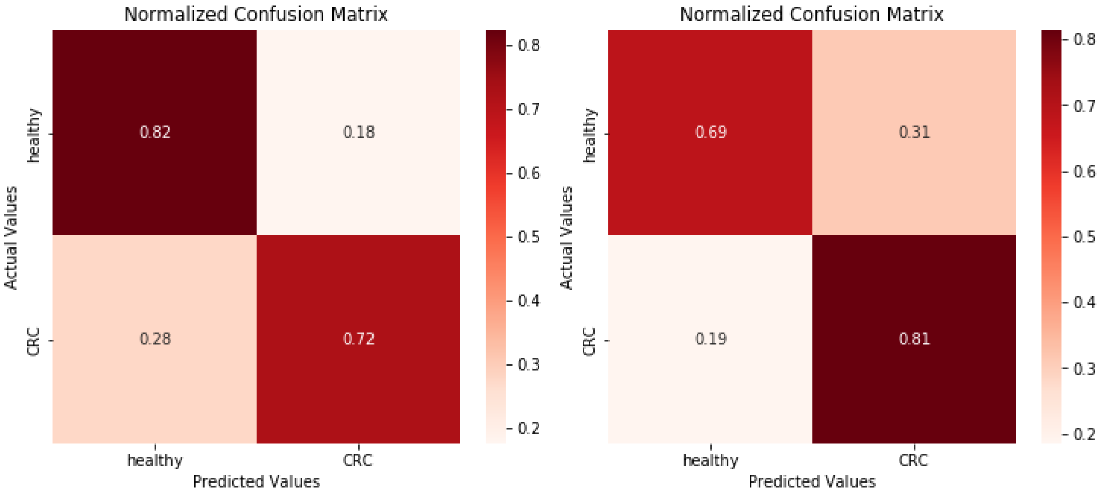


(a) ROC AUC

(b) PRC

Figure 3.12: Area under the receiver operating characteristic curve (ROC AUC) (a) and precision-recall curve (PRC) (b) for the CRC-healthy classification (SVM model on a genus level with Maaslin2 feature selection). The individual 25 curves resulting from each cross-validation step and the overall mean are shown in the graphs.

The support vector machine algorithm with radial kernel at genus level and with Maaslin2 feature selection yielded the highest AUC (0.843). Looking at the AUC values in Table 3.12, we can see that the AUC is high for all algorithms (>0.8). This is a good sign, because AUC is a metric for the overall ability of the model to classify the observations into their respective groups. The algorithms used are quite simple (some of them are even linear), which makes them easy to interpret, which is an important attribute of a potential diagnostic product in medicine. More specific metrics are sensitivity and specificity. Sensitivity denotes the proportion of correctly classified observations from class 1 (here CRC, the true positives), while specificity denotes the proportion of true negatives (here healthy). The plot of the area under the receiver operating characteristic curve (AUC ROC) shows the relationship between the true positive rate (sensitivity) and the false positive rate (1-specificity). In a diagnostic model such as this, sensitivity is of greater importance because we do not want the true positives (people with colorectal cancer) to go under the radar. We would rather have a higher proportion of false positives because we can always perform colonoscopy on healthy individuals and confirm that they are indeed healthy, even though the ML model has classified them as having CRC. This is the reason why a confusion matrix with an adjusted cutoff was implemented in the ML pipeline. We see that setting the cutoff to 0.4 results in a higher proportion of true positives (higher sensitivity), but at the cost of lower specificity. The cutoff adjusted for the training dataset can be used to predict unknown samples. This was done for the test dataset. The values for sensitivity and specificity in Table 3.12 are the "original" values obtained when the cutoff is defined by the function `predict()`. The values in the confusion matrix in Figure 3.13b are the result of the probabilities calculated by `predict_proba()` and classified using an adjusted cutoff value (defined manually, here 0.4). This means that all samples with probabilities > 0.4 were classified as CRC and ≤ 0.4 as healthy.



(a) CM with default cutoff

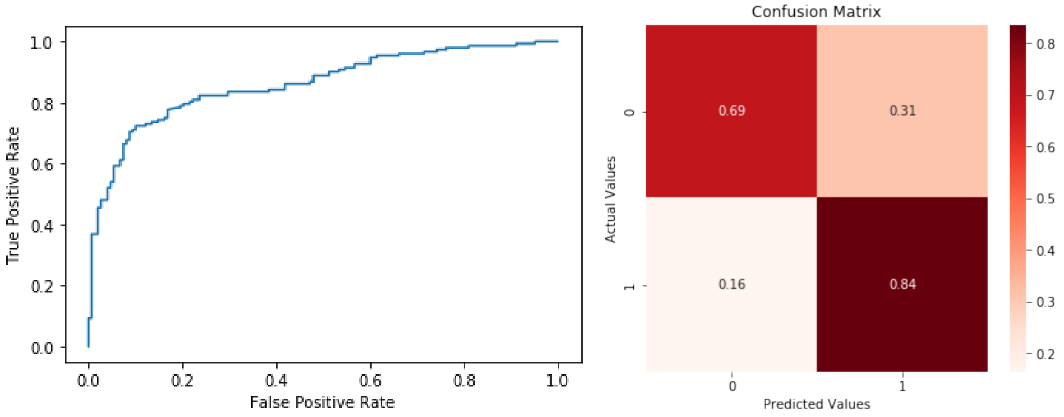
(b) CM with adjusted cutoff (0.4)

Figure 3.13: Confusion matrix (CM) for CRC-healthy classification: with default cutoff determined by the `predict()` function (a), and with manually adjusted cutoff (set to 0.4) to increase the sensitivity. The non-normalized CM with mean and standard deviation can be found in the appendix (Figure A.3a).

The results of the performance of the most successful model on the test dataset show that a higher sensitivity was indeed achieved with an adjusted cutoff even for the unknown data (sensitivity in Table 3.13 versus in Figure 3.14b). Based on the AUC values for the test dataset for all algorithms (Table 3.12), we can see that the models performed similarly for the training and test data, indicating that they are unlikely to be overfitted. However, a more reliable assessment of performance would be to use an external test dataset. Here, a portion of the meta-analysis dataset was split off from the training data to later serve as the (internal) test dataset. This dataset contained samples from the four study datasets (in random proportions), making it similar in "quality" to the training data. External datasets with their own aspects, such as sample preparation, inclusion of a particular race/geographic region, etc., would be necessary to test performance in an unbiased manner.

Table 3.13: Evaluation metrics for the best performing model of the CRC-healthy classification (SVM on a genus level with Maaslin2 feature selection) on the test dataset. Sensitivity and specificity are based on the default cutoff.

AUC	Sensitivity	Specificity
0.794	0.750	0.838



(a) ROC curve

(b) CM with adjusted cutoff (0.4)

Figure 3.14: CRC-healthy classification of the test dataset with the best performing model: ROC curve (a) and the confusion matrix with an adjusted cutoff (set to 0.4) (b) for the test dataset.

Figure 3.15 shows the taxa with the highest contributions to the output of the best performing machine learning model. Since Maaslin2 was the feature selection method, we again see the taxa already identified by the DAA tools in Section 3.3.1: genera *Peptostreptococcus, Parvimonas, Porphyromonas, Fusobacterium, Dialister, Ruminococcaceae UBA1819* and family *Prevotellaceae*, which were enriched in the CRC samples, and family *Peptostreptococcaceae*, genera *Lachnospira, Erysipelotrichaceae UCG-003, Ruminococcus, Megamonas* and *Anaerostipes*, which were enriched in the healthy samples. Although there were 99 features selected by Maaslin2, the ones that contribute the most to the model are also the ones with the highest effect size determined by the DAA algorithms (Table 3.6, Table 3.7, Table 3.8).

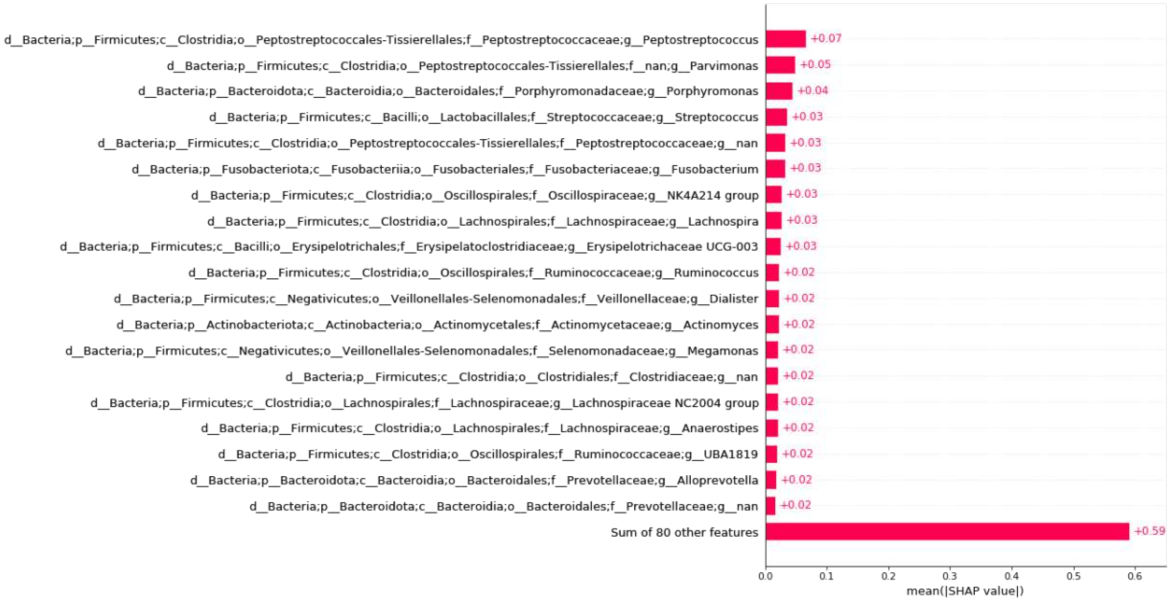
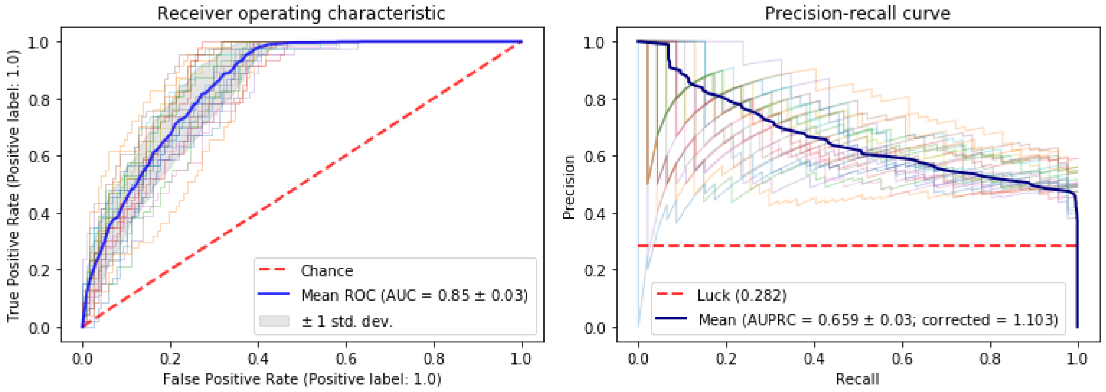


Figure 3.15: Interpretation of the best performing CRC-healthy classification model (SVM with Maaslin2 feature selection). 20 taxa with the highest contribution based on Shapley values calculated by iterating through permutations of the features are shown [146].

3.4.2 Adenoma-healthy classification

Table 3.14: Adenoma vs. healthy classification: the best performing model (based on AUC) for each algorithm. FS- feature selection, AUC- area under the curve, SKB- SelectKBest; ML algorithm abbreviations can be found in Table 2.3. Data transformation for the best performing models was always compositional (centered log ratio).

Model	Taxonomy	FS	AUC	Sensitivity	Specificity	Test AUC
LR	species	SKB	0.826	0.534	0.827	0.762
LASSO	species	SKB	0.837	0.534	0.838	0.753
RIDGE	species	SKB	0.824	0.535	0.823	0.748
EN	species	SKB	0.833	0.533	0.834	0.762
SVM linear	species	SKB	0.825	0.558	0.828	0.755
SVM	species	None	0.839	0.499	0.849	0.693
RF	species	SKB	0.838	0.508	0.847	0.710
LGBM	species	SKB	0.853	0.556	0.852	0.715
GNB	genus	SKB	0.814	0.934	0.607	0.774

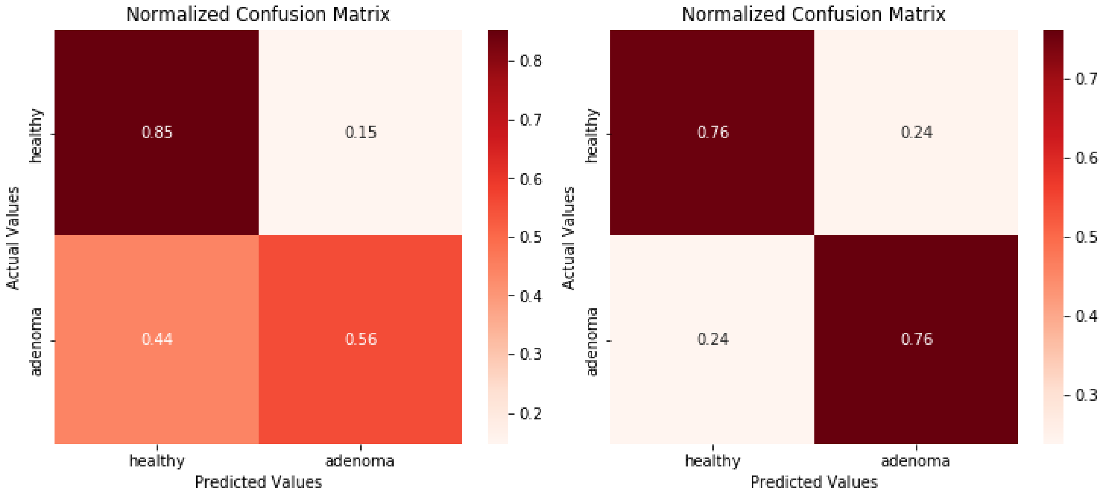


(a) ROC AUC

(b) PRC

Figure 3.16: Area under the receiver operating characteristic curve (ROC AUC) (a) and precision-recall curve (PRC) (b) for the adenoma-healthy classification (LGBM model on a species level with SKB feature selection). The individual 25 curves resulting from each cross-validation step and the overall mean are shown in the graphs.

The best performing model for adenoma-healthy classification is LGBM at the species level with SelectKBest feature selection (and compositional data transformation). The resulting AUC is even slightly higher than for CRC-healthy classification. While the sensitivity and specificity values were about the same for the CRC-healthy classification, here we have low sensitivity and higher specificity. The imbalance can be seen on the ROC curve in Figure 3.16a. The reason for the unbalanced results is a large imbalance of classes in the training dataset - 232 adenoma and 591 healthy samples (ratio $\approx 1 : 2.5$). To address this problem, class weights could be added when defining the classifier functions in future optimization of the models. The imbalance also resulted in the need for a rather extreme cutoff value to obtain a satisfactory ratio between sensitivity and specificity. Setting it to 0.2, we obtain approximately equal values (Figure 3.17b).



(a) CM with default cutoff

(b) CM with adjusted cutoff (0.2)

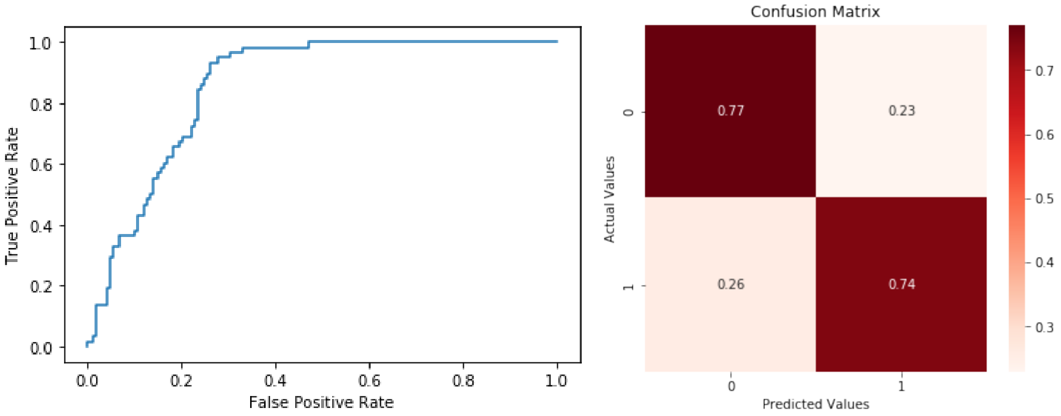
Figure 3.17: Confusion matrix (CM) for adenoma-healthy classification: with default cutoff determined by the `predict()` function (a), and with manually adjusted cutoff (set to 0.2) to increase the sensitivity. The non-normalized CM with mean and standard deviation can be found in the appendix (Figure A.3b).

It is interesting to look at the results using Maaslin2 feature selection. The q-value threshold for significance in the `Maaslin2()` function was not explicitly set, but was left at the default value of 0.25. This threshold was higher than that used for differential abundance analysis (Section 3.3), where the overall Maaslin2 results table was used and only the features with q< 0.05 were considered significant (because this threshold was taken for the other two DAA methods). With this more liberal q-value threshold, Maaslin2 yielded 4 DA taxa at the genus level (genus *Desulfovibrio*, family *Erysipelatoclostridiaceae* genus *Frisingicoccus*, and genus *Mogibacterium*) and 3 DA species-level taxa (*Alistipes obesi*, family *Erysipelatoclostridiaceae*, and genus *Mogibacterium*) that were used as features for the ML models. Although the number of variables was so small, the performance of the algorithms was surprisingly good: the average AUC of the 9 species-level models was 0.698 ± 0.020 (with the highest AUC of 0.718 for the GNB model) and 0.708 ± 0.026 at the genus level (with the highest AUC of 0.732 again for the GNB model). This shows that a reduction to fewer taxa can indeed lead to reliable models (of course, a reduction to only 3/4 taxa is a bit too drastic to obtain accurate models). Future optimization of the ML pipeline should therefore focus on exploring further methods for feature selection, including implementing the other two DAA tools used here and adjusting the cutoff threshold of the corrected p-value to obtain a desirable number of features.

As with the CRC-healthy classification, performance on the test dataset is satisfactory for all adenoma-healthy classifiers (Table 3.14). Table 3.15 and Figure 3.18b again demonstrate the improvement in sensitivity when using the cutoff value for classification determined on the training dataset (here 0.2).

Table 3.15: Evaluation metrics for the best performing model of the adenoma-healthy classification (LGBM on a species level with SKB feature selection) on the test dataset. Sensitivity and specificity are based on the default cutoff.

AUC	Sensitivity	Specificity
0.715	0.586	0.845



(a) ROC curve

(b) CM with adjusted cutoff (0.2)

Figure 3.18: Adenoma-healthy classification of the test dataset with the best performing model: ROC curve (a) and the confusion matrix with an adjusted cutoff (set to 0.2) (b) for the test dataset.

The beeswarm plot implemented for the tree-based models is even more informative than the permutation explainer (suitable for all model types, Figure 3.15). Here we can see not only the features that contribute most to the output of the model, but also how the feature value affects the model output. Figure 3.19 shows that a low value of the features *Anaerostipes hadrus, Escherichia Shigella, Bifidobacterium longum* (i.e. their low count values) and a high value of the features *Eubacterium halii, Gammaproteobacteria, Alistipes putredinis, Ruminococcus torques group, Pasteurellaceae, Methanobrevibacter, Akkermansia muciniphila* (i.e. their high count values) contribute to the classification of a sample to class 1 (adenoma). The bacteria from the first group of taxa (whose low abundance contributes to classification as adenoma) were found to be significantly enriched in the healthy group, and the bacteria from the second group of taxa (whose high abundance contributes to classification as adenoma) were found to be significantly enriched in the adenoma group in DAA (Section 3.3.2). The feature selection method used for this classification is SKB, and the results of the ML interpretation show that the features with the highest effect size values detected in DAA are not only selected by a different algorithm (SKB) but are also the most relevant to the ML classification. This highlights that the identified biomarkers in Section 3.3 are a reliable list of organisms indicative of the dysbiosis of the gut microbiome during the development of colorectal cancer and adenoma.

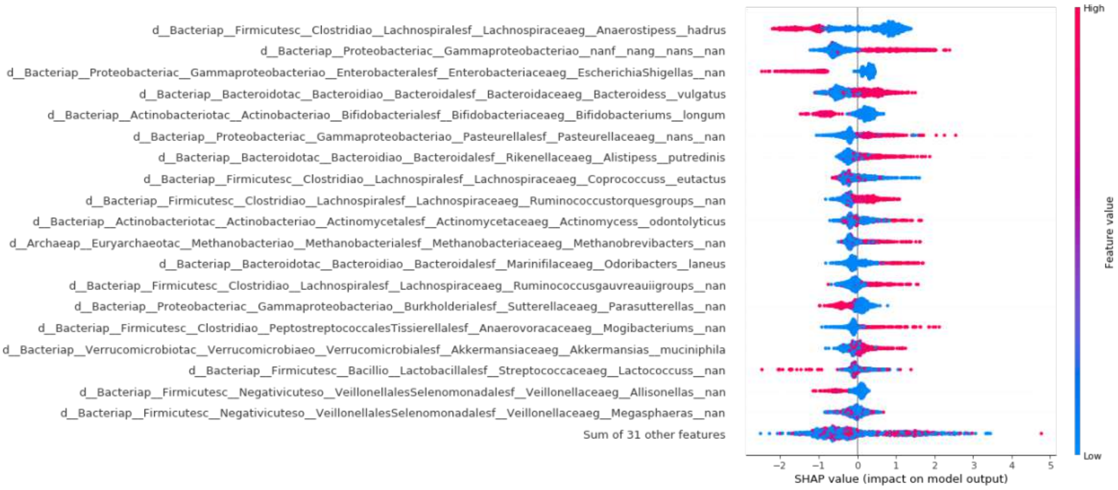


Figure 3.19: Interpretation of the best performing adenoma-healthy classification model (LGBM with SKB feature selection). 20 taxa with the highest contribution based on Shapley values and their effect on model output.

CHAPTER 4

Conclusion

Examination of the differences in alpha diversity between the healthy, adenoma, and colorectal cancer groups in the meta-analysis dataset did not yield meaningful outcomes. Combined with the inconsistent results from the literature, it was concluded that alpha diversity was not a relevant metric for comparing the microbiome composition of healthy controls and colorectal cancer/adenoma patients. The differences in beta diversity were statistically significant between the three diagnosis groups and the four study datasets. The significant difference in taxonomic composition between the diagnosis groups was further investigated using differential abundance analysis tools and machine learning models.

Differential abundance analysis (DAA) detected taxa enriched in healthy controls and in groups with a disease (colorectal cancer/adenoma). The organisms were identified using three different DAA methods, and the overlapping results of these tools represent a set of biomarkers indicative of gut microbiome dysbiosis in colorectal cancer and adenoma. Robustness was achieved by using multiple methods and a large and diverse dataset. Because all of the work here was based on publicly available data, the lack of metadata precluded more detailed analysis of the microbial signatures characteristic of these diseases. The only complete information available for all samples was the study dataset from which they originated. There are reports of unique taxa representative of different cancer stages, and it would thus be important and exciting to identify them in a large meta-analysis dataset such as this. This is not possible without the detailed metadata for the individual datasets. DA organisms have mostly been detected by other researchers as well, but not always based on fecal amplicon sequencing. Some bacteria have never been mentioned in the literature in the context of colorectal cancer or adenoma. The organisms detected here are therefore a good basis for further research, which should perhaps focus on understanding the heterogeneity of CRC occurrence and trying to find microbial clusters, rather than single organisms, indicative of different CRC gut dysbiosis types.

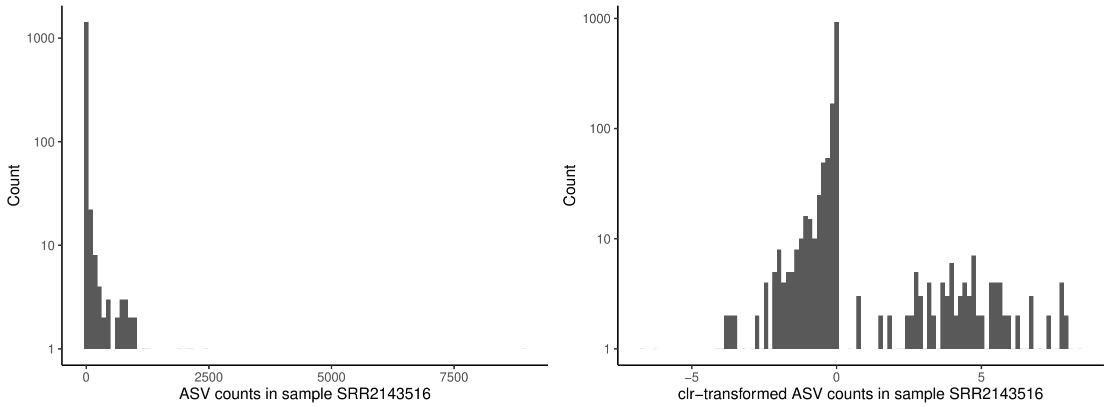
The machine learning models were very successful in classifying the samples for both cases (CRC vs. healthy and adenoma vs. healthy). For the internal test data set, the evaluation metrics were slightly lower than for cross-validation, but still good. Performance is comparable to that of the commonly used non-invasive tool, the fecal immunochemical test (FIT), for detecting CRC and superior to this method for detecting adenomas. Future work should include evaluation of the models using external test datasets. As the gut microbiome and its use for developing diagnostic tools has become a popular field in recent years, hopefully new datasets with more detailed metadata will become available, which will certainly make the models more accurate. With adequate metadata, it would also be possible to develop a tool that not only screens for cancer but can also reliably identify the cancer stage. The algorithms tested here are quite simple, but still managed to perform very well. Further optimization should involve hyperparameter tuning, where the different values of the ML classifier function parameters (regularisation strength for linear models (LR, LASSO, RIDGE, EN), maximum depth of the tree for tree-based methods (RF, LGBM), kernel coefficient for SVM...) should be tested during the automated runs to find optimal conditions under which the models could perform even better. With more high-quality data (and metadata) and detailed model optimization strategies, performance can be improved. Therefore, there is probably no need to use more complex approaches (e.g. deep learning), as these may produce less explainable solutions with a higher risk of overfitting. In addition to hyperparameter tuning, further work should focus on feature selection in order to reduce the dimension to the relevant features and obtain reliable models with as few variables as possible. The results of the model interpretation were consistent with the DAA results, as the features that contributed most to the models were also those with the highest absolute effect size in the DAA. Thus, the DAA tools used here would be good candidates for feature selection methods, and the (corrected) p-value significance threshold and/or the absolute effect size value could serve as adjustable hyperparameters. It is also worth noting that depending on the results of future research, identified biomarkers could also serve as a basis for the development of cheaper and simpler non-invasive tests, e.g., based on real-time polymerase chain reaction.

To summarize, the most important outcomes of the work presented in this thesis are a list of fecal biomarkers of colorectal cancer and adenoma dysbiosis as a result of a robust statistical analysis, and the classification models developed for screening these diseases. The enhancement of these models should be continued, as they could not only serve as a screening tool in the future, but their interpretation also offers insight into which organisms play a significant role during CRC/adenoma gut dysbiosis. The advantages of such a screening method are that it is completely non-invasive, as only a stool sample is needed to formulate a diagnosis (classification), it is also independent of the practitioner's experience, and it can offer disease screening results promptly and relatively accurately.

Ultimately, the results presented here provide valuable information for understanding the imbalance of the gut microbiome in colorectal carcinoma and adenoma patients. Recent advances in sequencing technology combined with artificial intelligence have made this possible. With further optimization, these types of models, or simpler alternatives based on their results, could serve as a non-invasive addition to colonoscopy, the gold standard in the diagnosis of colorectal cancer and polyps.

APPENDIX A

Appendix



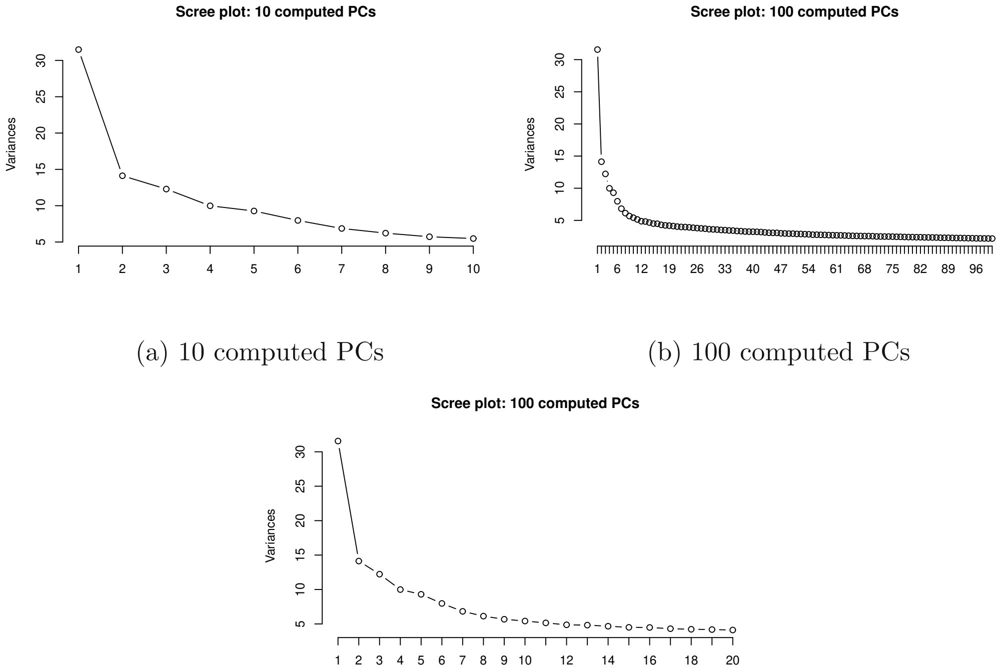
(a) Raw ASV counts

(b) clr-transformed counts

Figure A.1: Comparison of the ASV counts distribution between raw data (a) and centered log-ratio (clr) transformed data (b) for the first sample of the meta-analysis dataset. Raw data is strongly left-skewed, whereas the distribution becomes more centered after the clr transformation.

Table A.1: Cumulative proportion of explained variance for different numbers of calculated principal components (PCs).

Number of PCs	Cumulative explained variance (%)
10	12.6
30	22.2
100	43.6
200	63.3

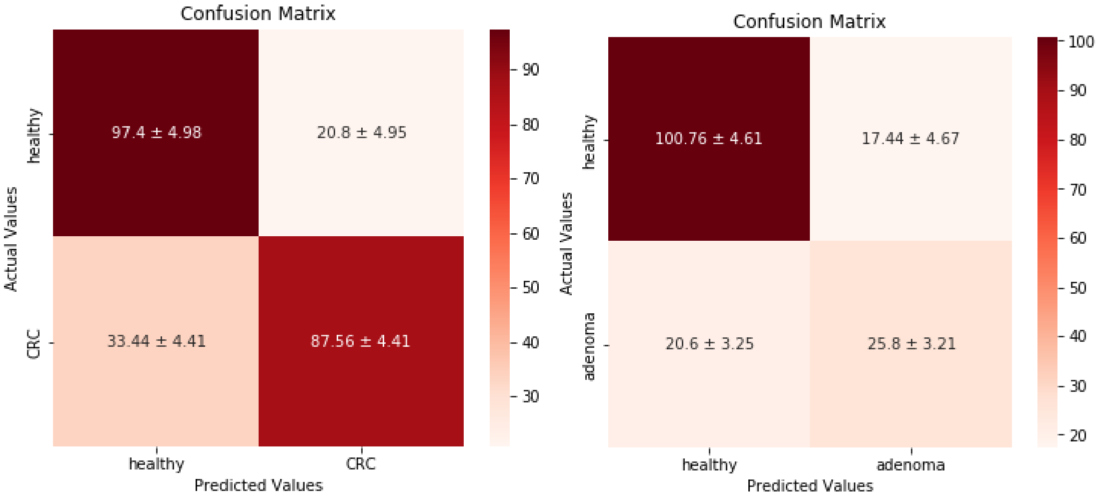


(a) 10 computed PCs

(b) 100 computed PCs

(c) 100 computed PCs, first 20 PCs shown

Figure A.2: Scree plots of computed 10 principal components (PCs) (a), 100 PCs (b) and 100 PCs with 20 PCs shown on the plot (c). Variance on the y-axis represents the absolute variance.



(a) CRC-healthy classification

(b) adenoma-healthy classification

Figure A.3: The original (non-normalized and with default cutoff) confusion matrix with mean and standard deviation for CRC-healthy classification (a) and adenoma-healthy classification (b).

List of Figures

List of Tables

Bibliography

[1] J. C. Setubal and E. Dias-Neto, “Microbiomes,” in *Reference Module in Life Sciences*, Elsevier, 2022.

[2] A. Heintz-Buschart and P. Wilmes, “Human gut microbiome: Function matters,” *Trends in Microbiology*, vol. 26, no. 7, pp. 563–574, 2018.

[3] S. V. Lynch and O. Pedersen, “The human intestinal microbiome in health and disease,” *New England Journal of Medicine*, vol. 375, no. 24, pp. 2369–2379, 2016.

[4] A. B. Shreiner, J. Y. Kao, and V. B. Young, “The gut microbiome in health and in disease,” *Current opinion in gastroenterology*, vol. 31, no. 1, p. 69, 2015.

[5] V. D’Argenio and F. Salvatore, “The role of the gut microbiome in the healthy adult status,” *Clinica chimica acta*, vol. 451, pp. 97–102, 2015.

[6] L. Zhang, F. Chen, Z. Zeng, M. Xu, F. Sun, L. Yang, X. Bi, Y. Lin, Y. Gao, H. Hao, *et al.*, “Advances in metagenomics and its application in environmental microorganisms,” *Frontiers in Microbiology*, p. 3847, 2021.

[7] F. Sambo, F. Finotello, E. Lavezzo, G. Baruzzo, G. Masi, E. Peta, M. Falda, S. Toppo, L. Barzon, and B. Di Camillo, “Optimizing pcr primers targeting the bacterial 16s ribosomal rna gene,” *BMC bioinformatics*, vol. 19, pp. 1–10, 2018.

[8] “Introduction to SBS Technology.” `https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html`. Accessed: 2023-05-12.

[9] B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes, “Dada2: High-resolution sample inference from illumina amplicon data,” *Nature methods*, vol. 13, no. 7, pp. 581–583, 2016.

[10] “Machine learning.” `https://en.wikipedia.org/wiki/Machine_learning`. Accessed: 2023-05-15.

[11] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.

[12] "Linear models - scikit learn." https://scikit-learn.org/stable/modules/linear_model.html. Accessed: 2023-05-14.

[13] "Naive Bayes - scikit learn." https://scikit-learn.org/stable/modules/naive_bayes.html. Accessed: 2023-05-14.

[14] "Bayes' theorem - Wikipedia." https://en.wikipedia.org/wiki/Bayes%27_theorem. Accessed: 2023-05-14.

[15] "LightGBM's documentation." https://lightgbm.readthedocs.io/en/v3.3.2/. Accessed: 2023-05-15.

[16] "CANCER FACT SHEETS." https://gco.iarc.fr/today/fact-sheets-cancers. Accessed: 2023-02-26.

[17] A. Leslie, F. Carey, N. Pratt, and R. Steele, "The colorectal adenoma–carcinoma sequence," *British Journal of Surgery*, vol. 89, no. 7, pp. 845–860, 2002.

[18] "Colon Adenoma." https://www.mercy.com/health-care-services/cancer-care-oncology/specialties/colorectal-cancer-treatment/conditions/colon-adenoma. Accessed: 2023-05-15.

[19] T. J. Eide, "Natural history of adenomas," *World journal of surgery*, vol. 15, pp. 3–6, 1991.

[20] T. Irrazábal, A. Belcheva, S. E. Girardin, A. Martin, and D. J. Philpott, "The multifaceted role of the intestinal microbiota in colon cancer," *Molecular cell*, vol. 54, no. 2, pp. 309–320, 2014.

[21] F. Stracci, M. Zorzi, and G. Grazzini, "Colorectal cancer screening: tests, strategies, and perspectives," *Frontiers in public health*, vol. 2, p. 210, 2014.

[22] "gFOBT." https://www.cancer.gov/publications/dictionaries/cancer-terms/def/gfobt. Accessed: 2023-05-15.

[23] "Fecal immunochemical test (FIT)." https://medlineplus.gov/ency/patientinstructions/000704.htm. Accessed: 2023-05-15.

[24] H. Brenner, L. Jansen, A. Ulrich, J. Chang-Claude, and M. Hoffmeister, "Survival of patients with symptom-and screening-detected colorectal cancer," *Oncotarget*, vol. 7, no. 28, p. 44695, 2016.

[25] J. Martín-López, C. Beltrán-Calvo, R. Rodríguez-López, and T. Molina-López, "Comparison of the accuracy of ct colonography and colonoscopy in the diagnosis of colorectal cancer," *Colorectal Disease*, vol. 16, no. 3, pp. O82–O89, 2014.

[26] C. J. Kahi, J. C. Anderson, and D. K. Rex, “Screening and surveillance for colorectal cancer: state of the art,” *Gastrointestinal endoscopy*, vol. 77, no. 3, pp. 335–350, 2013.

[27] B. Bressler, L. F. Paszat, Z. Chen, D. M. Rothwell, C. Vinden, and L. Rabeneck, “Rates of new or missed colorectal cancers after colonoscopy and their risk factors: A population-based analysis,” *Gastroenterology*, vol. 132, no. 1, pp. 96–102, 2007.

[28] S. H. Elsafi, N. I. Alqahtani, N. Y. Zakary, and E. M. Al Zahrani, “The sensitivity, specificity, predictive values, and likelihood ratios of fecal occult blood test for the detection of colorectal cancer in hospital settings,” *Clinical and experimental gastroenterology*, pp. 279–284, 2015.

[29] S. Zou, L. Fang, and M.-H. Lee, “Dysbiosis of gut microbiota in promoting the development of colorectal cancer,” *Gastroenterology Report*, vol. 6, pp. 1–12, 10 2017.

[30] K. J. Stott, B. Phillips, L. Parry, and S. May, “Recent advancements in the exploitation of the gut microbiome in the diagnosis and treatment of colorectal cancer,” *Bioscience Reports*, vol. 41, no. 7, 2021.

[31] S. H. Wong, T. N. Kwong, T.-C. Chow, A. K. Luk, R. Z. Dai, G. Nakatsu, T. Y. Lam, L. Zhang, J. C. Wu, F. K. Chan, *et al.*, “Quantitation of faecal fusobacterium improves faecal immunochemical test in detecting advanced colorectal neoplasia,” *Gut*, vol. 66, no. 8, pp. 1441–1448, 2017.

[32] S. Guo, L. Li, B. Xu, M. Li, Q. Zeng, H. Xiao, Y. Xue, Y. Wu, Y. Wang, W. Liu, *et al.*, “A simple and novel fecal biomarker for colorectal cancer: ratio of fusobacterium nucleatum to probiotics populations, based on their antagonistic effect,” *Clinical chemistry*, vol. 64, no. 9, pp. 1327–1337, 2018.

[33] M. Zhang, Y. Lv, S. Hou, Y. Liu, Y. Wang, and X. Wan, “Differential mucosal microbiome profiles across stages of human colorectal cancer,” *Life*, vol. 11, no. 8, p. 831, 2021.

[34] Y. Yang, L. Du, D. Shi, C. Kong, J. Liu, G. Liu, X. Li, and Y. Ma, “Dysbiosis of human gut microbiome in young-onset colorectal cancer,” *Nature communications*, vol. 12, no. 1, pp. 1–13, 2021.

[35] X. Fan, Y. Jin, G. Chen, X. Ma, and L. Zhang, “Gut microbiota dysbiosis drives the development of colorectal cancer,” *Digestion*, vol. 102, no. 4, pp. 508–515, 2021.

[36] R. L. Siegel, K. D. Miller, S. A. Fedewa, D. J. Ahnen, R. G. Meester, A. Barzi, and A. Jemal, “Colorectal cancer statistics, 2017,” *CA: a cancer journal for clinicians*, vol. 67, no. 3, pp. 177–193, 2017.

[37] N. T. Baxter, M. T. Ruffin, M. A. Rogers, and P. D. Schloss, "Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions," *Genome medicine*, vol. 8, no. 1, pp. 1–10, 2016.

[38] J. P. Zackular, M. A. Rogers, M. T. Ruffin IV, and P. D. Schloss, "The human gut microbiome as a screening tool for colorectal cancer," *Cancer prevention research*, vol. 7, no. 11, pp. 1112–1121, 2014.

[39] G. Zeller, J. Tap, A. Y. Voigt, S. Sunagawa, J. R. Kultima, P. I. Costea, A. Amiot, J. Böhm, F. Brunetti, N. Habermann, *et al.*, "Potential of fecal microbiota for early-stage detection of colorectal cancer," *Molecular systems biology*, vol. 10, no. 11, p. 766, 2014.

[40] "Microbiome CRC Biomarker Study." `http://mothur.org/MicrobiomeBiomarkerCRC/`. Accessed: 2022-11-17.

[41] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet. journal*, vol. 17, no. 1, pp. 10–12, 2011.

[42] "FIGARO - GitHub." `https://github.com/Zymo-Research/figaro`. Accessed: 2022-11-18.

[43] "IdTaxa: Assign Sequences a Taxonomic Classification." `https://rdrr.io/bioc/DECIPHER/man/IdTaxa.html`. Accessed: 2022-11-18.

[44] "addSpecies: Add species-level annotation to a taxonomic table.." `https://rdrr.io/bioc/dada2/man/addSpecies.html`. Accessed: 2022-11-18.

[45] "Analysis of colorectal cancer and adenoma microbiome signatures and the application of machine learning classification as a potential screening tool - Diploma Thesis GitHub repository." `https://github.com/kejt312/diplomathesis`. Accessed: 2023-03-10.

[46] "PcaHubert: ROBPCA - ROBust method for Principal Components Analysis." `https://www.rdocumentation.org/packages/rrcov/versions/1.7-1/topics/PcaHubert`. Accessed: 2022-10-29.

[47] T. P. Quinn, I. Erb, M. F. Richardson, and T. M. Crowley, "Understanding sequencing data as compositions: an outlook and review," *Bioinformatics*, vol. 34, no. 16, pp. 2870–2878, 2018.

[48] J. Aitchison, "The statistical analysis of compositional data," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 44, no. 2, pp. 139–160, 1982.

[49] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue, "Microbiome datasets are compositional: and this is not optional," *Frontiers in microbiology*, vol. 8, p. 2224, 2017.

[50] M. J. Anderson, “Permutational multivariate analysis of variance (permanova),” *Wiley statsref: statistics reference online*, pp. 1–15, 2014.

[51] L. Lahti, S. Shetty, and F. Ernst, “Orchestrating microbiome analysis,” 2021.

[52] J. T. Nearing, G. M. Douglas, M. G. Hayes, J. MacDonald, D. K. Desai, N. Allward, C. Jones, R. J. Wright, A. S. Dhanani, A. M. Comeau, *et al.*, “Microbiome differential abundance methods produce different results across 38 datasets,” *Nature communications*, vol. 13, no. 1, pp. 1–16, 2022.

[53] “Introduction to Katib.” `https://www.kubeflow.org/docs/components/katib/overview/`. Accessed: 2022-12-02.

[54] B. D. Topçuoğlu, N. A. Lesniak, M. T. Ruffin IV, J. Wiens, and P. D. Schloss, “A framework for effective application of machine learning to microbiome-based classification problems,” *MBio*, vol. 11, no. 3, pp. e00434–20, 2020.

[55] “Univariate feature selection.” `https://scikit-learn.org/stable/modules/feature_selection.html#univariate-feature-selection`. Accessed: 2022-12-02.

[56] T. Yatsunenko, F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R. N. Baldassano, A. P. Anokhin, *et al.*, “Human gut microbiome viewed across age and geography,” *nature*, vol. 486, no. 7402, pp. 222–227, 2012.

[57] V. D. Badal, E. D. Vaccariello, E. R. Murray, K. E. Yu, R. Knight, D. V. Jeste, and T. T. Nguyen, “The gut microbiome, aging, and longevity: a systematic review,” *Nutrients*, vol. 12, no. 12, p. 3759, 2020.

[58] J. Ahn, R. Sinha, Z. Pei, C. Dominianni, J. Wu, J. Shi, J. J. Goedert, R. B. Hayes, and L. Yang, “Human gut microbiome and risk for colorectal cancer,” *Journal of the National Cancer Institute*, vol. 105, no. 24, pp. 1907–1911, 2013.

[59] Q. Sheng, H. Du, X. Cheng, X. Cheng, Y. Tang, L. Pan, Q. Wang, and J. Lin, “Characteristics of fecal gut microbiota in patients with colorectal cancer at different stages and different sites,” *Oncology letters*, vol. 18, no. 5, pp. 4834–4844, 2019.

[60] D. Ai, H. Pan, X. Li, Y. Gao, G. Liu, and L. C. Xia, “Identifying gut microbiota associated with colorectal cancer using a zero-inflated lognormal model,” *Frontiers in microbiology*, vol. 10, p. 826, 2019.

[61] C. Lin, B. Li, C. Tu, X. Chen, and M. Guo, “Correlations between intestinal microbiota and clinical characteristics in colorectal adenoma/carcinoma,” *BioMed Research International*, vol. 2022, 2022.

[62] E. Russo, G. Bacci, C. Chiellini, C. Fagorzi, E. Niccolai, A. Taddei, F. Ricci, M. N. Ringressi, R. Borrelli, F. Melli, *et al.*, "Preliminary comparison of oral and intestinal human microbiota in patients with colorectal cancer: a pilot study," *Frontiers in microbiology*, vol. 8, p. 2699, 2018.

[63] W. Chen, F. Liu, Z. Ling, X. Tong, and C. Xiang, "Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer," *PloS one*, vol. 7, no. 6, p. e39743, 2012.

[64] N. Wu, X. Yang, R. Zhang, J. Li, X. Xiao, Y. Hu, Y. Chen, F. Yang, N. Lu, Z. Wang, *et al.*, "Dysbiosis signature of fecal microbiota in colorectal cancer patients," *Microbial ecology*, vol. 66, no. 2, pp. 462–470, 2013.

[65] C. A. Gaulke and T. J. Sharpton, "The influence of ethnicity and geography on human gut microbiome composition," *Nature medicine*, vol. 24, no. 10, pp. 1495–1496, 2018.

[66] L. Mancabelli, C. Milani, G. A. Lugli, F. Turroni, D. Cocconi, D. van Sinderen, and M. Ventura, "Identification of universal gut microbial biomarkers of common human intestinal diseases by meta-analysis," *FEMS microbiology ecology*, vol. 93, no. 12, p. fix153, 2017.

[67] A. Mukherjee, C. Lordan, R. P. Ross, and P. D. Cotter, "Gut microbes from the phylogenetically diverse genus eubacterium and their various contributions to gut health," *Gut Microbes*, vol. 12, no. 1, p. 1802866, 2020.

[68] M. Vacca, G. Celano, F. M. Calabrese, P. Portincasa, M. Gobbetti, and M. De Angelis, "The controversial role of human gut lachnospiraceae," *Microorganisms*, vol. 8, no. 4, p. 573, 2020.

[69] S. H. Duncan, A. Belenguer, G. Holtrop, A. M. Johnstone, H. J. Flint, and G. E. Lobley, "Reduced dietary intake of carbohydrates by obese subjects results in decreased concentrations of butyrate and butyrate-producing bacteria in feces," *Applied and environmental microbiology*, vol. 73, no. 4, pp. 1073–1078, 2007.

[70] I. J. Dikeocha, A. M. Al-Kabsi, H.-T. Chiu, and M. A. Alshawsh, "Faecalibacterium prausnitzii ameliorates colorectal tumorigenesis and suppresses proliferation of hct116 colorectal cancer cells," *Biomedicines*, vol. 10, no. 5, p. 1128, 2022.

[71] M. Loftus, S. A.-D. Hassouneh, and S. Yooseph, "Bacterial community structure alterations within the colorectal cancer gut microbiome," *BMC microbiology*, vol. 21, no. 1, pp. 1–18, 2021.

[72] M. Mohammadi, H. Mirzaei, and M. Motallebi, "The role of anaerobic bacteria in the development and prevention of colorectal cancer: A review study," *Anaerobe*, p. 102501, 2021.

[73] L. Zhao, X. Zhang, Y. Zhou, K. Fu, H. C.-H. Lau, T. W.-Y. Chun, A. H.-K. Cheung, O. O. Coker, H. Wei, W. K.-K. Wu, *et al.*, “Parvimonas micra promotes colorectal tumorigenesis and is associated with prognosis of colorectal cancer patients,” *Oncogene*, vol. 41, no. 36, pp. 4200–4210, 2022.

[74] J. Ren, H. Sui, F. Fang, Q. Li, and B. Li, “The application of apcmin/+ mouse model in colorectal tumor researches,” *Journal of Cancer Research and Clinical Oncology*, vol. 145, no. 5, pp. 1111–1122, 2019.

[75] C.-H. Sun, B.-B. Li, B. Wang, J. Zhao, X.-Y. Zhang, T.-T. Li, W.-B. Li, D. Tang, M.-J. Qiu, X.-C. Wang, *et al.*, “The role of fusobacterium nucleatum in colorectal cancer: from carcinogenesis to clinical management,” *Chronic diseases and translational medicine*, vol. 5, no. 03, pp. 178–187, 2019.

[76] Y. Suehiro, K. Sakai, M. Nishioka, S. Hashimoto, T. Takami, S. Higaki, Y. Shindo, S. Hazama, M. Oka, H. Nagano, *et al.*, “Highly sensitive stool dna testing of fusobacterium nucleatum as a marker for detection of colorectal tumours in a japanese population,” *Annals of clinical biochemistry*, vol. 54, no. 1, pp. 86–91, 2017.

[77] Y.-Y. Li, Q.-X. Ge, J. Cao, Y.-J. Zhou, Y.-L. Du, B. Shen, Y.-J. Y. Wan, and Y.-Q. Nie, “Association of fusobacterium nucleatum infection with colorectal cancer in chinese patients,” *World journal of gastroenterology*, vol. 22, no. 11, p. 3227, 2016.

[78] R. V. Purcell, M. Visnovska, P. J. Biggs, S. Schmeier, and F. A. Frizelle, “Distinct gut microbiome patterns associate with consensus molecular subtypes of colorectal cancer,” *Scientific reports*, vol. 7, no. 1, pp. 1–12, 2017.

[79] K. J. Flynn, N. T. Baxter, and P. D. Schloss, “Metabolic and community synergy of oral bacteria in colorectal cancer,” *Msphere*, vol. 1, no. 3, pp. e00102–16, 2016.

[80] W. Dai, C. Li, T. Li, J. Hu, and H. Zhang, “Super-taxon in human microbiome are identified to be associated with colorectal cancer,” *BMC bioinformatics*, vol. 23, no. 1, pp. 1–18, 2022.

[81] X. Long, C. C. Wong, L. Tong, E. S. Chu, C. Ho Szeto, M. Y. Go, O. O. Coker, A. W. Chan, F. K. Chan, J. J. Sung, *et al.*, “Peptostreptococcus anaerobius promotes colorectal carcinogenesis and modulates tumour immunity,” *Nature microbiology*, vol. 4, no. 12, pp. 2319–2330, 2019.

[82] H. Tsoi, E. S. Chu, X. Zhang, J. Sheng, G. Nakatsu, S. C. Ng, A. W. Chan, F. K. Chan, J. J. Sung, and J. Yu, “Peptostreptococcus anaerobius induces intracellular cholesterol biosynthesis in colon cells to induce proliferation and causes dysplasia in mice,” *Gastroenterology*, vol. 152, no. 6, pp. 1419–1433, 2017.

[83] N. Kelley, D. Jeltema, Y. Duan, and Y. He, “The nlrp3 inflammasome: an overview of mechanisms of activation and regulation,” *International journal of molecular sciences*, vol. 20, no. 13, p. 3328, 2019.

[84] Z. Wang, X. Wang, and Y. Jia, “Porphyromonas gingivalis promotes colorectal cancer development by regulating nlrp3 inflammasome signaling,” *Cancer Research*, vol. 79, no. 13_Supplement, pp. 2358–2358, 2019.

[85] M. S. Shah, T. Z. DeSantis, T. Weinmaier, P. J. McMurdie, J. L. Cope, A. Altrichter, J.-M. Yamal, and E. B. Hollister, “Leveraging sequence-based faecal microbial community survey data to identify a composite biomarker for colorectal cancer,” *Gut*, vol. 67, no. 5, pp. 882–891, 2018.

[86] D. Ternes, J. Karta, M. Tsenkova, P. Wilmes, S. Haan, and E. Letellier, “Microbiome in colorectal cancer: how to get from meta-omics to mechanism?,” *Trends in microbiology*, vol. 28, no. 5, pp. 401–423, 2020.

[87] A. Ribeiro Sobrinho, S. de Melo Maltos, L. Farias, M. De Carvalho, J. Nicoli, M. De Uzeda, and L. Vieira, “Cytokine production in response to endodontic infection in germ-free mice,” *Oral microbiology and Immunology*, vol. 17, no. 6, pp. 344–353, 2002.

[88] J. A. Lomholt and M. Kilian, “Immunoglobulin a1 protease activity in gemella haemolysans,” *Journal of clinical microbiology*, vol. 38, no. 7, pp. 2760–2762, 2000.

[89] X. Xia, W. K. K. Wu, S. H. Wong, D. Liu, T. N. Y. Kwong, G. Nakatsu, P. S. Yan, Y.-M. Chuang, M. W.-Y. Chan, O. O. Coker, *et al.*, “Bacteria pathogens drive host colonic epithelial cell promoter hypermethylation of tumor suppressor genes in colorectal cancer,” *Microbiome*, vol. 8, no. 1, pp. 1–13, 2020.

[90] F. Lu, J. Zhou, D. Lu, L. Ye, H. Liang, L. Lan, X. Yang, P. Cui, and J. Huang, “Changes of intestinal microbiota in colorectal cancer and its potential ability to predict disease,” 2022.

[91] T. L. Weir, D. K. Manter, A. M. Sheflin, B. A. Barnett, A. L. Heuberger, and E. P. Ryan, “Stool microbiome and metabolome differences between colorectal cancer patients and healthy adults,” *PloS one*, vol. 8, no. 8, p. e70803, 2013.

[92] J. Shimizu, T. Kubota, E. Takada, K. Takai, N. Fujiwara, N. Arimitsu, Y. Ueda, S. Wakisaka, T. Suzuki, and N. Suzuki, “Relative abundance of megamonas hypermegale and butyrivibrio species decreased in the intestine and its possible association with the t cell aberration by metabolite alteration in patients with behcet’s disease (210 characters),” *Clinical rheumatology*, vol. 38, no. 5, pp. 1437–1445, 2019.

[93] M. Zhuang, W. Shang, Q. Ma, P. Strappe, and Z. Zhou, “Abundance of probiotics and butyrate-production microbiome manages constipation via short-chain fatty acids production and hormones secretion,” *Molecular Nutrition & Food Research*, vol. 63, no. 23, p. 1801187, 2019.

[94] F. Beghini, L. J. McIver, A. Blanco-Míguez, L. Dubois, F. Asnicar, S. Maharjan, A. Mailyan, P. Manghi, M. Scholz, A. M. Thomas, *et al.*, "Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with biobakery 3," *Elife*, vol. 10, p. e65088, 2021.

[95] L. DongCheng, S. Liao, Y. Li, H. Lai, Y. Lin, and X. Liao, "Metagenomic meta-analysis of the gut microbiome in the different primary locations of colorectal cancer," 2022.

[96] L. Fan, C. Xu, Q. Ge, Y. Lin, C. C. Wong, Y. Qi, B. Ye, Q. Lian, W. Zhuo, J. Si, *et al.*, "A. muciniphila suppresses colorectal tumorigenesis by inducing tlr2/nlrp3-mediated m1-like tamsa. muciniphila induces m1-like tams that suppress crc," *Cancer Immunology Research*, vol. 9, no. 10, pp. 1111–1124, 2021.

[97] F. Wang, K. Cai, Q. Xiao, L. He, L. Xie, and Z. Liu, "Akkermansia muciniphila administration exacerbated the development of colitis-associated colorectal cancer in mice," *Journal of Cancer*, vol. 13, no. 1, p. 124, 2022.

[98] A. Cuevas-Sierra, J. I. Riezu-Boj, E. Guruceaga, F. I. Milagro, and J. A. Martínez, "Sex-specific associations between gut prevotellaceae and host genetics on adiposity," *Microorganisms*, vol. 8, no. 6, p. 938, 2020.

[99] J. Fernández, E. Ledesma, J. Monte, E. Millán, P. Costa, V. G. de la Fuente, M. T. F. García, P. Martínez-Camblor, C. J. Villar, and F. Lombó, "traditional processed meat products re-designed towards inulin-rich functional foods reduce polyps in two colorectal cancer animal models," *Scientific reports*, vol. 9, no. 1, pp. 1–17, 2019.

[100] T. Ju, J. Y. Kong, P. Stothard, and B. P. Willing, "Defining the role of parasutterella, a previously uncharacterized member of the core gut microbiota," *The ISME journal*, vol. 13, no. 6, pp. 1520–1534, 2019.

[101] I. Sobhani, E. Bergsten, S. Couffin, A. Amiot, B. Nebbad, C. Barau, N. de'Angelis, S. Rabot, F. Canoui-Poitrine, D. Mestivier, *et al.*, "Colorectal cancer-associated microbiota contributes to oncogenic epigenetic signatures," *Proceedings of the National Academy of Sciences*, vol. 116, no. 48, pp. 24285–24295, 2019.

[102] Y. Yang, Q. Cai, X.-O. Shu, M. D. Steinwandel, W. J. Blot, W. Zheng, and J. Long, "Prospective study of oral microbiome and colorectal cancer risk in low-income and african american populations," *International journal of cancer*, vol. 144, no. 10, pp. 2381–2389, 2019.

[103] J. Yu, Q. Feng, S. H. Wong, D. Zhang, Q. yi Liang, Y. Qin, L. Tang, H. Zhao, J. Stenvang, Y. Li, *et al.*, "Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer," *Gut*, vol. 66, no. 1, pp. 70–78, 2017.

[104] Y. Uchino, Y. Goto, Y. Konishi, K. Tanabe, H. Toda, M. Wada, Y. Kita, M. Beppu, S. Mori, H. Hijioka, *et al.*, "Colorectal cancer patients have four specific bacterial species in oral and gut microbiota in common—a metagenomic comparison with healthy subjects," *Cancers*, vol. 13, no. 13, p. 3332, 2021.

[105] H. N. Tran, T. N. H. Thu, P. H. Nguyen, C. N. Vo, K. Van Doan, C. N. N. Minh, N. T. Nguyen, K. A. Vu, T. D. Hua, T. N. T. Nguyen, *et al.*, "Mucosal microbiomes and fusobacterium genomics in vietnamese colorectal cancer patients," *bioRxiv*, 2022.

[106] J. Kharofa, S. Apewokin, T. Alenghat, and N. J. Ollberding, "Metagenomic analysis of the fecal microbiome in colorectal cancer patients compared to healthy controls as a function of age," *Cancer Medicine*, 2022.

[107] H. J. Freeman, "Effects of differing concentrations of sodium butyrate on 1, 2-dimethylhydrazine-induced rat intestinal neoplasia," *Gastroenterology*, vol. 91, no. 3, pp. 596–602, 1986.

[108] A. Belcheva, T. Irrazabal, S. J. Robertson, C. Streutker, H. Maughan, S. Rubino, E. H. Moriyama, J. K. Copeland, A. Surendra, S. Kumar, *et al.*, "Gut microbial metabolism drives transformation of msh2-deficient colon epithelial cells," *Cell*, vol. 158, no. 2, pp. 288–299, 2014.

[109] K. Yoon and N. Kim, "The effect of microbiota on colon carcinogenesis," *Journal of cancer prevention*, vol. 23, no. 3, p. 117, 2018.

[110] R. Kant, P. Rasinkangas, R. Satokari, T. E. Pietilä, and A. Palva, "Genome sequence of the butyrate-producing anaerobic bacterium anaerostipes hadrus pel 85," *Genome announcements*, vol. 3, no. 2, pp. e00224–15, 2015.

[111] J. Rafter, "Probiotics and colon cancer," *Best Practice & Research Clinical Gastroenterology*, vol. 17, no. 5, pp. 849–859, 2003.

[112] M. A. Zocco, M. E. Ainora, G. Gasbarrini, and A. Gasbarrini, "Bacteroides thetaiotaomicron in the gut: molecular aspects of their interaction," *Digestive and Liver Disease*, vol. 39, no. 8, pp. 707–712, 2007.

[113] K. Li, Z. Hao, J. Du, Y. Gao, S. Yang, and Y. Zhou, "Bacteroides thetaiotaomicron relieves colon inflammation by activating aryl hydrocarbon receptor and modulating cd4+ t cell homeostasis," *International immunopharmacology*, vol. 90, p. 107183, 2021.

[114] I. Belotserkovsky and P. J. Sansonetti, "Shigella and enteroinvasive escherichia coli," *Escherichia coli, a Versatile Pathogen*, pp. 1–26, 2018.

[115] Z. Zhang, L. Taylor, N. Shommu, S. Ghosh, R. Reimer, R. Panaccione, S. Kaur, J. E. Hyun, C. Cai, E. C. Deehan, *et al.*, "A diversified dietary pattern is associated with

a balanced gut microbial composition of faecalibacterium and escherichia/shigella in patients with crohn's disease in remission," *Journal of Crohn's and Colitis*, vol. 14, no. 11, pp. 1547–1557, 2020.

[116] C. Kasai, K. Sugimoto, I. Moritani, J. Tanaka, Y. Oya, H. Inoue, M. Tameda, K. Shiraki, M. Ito, Y. Takei, *et al.*, "Comparison of human gut microbiota in control subjects and patients with colorectal carcinoma in adenoma: Terminal restriction fragment length polymorphism and next-generation sequencing analyses," *Oncology reports*, vol. 35, no. 1, pp. 325–333, 2016.

[117] R.-X. Huo, Y.-J. Wang, S.-B. Hou, W. Wang, C.-Z. Zhang, and X.-H. Wan, "Gut mucosal microbiota profiles linked to colorectal cancer recurrence," *World Journal of Gastroenterology*, vol. 28, no. 18, p. 1946, 2022.

[118] J. C. Ezeji, D. K. Sarikonda, A. Hopperton, H. L. Erkkila, D. E. Cohen, S. P. Martinez, F. Cominelli, T. Kuwahara, A. E. Dichosa, C. E. Good, *et al.*, "Parabacteroides distasonis: intriguing aerotolerant gut anaerobe with emerging antimicrobial resistance and pathogenic and probiotic roles in human health," *Gut Microbes*, vol. 13, no. 1, p. 1922241, 2021.

[119] L. Polimeno, M. Barone, A. Mosca, M. T. Viggiani, F. Joukar, F. Mansour-Ghanaei, S. Mavaddati, A. Daniele, L. Debellis, M. Bilancia, *et al.*, "Soy metabolism by gut microbiota from patients with precancerous intestinal lesions," *Microorganisms*, vol. 8, no. 4, p. 469, 2020.

[120] Z. Gao, B. Guo, R. Gao, Q. Zhu, and H. Qin, "Microbiota disbiosis is associated with colorectal cancer," *Frontiers in microbiology*, vol. 6, p. 20, 2015.

[121] P. H. Leung, R. Subramanya, Q. Mou, K. T.-w. Lee, F. Islam, V. Gopalan, C.-t. Lu, and A. K.-y. Lam, "Characterization of mucosa-associated microbiota in matched cancer and non-neoplastic mucosa from patients with colorectal cancer," *Frontiers in microbiology*, vol. 10, p. 1317, 2019.

[122] D. Demeyer, B. Mertens, S. De Smet, and M. Ulens, "Mechanisms linking colorectal cancer to the consumption of (processed) red meat: a review," *Critical reviews in food science and nutrition*, vol. 56, no. 16, pp. 2747–2766, 2016.

[123] F. N. Alkanany, S. A. Gmais, A. A. Maki, A. M. Altaee, *et al.*, "Estimation of bacterial biodegradability of pah in khor al-zubair channel, southern iraq," *International Journal of Marine Science*, vol. 7, 2017.

[124] F. Wu, X. Guo, J. Zhang, M. Zhang, Z. Ou, and Y. Peng, "Phascolarctobacterium faecium abundant colonization in human gastrointestinal tract," *Experimental and therapeutic medicine*, vol. 14, no. 4, pp. 3122–3126, 2017.

[125] L. Li, Q. Su, B. Xie, L. Duan, W. Zhao, D. Hu, R. Wu, and H. Liu, "Gut microbes in correlation with mood: case study in a closed experimental human life support system," *Neurogastroenterology & Motility*, vol. 28, no. 8, pp. 1233–1240, 2016.

[126] B. J. Parker, P. A. Wearsch, A. C. Veloo, and A. Rodriguez-Palacios, “The genus alistipes: gut bacteria with emerging implications to inflammation, cancer, and mental health,” *Frontiers in immunology*, vol. 11, p. 906, 2020.

[127] A. R. Moschen, R. R. Gerner, J. Wang, V. Klepsch, T. E. Adolph, S. J. Reider, H. Hackl, A. Pfister, J. Schilling, P. L. Moser, *et al.*, “Lipocalin 2 protects from inflammation and tumorigenesis associated with gut microbiota alterations,” *Cell host & microbe*, vol. 19, no. 4, pp. 455–469, 2016.

[128] C. Wang, J. Zhao, H. Zhang, Y.-K. Lee, Q. Zhai, and W. Chen, “Roles of intestinal bacteroides in human health and diseases,” *Critical reviews in food science and nutrition*, vol. 61, no. 21, pp. 3518–3536, 2021.

[129] T. Wang, G. Cai, Y. Qiu, N. Fei, M. Zhang, X. Pang, W. Jia, S. Cai, and L. Zhao, “Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers,” *The ISME journal*, vol. 6, no. 2, pp. 320–329, 2012.

[130] Y. Wu, N. Jiao, R. Zhu, Y. Zhang, D. Wu, A.-J. Wang, S. Fang, L. Tao, Y. Li, S. Cheng, *et al.*, “Identification of microbial markers across populations in early detection of colorectal cancer,” *Nature communications*, vol. 12, no. 1, pp. 1–13, 2021.

[131] T. Takada, K. Watanabe, H. Makino, and A. Kushiro, “Reclassification of eubacterium desmolans as butyricicoccus desmolans comb. nov., and description of butyricicoccus faecihominis sp. nov., a butyrate-producing bacterium from human faeces,” *International Journal of Systematic and Evolutionary Microbiology*, vol. 66, no. 10, pp. 4125–4131, 2016.

[132] W. Moore and L. H. Moore, “Intestinal floras of populations that have a high risk of colon cancer,” *Applied and environmental microbiology*, vol. 61, no. 9, pp. 3202–3207, 1995.

[133] Q. Xiao, W. Lu, X. Kong, Y. W. Shao, Y. Hu, A. Wang, H. Bao, R. Cao, K. Liu, X. Wang, *et al.*, “Alterations of circulating bacterial dna in colorectal cancer and adenoma: A proof-of-concept study,” *Cancer letters*, vol. 499, pp. 201–208, 2021.

[134] A. A. Hibberd, A. Lyra, A. C. Ouwehand, P. Rolny, H. Lindegren, L. Cedgård, and Y. Wettergren, “Intestinal microbiota is altered in patients with colon cancer and modified by probiotic intervention,” *BMJ open gastroenterology*, vol. 4, no. 1, p. e000145, 2017.

[135] L. Mira-Pascual, R. Cabrera-Rubio, S. Ocon, P. Costales, A. Parra, A. Suarez, F. Moris, L. Rodrigo, A. Mira, and M. Collado, “Microbial mucosal colonic shifts associated with the development of colorectal cancer reveal the presence of different bacterial and archaeal biomarkers,” *Journal of gastroenterology*, vol. 50, no. 2, pp. 167–179, 2015.

[136] B. A. Peters, C. Dominianni, J. A. Shapiro, T. R. Church, J. Wu, G. Miller, E. Yuen, H. Freiman, I. Lustbader, J. Salik, *et al.*, "The gut microbiota in conventional and serrated precursors of colorectal cancer," *Microbiome*, vol. 4, no. 1, pp. 1–14, 2016.

[137] E. Allen-Vercoe and C. Jobin, "Fusobacterium and enterobacteriaceae: important players for crc?," *Immunology letters*, vol. 162, no. 2, pp. 54–61, 2014.

[138] J. Park, N.-E. Kim, H. Yoon, C. M. Shin, N. Kim, D. H. Lee, J. Y. Park, C. H. Choi, J. G. Kim, Y.-K. Kim, *et al.*, "Fecal microbiota and gut microbe-derived extracellular vesicles in colorectal cancer," *Frontiers in oncology*, p. 3351, 2021.

[139] V. Sarhadi, L. Lahti, F. Saberi, O. Youssef, A. Kokkola, T. Karla, M. Tikkanen, H. Rautelin, P. Puolakkainen, R. Salehi, *et al.*, "Gut microbiota and host gene mutations in colorectal cancer patients and controls of iranian and finnish origin," *Anticancer research*, vol. 40, no. 3, pp. 1325–1334, 2020.

[140] A. J. La Reau and G. Suen, "The ruminococci: key symbionts of the gut ecosystem," *Journal of microbiology*, vol. 56, no. 3, pp. 199–208, 2018.

[141] B. Dassa, I. Borovok, R. Lamed, N. Koropatkin, E. Martens, B. White, A. Bernalier-Donadille, S. Duncan, H. Flint, E. Bayer, *et al.*, "Ruminococcal cellulosome systems from rumen to human.," *Environmental Microbiology*, vol. 17, no. 9, pp. 3407–3426, 2015.

[142] F. Zakham, T. Pillonel, A.-S. Brunel, P.-Y. Zambelli, G. Greub, A. Croxatto, and C. Bertelli, "Molecular diagnosis and enrichment culture identified a septic pseudoarthrosis due to an infection with erysipelatoclostridium ramosum," *International Journal of Infectious Diseases*, vol. 81, pp. 167–169, 2019.

[143] A. C. Ouwehand, S. Salminen, T. Arvola, T. Ruuska, and E. Isolauri, "Microbiota composition of the intestinal mucosa: association with fecal microbiota?," *Microbiology and immunology*, vol. 48, no. 7, pp. 497–500, 2004.

[144] B. Flemer, D. B. Lynch, J. M. Brown, I. B. Jeffery, F. J. Ryan, M. J. Claesson, M. O'Riordain, F. Shanahan, and P. W. O'Toole, "Tumour-associated and non-tumour-associated microbiota in colorectal cancer," *Gut*, vol. 66, no. 4, pp. 633–643, 2017.

[145] T.-W. Yang, W.-H. Lee, S.-J. Tu, W.-C. Huang, H.-M. Chen, T.-H. Sun, M.-C. Tsai, C.-C. Wang, H.-Y. Chen, C.-C. Huang, *et al.*, "Enterotype-based analysis of gut microbiota along the conventional adenoma-carcinoma colorectal cancer pathway," *Scientific reports*, vol. 9, no. 1, pp. 1–13, 2019.

[146] "SHAP explainers: Permutation explainer." `https://shap.readthedocs.io/en/latest/generated/shap.explainers.Permutation.html`. Accessed: 2023-02-24.