Bioinformatics Tools and Explainable Machine Learning Approaches for Colorectal Cancer Genomic and Metagenomic Data Analysis

A thesis presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

School of Mathematical and Computational Sciences at Massey University, Albany Campus,

Auckland, New Zealand

Tyler Kolisnik

2024

Copyright:

Supervisory Team:

Dr. Olin K. Silander, PhD,
Senior Research Fellow, Liggins Institute, University of Auckland, Auckland, New Zealand

Dr. Sebastian Schmeier, PhD,
Senior Lecturer in Bioinformatics, School of Natural and Computational Sciences, Massey University, Auckland, New Zealand *(former)*

Dr. Steven Jones, PhD,
Head of Bioinformatics and Co-Director, Canada's Michael Smith Genome Sciences Centre at BC Cancer & Director, Bioinformatics Graduate Program, University of British Columbia, Vancouver, BC, Canada

Dr. Adam N. H. Smith, PhD,
Senior Lecturer in Statistics, School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand

Examination Panel:

Dr. Justin O'Sullivan, PhD,
Director, Liggins Institute, University of Auckland, Auckland, New Zealand

Dr. Xochitl Morgan, PhD,
Director, Harvard Chan Microbiome Analysis Core, Harvard University, USA

Dr. Jonathan Marshall, PhD,
Associate Professor in Statistics, School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand

Abstract

Colorectal cancer (CRC) is a leading cause of cancer-related mortality worldwide and is influenced by complex interactions between genetic factors and the microbiome. The advent of high-throughput sequencing technologies has led to the generation of vast amounts of genomic and metagenomic data, providing opportunities to uncover novel biological markers (biomarkers) for CRC diagnosis, prognosis, and treatment. However, analyzing such datasets poses significant computational and interpretive challenges, necessitating the development of efficient and user-friendly bioinformatics tools.

Recent advancements in machine learning, particularly Random Forest (RF) models, have shown promise in identifying predictive features in genomic data. Yet, existing implementations often face limitations in scaling and interpretability, especially when applied to large genomic studies. Additionally, integrating host genomic and microbial metagenomic data remains a complex task due to the heterogeneity of data types and sophisticated analytical methods required.

This thesis focuses on the development of computational tools and the application of machine learning techniques to enhance the analysis of genomic and metagenomic data for colorectal cancer research. Firstly, I present the MetaFunc App, an interactive Shiny application designed to facilitate the exploration of data generated from the MetaFunc pipeline, an analysis

pipeline for host and microbiome transcriptome data. The app provides a user-friendly interface for visualizing and analyzing microbial taxonomic profiles alongside host gene expression data, linking functional annotations to specific microbial taxa. This integration facilitates a deeper understanding of microbial contributions to a designated target outcome, e.g. cancer versus normal, and aids in identifying potential microbial biomarkers and eliciting their functions.

Secondly, I apply Random Forest machine learning models to identify genomic and microbial biomarkers that differentiate right-sided colorectal cancer (RCC) from left-sided colorectal cancer (LCC). Utilizing RNA-seq data for 58,677 coding and non-coding human genes, and count data for 28,577 microbial taxa from 308 patient tumour samples, I develop three models: a genes-only model, a microbes-only model, and a combined genes-and-microbes model. The genes-only model achieves an accuracy of 90%, identifying significant genomic features such as *PRAC1*, *HOXB13*, *HOXC4*, and *HOXC6*, which are associated with colorectal cancer location and development. The microbes-only model achieves an accuracy of 70%, identifying significant microbial features including *Ruminococcus gnavus* and *Fusobacterium nucleatum.* The combined model achieves an accuracy of 87%, which may reflect an association between microbial communities and host gene expression in CRC.

Finally, I present pyRforest, an R package that integrates Python's scikit-learn RandomForestClassifier into the R environment via the reticulate package to enhance computational efficiency and memory management when using Random Forest models in R to

analyze large genomic datasets. pyRforest also includes a novel rank-based permutation method for calculating p-values of individual features for feature identification. Additionally, it includes the capacity for calculating and plotting SHapley Additive exPlanations (SHAP) to interpret the contribution of each feature to model predictions, enhancing the explainability of Random Forest model results. The utility of pyRforest is demonstrated through a case study, where it is used to identify candidate biomarkers and provide insights into biological significance.

Collectively, this work advances the understanding of genomic and microbial factors influencing colorectal cancer and provides advanced computational tools that can be used in other analyses. The MetaFunc App and pyRforest package facilitate the integration and interpretation of complex genomic and metagenomic data, and represent valuable resources for biomarker discovery. By addressing current challenges in data analysis and in bioinformatics software development, this thesis lays the groundwork for future research in bioinformatics and oncology, ultimately aiming to create and implement tools for improved genomic and metagenomic cancer dataset analysis.

Acknowledgements

I would like to express my sincerest thanks to my doctoral supervisors, Dr. Olin Silander, Dr. Adam Smith, Dr. Sebastian Schmeier, and Dr. Steven Jones, this thesis would not have been possible without their guidance and support. I'm profoundly thankful to Dr. Jones, whose supervision at the British Columbia Genome Sciences Centre allowed me to return home to Canada to continue my studies at Massey from "abroad" during the uncertain times of the global COVID-19 pandemic. I am especially thankful to all my fellow students and colleagues at Massey University, especially Arielle Sulit, and my fellow students at the Genome Sciences Centre, especially Faeze Keshavarz-Rahaghi. I would like to also thank Lucia Lam, who has been an amazing mentor and has helped me so much with my career in bioinformatics. I would like to thank my high school math and physics teacher, Roger Pavey, for everything he taught me that set me up for success with my education. I would like to thank my amazing neighbours and friends Samina and Aodhan for making us feel welcome in New Zealand.

My PhD studies would not have been possible without a PhD scholarship from the Massey University School of Natural Sciences, for which I am profoundly grateful.

Thank you to my parents Teresa and Steven, and my brother Kelly for always providing me with encouragement, and for always being supportive of my decisions. Thank you to my partner Layton and my cat Fry for their love and support and for coming on the journey abroad to New Zealand with me.

I would also like to express my deepest appreciation for the support of my friends and extended family throughout my PhD. Thank you to my amazing friends Shyy, Stefanie, Rachel, and Naomi. Thank you to my friend Jeff for making sure that not all of my time on computers is spent coding. Thank you to my in-laws Caron and Blaine for always being there to help us during tough times. A special thank you goes out to my cousins, Elise, Mercedes, Mike, Troy, Jen, and Jade, my aunts and uncles, especially my aunts Rita and Rose, and my uncle Aaron, and all those who sent me packages from Canada while I was studying abroad, it made me feel a lot closer to home.

"Things are only impossible until they are not." - Captain Jean-Luc Picard
"Tea, earl grey, hot." - Also Captain Jean-Luc Picard

List of Publications

The following published works were completed during my doctoral candidature.

The results presented in Chapter 2 of this thesis contributed to the following publication, which is attached in **Appendix A**:

Sulit, A. K., **Kolisnik, T.**, Frizelle, F. A., Purcell, R., & Schmeier, S. (2023). MetaFunc: taxonomic and functional analyses of high throughput sequencing for microbiomes. *Gut Microbiome, 4*(4), 1-21. https://doi.org/10.1017/gmb.2022.12.

The results presented in Chapter 3 of this thesis were published as:

Kolisnik, T., Sulit, A. K., Schmeier, S., Frizelle, F., Purcell, R., Smith, A., & Silander, O. (2023). Identifying important microbial and genomic biomarkers for differentiating right- versus left-sided colorectal cancer using random forest models. *BMC Cancer, 23*(647), 1-11. https://doi.org/10.1186/s12885-023-10848-9.

The results presented in Chapter 4 of this thesis were published as:

Kolisnik, T., Keshavarz-Rahaghi, F., Purcell, R., Smith, A., & Silander, O. (2024). pyRforest: A comprehensive R package for genomic data analysis featuring scikit-learn Random Forests in R. *Briefings in Functional Genomics, 2024*(38), 1-9. https://doi.org/10.1093/bfgp/elae038.

In addition, during my time studying at the BC Genome Sciences Centre I contributed to the following publication:

Keshavarz-Rahaghi, F., Pleasance, E., **Kolisnik, T.**, & Jones, S. J. M. (2022). A p53 transcriptional signature in primary and metastatic cancers derived using machine learning. Frontiers in Genetics, 13, 987238. https://doi.org/10.3389/fgene.2022.987.

List of Abbreviations

AI	Artificial Intelligence
APC	Adenomatous Polyposis Coli (Gene)
AUC	Area Under the Curve
AUROC	Area Under the Receiver Operator Characteristic Curve
BRAF	B-Raf proto-oncogene (Gene)
CMS	Consensus Molecular Subtypes
CPM	Counts-Per-Million
CRC	Colorectal Cancer
DE	Differential Expression
DNA	Deoxyribonucleic Acid
EMT	Epithelial-Mesenchymal Transition
EGFR	Epidermal Growth Factor Receptor (Gene)
FMT	Fecal Microbiota Transplant
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
GB	Gigabyte
GHz	Gigahertz
HOXB13	Homeobox B13 (Gene)
HOXC4	Homeobox C4 (Gene)
HOXC6	Homeobox C6 (Gene)
HOXC8	Homeobox C8 (Gene)
KEGG	Kyoto Encyclopedia of Genes and Genomes
KNN	k-Nearest Neighbours
KRAS	Kirsten RAt Sarcoma
LCC	Left-sided Colorectal Cancer
ML	Machine Learning
MLH1	MultiL Homolog 1 (Gene)
MSH1	MutS Homolog 1 (Gene)
MSH6	MutS Homolog 6 (Gene)

MSI	Microsatellite Instability
NCBI	National Center for Biotechnology Information
NeSI	New Zealand e-Science Infrastructure
PCA	Principal Component Analysis
PIK3CA	Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha (Gene)
PMS2	Postmeiotic Segregation Increased 2 (Gene)
PRAC1	Prostate Cancer Susceptibility Candidate 1 (Gene)
RCC	Right-sided Colorectal Cancer
RF	Random Forest
RNA	Ribonucleic Acid
RNA-seq	Ribonucleic Acid Sequencing
RNN	Recurrent Neural Networks
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machines
TGFBR2	Transforming Growth Factor Beta Receptor 2 (Gene)
TNM	Tumour Node Metastasis (Cancer staging system)
TPM	Transcripts-Per-Million
WES	Whole-Exome Sequencing
WGS	Whole-Genome Sequencing

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Colorectal Cancer

1.1.1 Epidemiology

Worldwide, colorectal cancer (CRC) is the second most prevalent cancer in women, and the third most prevalent cancer in men, with approximately 2 million newly diagnosed cases occurring every year (WCRF International, 2022). CRC represents 10% of all global cancers (IARC, 2019). Age and family history are the greatest risk factors for CRC, with over 99% of cases occurring in people over 40. CRC is known to arise from polyps, which are groups of cells that begin as benign and then later become malignant (Ballinger & Anggiansah, 2007). Due to the invasive nature of colonoscopy, which is the primary diagnostic method for colorectal cancer, and the non-specific symptoms of CRC, many cases go undiagnosed until the disease has progressed to an advanced stage.

1.1.2 Colorectal Cancer Nomenclature

Colorectal cancer refers to cancers that can arise in the colon, rectosigmoid junction or rectum (IARC, 2019; Mayo Clinic, 2022). The term CRC has become the preferred term by many clinicians and researchers because it encompasses all malignancies across the lower gastrointestinal tract. However, historically, there has been a lack of standardization in the nomenclature for CRC, which presents an issue when reconciling data from older studies with modern research (Paschke et al., 2018). Some clinicians prefer to refer to specific anatomical locations, leading to distinctions such as ‘colon cancer’ versus ‘rectal cancer’, while others continue to use the more general terms ‘bowel cancer’ or ‘colorectal cancer’. Recently, further distinctions within CRC have become an important topic in the research community. Terms such

as ‘proximal’ or ‘right-sided’ and ‘distal’ or ‘left-sided’ colorectal cancer are now becoming more widely used to describe tumours based on their specific location within the colon. This is due to emerging evidence that right-sided and left-sided CRCs exhibit distinct molecular characteristics, clinical presentations, and responses to treatment (Lee et al., 2015).

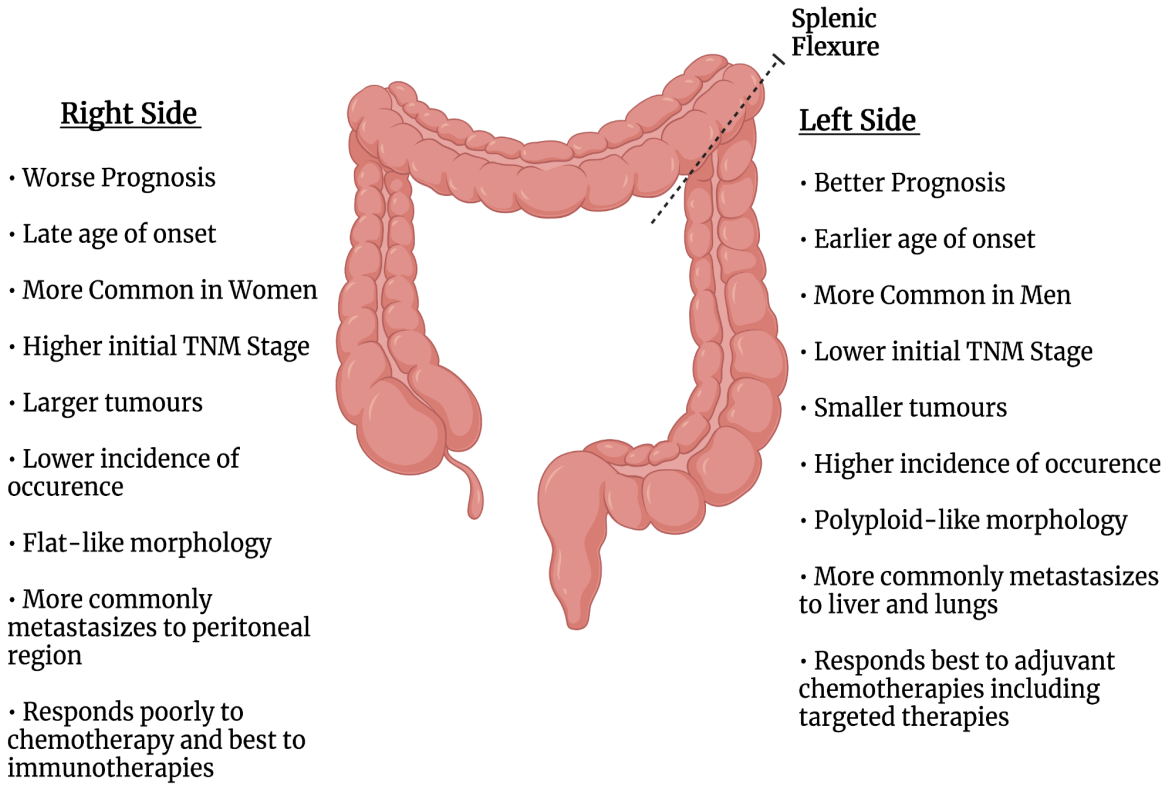
1.1.3 CRC Sidedness and Clinical Implications

Cancers located in the proximal part of the colon, comprising cancers of the cecum, ascending colon, hepatic flexure, and proximal part of the transverse colon are referred to as right-sided colorectal cancers (RCC) (Mukund et al., 2020, Stintzing et al., 2017). Cancers located in the distal portion of the colon (beyond the splenic flexure), descending colon, sigmoid colon, and rectum, are referred to as left-sided colorectal cancers (LCC). These two portions of the colon, while linked in sequence, play different roles in the digestive process and have different developmental origins, with the right-sided colon developing from the embryonic midgut, and the left-sided colon developing from the embryonic hindgut (Kostouros et al., 2020). **Figure 1.1** illustrates several differences in the pathology and treatment of right versus left-sided colorectal cancer (Baran et al., 2018; Lee et al., 2015).

In terms of epidemiology, RCC is more common in women than in men, and typically presents with larger tumours, higher tumour node metastasis (TNM) classification stage and poorer overall survival outcomes (Liang et al., 2018; Yang et al., 2016; Zhao et al., 2020). In contrast, LCC is more common overall than RCC, especially in men, and it is often identified at a lower TNM (Tumour Node Metastasis) stage, with the tumours being smaller. LCCs generally have a more favourable prognosis than RCCs.

Figure 1.1

Summary of differences between left and right-side colorectal cancer.



Note. This figure is a derivative work based on information from Lee et al. 2015, with additional information added from Baran et al. 2018. Permission to create a derivative work was obtained from Elsevier under license number 5896941135490.

1.1.4 Molecular and Genetic Differences between RCC and LCC

The differences between RCC and LCC are not merely anatomical but extend to their molecular and genetic landscapes. RCC is characterized by a higher frequency of mutations caused by defects in DNA repair called microsatellite instability (MSI), *BRAF* (B-Raf proto-oncogene) mutations, and CpG island methylator phenotype positivity, all of which contribute to its rapid progression and poorer prognosis (Baran et al., 2018; Lee et al., 2015). RCC tumours tend to be more resistant to standard chemotherapy and respond better to immunotherapy. In contrast, LCC tumours typically exhibit chromosomal instability and

mutations in genes like the proto-oncogenes *APC* (Adenomatous Polyposis Coli) and *KRAS* (Kirsten RAt Sarcoma), which are more commonly targeted by conventional treatments like chemotherapy and *EGFR* (Epidermal Growth Factor Receptor) inhibitors (Natsume et al., 2018). Understanding the molecular genetic distinctions between RCC and LCC is critical as it impacts prognosis and shapes the therapeutic strategies available to patients.

1.1.5 CRC Treatments

Treatment options for colorectal cancer can include surgery, chemotherapy, immunotherapy, and less commonly, radiotherapy (Kuipers et al., 2015). Early-stage CRC can often be treated with minimally invasive surgical techniques such as mucosal resection or submucosal dissection during colonoscopy. For later stage cancers, more radical surgical options such as partial colectomy are common.

Chemotherapy is a cornerstone of treatment for CRC, and may be delivered in a neoadjuvant (before surgery) or adjuvant (after surgery) setting. Frequently used chemotherapy drugs include 5-flurouracil, capacitabine, oxaliplatin, irinotecan, and targeted biologics such as bevacizumab, aflibercept, cetuximab, and panitumumab, depending on the genetic profile of the tumour (Kumar et al., 2023). Radiation therapy is not commonly used, except for low-stage rectal cancers, due to the sensitivity of the bowel to radiation (Häfner & Debus, 2016).

Among the most promising treatments in colorectal cancer are the immunotherapeutic options. A recent study on 12 patients with left-sided CRC showed that treatment with dostarlamib, an anti-PD-1 monoclonal antibody, achieved complete clinical remission for 12 months after their last dose (Cercek et al., 2022). While the clinically complete response in this trial is promising, there are two key caveats that these patients specifically had rectal adenocarcinoma (a form of LCC) with a genetic mutations for loss of expression of the *MLH1*,

MSH1, *MSH6*, and *PMS2* genes. This study placed an emphasis on anatomic location, the microbiome, and gene expression in shaping treatment decisions.

1.1.6 The Gut Microbiome in CRC

The human gastrointestinal tract hosts a diverse community of microorganisms, collectively known as the gut microbiome (Thursby & Juge, 2017). This microbial ecosystem plays an important role in maintaining homeostasis by aiding digestion, synthesizing vitamins, and modulating the immune system. The balance of microbial populations is crucial for health, and disruptions can lead to various diseases. In the past twenty years, projects such as The Human Microbiome Project (Human Microbiome Project Consortium, 2012) have started to characterize the importance of microbial interactions and colon epithelial cells in the human gut. However, the roles that specific microbes play in the development and progression of colorectal cancer remain poorly understood (Zou et al., 2018). Therapeutic interventions targeting the gut microbiome have shown significant success in other conditions; for instance, fecal microbiota transplants (FMT) are now a standard of care for recurrent *Clostridium difficile* infections, with cure rates of up to 90% (Rohlke & Stollman, 2012). This success raises the possibility of exploring similar microbiome-targeted therapies for CRC. Despite this success, stool transplants have not been extensively trialed for treatment of CRC, although some early studies such as those conducted in mice are showing promising results (Yu et al., 2023). In this study, FMT inhibited CRC progression in mice by reversing intestinal microbial dysbiosis, reducing tumour size and number, and significantly prolonging survival rates when compared to the group that did not receive FMT. The study also noted an increase in infiltration of CD8+ T cells, and inflammatory cytokines, suggesting an enhanced anti-cancer immune response in the FMT group.

The composition of the gut microbiome is unique to every individual, which presents a significant challenge when studying microbe-gene interactions (Gilbert et al., 2018). In addition to cancer, numerous other diseases, such as Crohn's disease, irritable bowel syndrome (IBS), obesity and autism, have been suggested to have links to the gut microbiome; however, therapeutic options targeting the microbiome remain largely unsuccessful at treating these diseases, and the specific microbes involved and their roles in influencing gene expression remain somewhat of a mystery (Gilbert et al., 2018). In colorectal cancer, one of the leading theories postulates that a dysbiosis of the gut microbiome results in inflammation which then leads to tumour initiation, progression, and proliferation (Kostic et al., 2013).

Inflammation is considered a hallmark of cancer development (Hanahan & Weinberg, 2011), and microbes have been shown to produce substances that induce cellular inflammation, contributing to tumourigenesis (Brennan et al., 2016; Grivennikov et al., 2010). In CRC, the gut microbiome plays a significant role by modulating inflammation, epithelial cell proliferation, and apoptosis (Gao et al., 2015). Certain microbial species have been implicated in CRC development through their ability to produce pro-inflammatory and genotoxic substances (Kostic et al., 2013).

One such microbe is *Fusobacterium nucleatum*, an anaerobic Gram-negative bacterium found in higher abundance in CRC tissues compared to normal tissues (Kostic et al., 2013; Wang et al., 2022). *F. nucleatum* has been shown to promote carcinogenesis by adhering to and invading epithelial cells via FadA adhesin to E-cadherin binding and activation of β-catenin signalling pathways which promote inflammatory responses (Rubinstein et al., 2013). Additionally, certain strains of *Escherichia coli* can produce colibactin, a genotoxin that induces

DNA double-strand breaks, potentially promoting tumour initiation (Rosendahl Huber et al., 2024).

Ruminococcus and *Akkermansia* species have also been associated with CRC. Both of these bacterial species are involved in the degradation of mucin and can disrupt the mucus layer, potentially facilitating contact between luminal contents and the epithelium, thereby promoting inflammation and subsequent carcinogenesis (Coleman et al., 2021).

Identifying specific microbes associated with CRC can aid in early diagnosis and risk assessment (Zackular et al., 2014). Exploring microbial biomarkers enhances our understanding of disease mechanisms and opens up possibilities for microbiome-targeted interventions, such as dietary modifications, probiotics, prebiotics, antibiotics, and fecal transplants. There remains a significant gap in our understanding of how complex microbial communities interact with host factors to influence CRC development and progression (Brennan et al., 2016). Current research often focuses on individual bacterial species or specific pathways, but the synergistic effects of the microbiome as a whole and its interplay with gene expression have not been fully elucidated, especially while also considering the implications of right-side versus left-side CRC tumour location. This demonstrates the growing recognition of the importance of integrating host genomic data with not only tumour location, but also microbiome profiles to understand the complex interactions between the tumour, host, and microbiome (Garza et al., 2020).

1.1.7 Advances in Genomic Technologies

1.1.7.1 High-Throughput Sequencing Technologies

The advent of next-generation sequencing (NGS) technologies has revolutionized genomic research by enabling rapid and cost-effective large-scale sequencing of DNA and RNA (Goodwin et al., 2016). Techniques such as whole-genome sequencing (WGS), whole-exome

sequencing (WES), and RNA sequencing (RNA-seq) have become indispensable tools in cancer genomics, facilitating the identification of genetic mutations, gene expression profiles, and alternative splicing events (Metzker, 2010).

Metagenomic sequencing allows for characterization of microbial communities, providing insights into microbial diversity, function, and interactions with the host without the use of sample cultures (Sharpton, 2014).

1.1.7.2 Challenges in Data Normalization and Scaling

The generation of large-scale genomic and metagenomic datasets presents significant analytical challenges. These datasets often contain tens of thousands of features (e.g. genes, microbial taxa) and require computational methods that can manage their complexity while minimizing potential sources of error. Data normalization is a critical step to adjust for technical variation such as sequencing depth, gene length, and library size, ensuring that comparisons reflect true biological differences (Conesa et al., 2016).

For RNA-seq data, normalization approaches like transcripts per million (TPM) and fragments per kilobase of transcript per million mapped reads (FPKM) are commonly used to account for sequencing depth and gene length (Conesa et al., 2016). TPM normalization facilitates comparison of gene expression levels across samples by scaling for both the length of the gene and the total number of reads across samples, thus providing a consistent framework for downstream analysis.

In metagenomic analyses, normalization techniques must account for differences in sequencing effort, sample composition, and the inherently uneven distribution of microbial taxa. As used in Chapter 2, microbial abundances can be expressed as percentages, calculated by dividing the count of each taxon by the total read counts mapped to all taxa in a sample and

multiplying by 100 (Sulit et al., 2023). This method provides a straightforward interpretation of the proportion of each microbial taxon within a sample. In Chapters 3 and 4, we utilize counts per million (CPM) normalization which adjusts for differences in library sizes by scaling raw counts to a common scale based on total counts, enabling comparison of relative abundances across samples, and ensuring consistency of scaling with the TPM normalized genomic data.

1.2 Computational Approaches in Genomic Research

1.2.1 Use of Applications for Exploring Metagenomic Data

The analysis of metagenomic data presents significant challenges due to the complexity and vast diversity present in microbial communities. Metagenomic datasets typically contain a mixture of sequences from different bacterial, viral, and eukaryotic organisms (Lapidus et al., 2021). These sequences can be analyzed to identify genes, proteins, and infer potential functions, providing insights into the functional profiles of the gut microbiome (Sharpton, 2014). Extracting meaningful biological insights from these datasets requires appropriate analytical tools. When it comes to building custom pipelines for genomic data analysis, R Shiny has become a popular tool amongst bioinformaticians for building applications that make it possible for researchers without coding expertise to display and interact with biological data.

Currently, there is an absence of integrated solutions that allow researchers to perform gene ontology-based functional microbiome analyses within a single platform. No pre-existing R Shiny tools enable the exploration of host and microbial abundances and their taxonomic profiles while linking these to gene ontology annotations. Consequently, researchers must employ multiple separate tools, resulting in fragmented workflows that take considerable time to set up. This lack of an integrated solution poses challenges for conducting holistic studies on

host-microbiome interactions, particularly in the context of diseases like colorectal cancer, where understanding the functional roles of microbial communities is essential.

1.2.2 Existing Applications for Exploring Metagenomic Data

Despite the lack of integrated solutions, some applications have been developed to make certain aspects of microbiome analysis more accessible and interactive. Phyloseq (McMurdie & Holmes, 2012) is a widely used R shiny app that has tools for analysis and display of microbiome census data. It enables the production of graphical plots for ecological analyses. However, Phyloseq is primarily focused on phylogenetic analyses and lacks comprehensive functional annotation capabilities. It does not provide functionalities for displaying host gene expression data or performing gene ontology analyses, which limits its utility in studies aiming to understand host-microbiome interactions at the functional level.

Animalcules (Zhao et al., 2020) is another R Shiny application designed for microbiome data analysis and visualization. It offers an interactive interface for exploring microbiome datasets, including functionalities for differential abundance analysis, diversity analysis, and visualization of microbiome community compositions. Animalcules does not offer functional insights into gene ontology analysis and requires that users preprocess their data from existing tools.

While these applications have made certain aspects of microbiome analysis more accessible and interactive, they lack the ability to perform gene ontology based functional microbiome analysis. In addition to accessible tools for host-microbiome exploration, new methods that can lead to candidate biomarker discovery and new biological insights are required.

1.2.3 Application of Machine Learning Approaches in CRC Research

Machine learning (ML) is a broad range of statistical techniques that enable computers to learn from data and fit models that can be used to make predictions, classify data, or to support decision-making (Davenport et al., 2019). Rather than being explicitly programmed to perform specific tasks, ML algorithms identify patterns from training data and can generalize to make predictions on unseen data. In biomedical research, ML excels at analyzing complex datasets to reveal insights that can improve diagnosis, prognosis, and inform treatment strategies. ML techniques have been applied to many aspects of CRC research, including early detection, classification of tumour subtypes, prediction of patient outcomes, and identification of candidate biomarkers (Alboaneen et al., 2023).

1.2.3.1 Historical use of ML for CRC Research

ML algorithms have been utilized in CRC research to analyze many types of data, including genomic, transcriptomic, proteomic, and imaging data (Kourou et al., 2015). ML algorithms fall into different non-exclusive categories including unsupervised, supervised, deep learning and ensemble, each with their own strengths and weaknesses (Jordan & Mitchell, 2015).

Supervised algorithms learn from labeled training data, where each input is associated with a known output label to predict labels on new unseen data. This includes approaches such as support vector machines (SVM) (Cortes et al., 1995), k-nearest neighbours (KNN) (Cover & Hart, 1967), decision trees (Quinlan, 1986), and Random Forests (Breiman, 2001). In unsupervised approaches, algorithms learn from data identifying patterns, data structures and relationships without any predefined outcome labels. Unsupervised techniques include clustering (e.g. hierarchical clustering (Lloyd, 1982) or k-means clustering (Johnson, 1967)), and dimensionality reduction techniques like principal component analysis (PCA) (Jolliffe, 2013).

Deep learning is another subset of machine learning which involves neural networks with multiple layers (called deep architectures) that can learn hierarchical representations of data (LeCun et al., 2015). This includes convolutional neural networks (CNNs) (O'Shea et al., 2015), recurrent neural networks (RNNs) (Sherstinsky, 2018). Lastly, ensemble approaches are another method which combine predictions from multiple models to improve overall performance, this includes methods like gradient boosting machines (Friedman, 2001), and Random Forests (Breiman, 2001), which are a primary topic of this thesis work.

1.2.3.2 Random Forest Models

The Random Forest (RF) model is a supervised ensemble ML method developed by Breiman (2001) which constructs multiple decision trees during training. Each tree is built from a random subset of the data and features, and the final prediction is made by aggregating the outcomes (i.e., taking the majority vote) of all trees. This approach enhances the stability and accuracy of the model while reducing overfitting. RF models are particularly well-suited for genomic research due to their ability to handle large, complex datasets with many features (genes, microbial taxa, etc.) (Chen & Ishwaran, 2012). RFs are non-parametric, making no assumptions about data distribution, and are robust to outliers and noise typical of genomic data (Miniati et al., 2016).

Historically, one of the most widely cited studies that makes use of RF models for colorectal cancer is the CMS stratification study titled "The consensus molecular subtypes of colorectal cancer" (Guinney et al., 2015). In this study, DNA microarray and RNA-sequencing data were used to develop a Random Forest classifier for colorectal cancer which classified it into four subtypes, now known as the consensus molecular subtypes (CMS).

This CMS classification is integrated into an R package, CMScaller (Eide et al., 2017), and provides a framework for understanding the molecular diversity of CRC by associating each sample with a subtype associated with specific molecular characteristics and clinical implications. **Table 1.1** summarizes the observed pathophysiological differences amongst the CMS subtypes.

Table 1.1

CMS Subtypes and their observed pathophysiological differences.

CMS Subtype	**Molecular Characteristics**	**Clinical Implications**
CMS 1 "MSI Immune"	MSI High, BRAF & TGFBR2 mutations; Immune infiltration & activation	Worse survival after relapse; better response to immunotherapy
CMS 2 "Canonical"	Chromosomal Instability, WNT/MYC activation; APC & TP53 mutations	Good prognosis; responsive to conventional chemotherapy
CMS 3 "Metabolic"	Metabolic dysregulation, KRAS & APC mutations	Variable prognosis; potential metabolic therapy targets
CMS 4 "Mesenchymal"	TGF-β activation, stromal invasion, EMT activation, angiogenesis	Poor prognosis; high recurrence; prone to metastasis

Note. This table is a derivative work based Guinney et al. 2015, with additional information from Inamura 2018 (Creative Commons Attribution License). Permission to create a derivative work from Guinney et al. 2015 was obtained from Springer Nature under license number 5896940821640.

While the CMS classification offers valuable insights into the molecular heterogeneity of CRC, it has recently come under scrutiny for its lack of reproducible results and poor prognostic values (Eide et al., 2021). Multiple samples from the same CRC tumours frequently have different CMS subtype classifications (Mouillet-Richard et al., 2024). In addition, no CMS subtypes consistently correlate with LCC and RCC locations, while other studies have demonstrated that RCC and LCC exhibit distinct molecular and clinical features that impact

prognosis and treatment response (Lee et al., 2017). This discrepancy highlights a potential gap in the CMS subtype classification. Therefore, there is an additional need for research that explicitly focuses and accounts for tumour location to develop more precise classifiers and improve personalized treatment strategies for CRC patients.

Moreover, while Guinney et al. (2015) utilized a RF classifier to develop the CMS subtypes, they provided little detail on the inner workings of their model. Specifically, the study treated the model as a "black box", and did not extensively explain the feature importance scores or the specific genes that contributed most significantly to the classification. This lack of model interpretability makes it challenging to understand the biological drivers of each subtype and limits the ability to translate these findings in clinical practice. Therefore, new studies in the area must not only account for tumour location, but also emphasize the interpretability of the ML model used.

One way to interpret RF models is to use feature importance, a type of scoring metric for each feature which can offer insights into which variables contribute most to the model predictions. The built-in feature importance scoring allows RF models to address some of the black-box issues by highlighting the most relevant features driving the predictions. This makes RF more interpretable compared to other ML methods such as Support Vector Machines (SVM) which relies on complex kernel functions, and deep learning which utilizes complex internal representations, limiting the ability to easily interpret how specific features influence final predictions (Breiman, 2001; Cortes et al., 1995; LeCun et al., 2015).

However, even with these advantages, RF models, like many other ML models, can suffer from overfitting. Overfitting occurs when a model learns not only the underlying patterns in the training data but also the random noise specific to the training set, leading to poor

generalization on new, unseen data (Hastie et al., 2009). Overfitted models perform well on training data but fail to predict accurately on testing or real-world data.

To mitigate overfitting and ensure model robustness, cross-validation techniques such as *k*-fold cross validation (as illustrated in **Figure 1.2**) can be applied. Cross-validation is a statistical method used to evaluate performance of a classifier by partitioning the data into multiple subsets (Hastie et al., 2009). In *k*-fold cross validation, the dataset is divided into *k* subsets (or folds). The model is trained on *k*-1 subsets and tested on the remaining one, and this process is repeated *k* times, with each subset serving as the test set once.

Figure 1.2

A visualization of the flow of testing and training data in 5-fold cross validation.

Note. This figure is a derivative work based on information from scikit-learn (2022) which is licensed under the Copyright BSD License 2007-2024, scikit-learn developers.

This method also allows for the calculation of an estimate of the model's performance on unseen data. Evaluating RF model performance typically employs six common statistical metrics: accuracy, precision, recall (sensitivity), specificity, F1 score, and the area under the

Receiver Operating Characteristic curve (ROC AUC). **Table 1.2** summarizes these metrics and their definitions.

Table 1.2

Common statistical metrics for evaluating performance of machine learning classifiers and their definitions.

Metric	Formula	Description
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	The percentage of overall correct classification.
Precision	$\frac{TP}{TP + FP}$	The proportion of positive cases which were classified correctly out of all positive cases.
Recall (Sensitivity)	$\frac{TP}{TP + FN}$	The proportion of correctly classified positive cases.
Specificity	$\frac{TN}{TN + FP}$	The proportion of correctly classified true negatives in negative cases.
F1 Score	$2 \times \frac{Precision \times Recall}{Precision + Recall}$	Conveys the balance between precision and recall.
ROC AUC (Area Under the Receiver Operating Characteristic Curve)	Calculated by taking the area under the receiver operating characteristic curve using the trapezoidal rule.	An accuracy-like measurement for classifier output quality across all possible classification thresholds.

Note. Where TP: True Positives; TN: True Negatives; FP: False Positives; FN: False Negatives. This table contains information derived from Pellegrino et al. 2021 with permission obtained under a Creative Commons Attribution 4.0 International License, with additional information added from Baratloo et al. 2015.

In developing RF models, the data are typically partitioned into training, validation, and testing sets. Performance metrics **(Table 1.2)** are often recalculated on the testing and/or

validation set many times during model optimization, and then calculated a single time on the testing set for final performance reporting.

1.2.3.3 Features in Random Forests

Improving the explainability of RF models fit to genomic data is crucial for translating the models into actionable biological insights. Feature selection plays a role in enhancing model interpretability by identifying the most relevant features (e.g., genes, proteins, microbes, or metabolites) that are significantly associated with patient outcomes, disease progression, or treatment responses. By focusing on the most important features to the RF model, it is possible to identify features that may have important biological mechanisms.

Feature selection can be categorized into pre-hoc and post-hoc methods, each with unique advantages and challenges. Pre-hoc feature selection involves selecting or reducing features before model training, based on certain criteria, such as variance, correlation with the target variable, or domain knowledge. This approach can help in reducing dimensionality, minimizing noise, and improving model performance, however pre-hoc methods are prone to erroneously excluding important features (Saeys et al., 2007).

In contrast, post-hoc feature selection involves evaluating the importance of features after the model has been trained. This method leverages the model's inherent ability to assess feature importance, allowing for a more informed selection of relevant candidate biomarkers based on their contribution to the model's predictive power. Unlike pre-hoc methods, post-hoc feature-selection methods can capture complex interactions and non-linear relationships between features and the target variable.

Despite the prevalence of pre-hoc feature-selection techniques in machine-learning based genomics research, post-hoc methods, particularly in the form of feature-importance scoring in

RF models can be a powerful alternative for identifying important features without risking erroneously excluding features prior to model building.

1.2.3.4 Feature Importance Scoring in Random Forests

RF models have the ability to provide feature-importance scores, indicating the contribution of each feature to the model's predictive power. A popular choice for feature importance calculation in RF models is based on the Gini impurity, which was first pioneered as a measure of statistical dispersion to quantify income inequality by Corrado Gini (1912), but made popular for use in ML models by Breiman et al. 2001 & 2017. Gini impurity measures the likelihood of an incorrect classification of a new instance if it were randomly labeled according to the distribution of labels in the dataset.

Gini impurity (Breiman et al. 2001, 2017; Homola 2017), at a node *n* is calculated by:

$$I_G(n) = 1 - \sum_{i=1}^{J} (P_i)^2$$

Where:

$I_G(n)$: Gini Impurity at a node (n)

J: Number of classification labels

P_i: Proportion of samples belonging to class i

This is first calculated for the root node, and then recursively for all subsequent child nodes. The further down the tree, the less the total weighted Gini Impurity, and the last layer of nodes has a Gini impurity of 0.

The decrease in Gini impurity when a feature is used to split a node contributes to that feature's importance. The importance ni_j of node j is calculated as:

$$ni_j = w_j C_j - wleft_{(j)} Cleft_{(j)} - Wright_{(j)} Cright_{(j)}$$

Where:

w_j: Weighted number of samples reaching node *j*

C_j: Impurity value of node *j*

$left_j$, $right_j$: Child node from split on node *j*

This formula calculates the importance for each individual node in the tree. This must then be summarized to the level of each feature (i.e., gene or microbe in our dataset) in the tree, as multiple nodes may base their decision on information from the same feature. Individual feature importance can then be calculated by:

$$fi_i = \frac{\sum j{:}node\ j\ splits\ on\ feature\ ini_j}{\Sigma k \epsilon\ all\ nodes\ ni_k}$$

fi_i: The importance of feature *i*

ni_j: The importance of node *j*

ni_k: The importance of node *k*

These values are then scaled to a value between 0 and 1 by dividing by the sum of all feature importance values:

$$normfi_i = \frac{fi_i}{\Sigma j \epsilon\ all\ features\ fi_j}$$

The final task is to calculate the feature's importance over the entire Random Forest. This is calculated as the sum of the normalized features importance in each individual tree, divided by the total number of trees.

$$RFfi_i = \frac{\Sigma j \epsilon\ all\ trees\ normfi_j}{T}$$

Where:

RFf_i: Importance of feature *i* calculated from all trees in the RF model

$normfi_{ij}$: Normalized feature importance for i in tree j

T: Total number of trees

This provides a normalized feature importance score between 0 and 1, indicating the relative importance of each feature across the entire forest. By identifying the most important features and quantifying their relative importance, where these features represent biological inputs (e.g., genes, microbes), researchers can gain insights into their biological relevance at predicting the target (outcome) variable (Breiman et al., 2001, 2017).

1.2.3.5 Permutation Importance

Despite the inherent feature importance scores included with most RF model implementations, there is often a need for more robust approaches to assessing the significance of each feature in the model. Unlike base feature importance scores, permutation importance offers an unbiased estimate of the feature's true contribution to the model's predictive performance. The traditional approach to permutation importance for RF models, as included in many implementations such as scikit-learn, involves randomly shuffling the values of each feature and observing the effect on the model's predictive performance (Pedregosa et al., 2011). This disrupts any existing relationships between the feature and the target variable, but also disrupts the relationships between features themselves. In cases where features may be correlated (such as genomics), this disruption can lead to inappropriate null values of importance scores, thus this type of importance is not suitable for such applications (Altmann et al., 2010).

Rank-based permutation importance approaches are more suitable for cases of complex interactions and correlations between features (Altmann et al., 2010). Rather than shuffling individual feature values, it ranks the features by their importance scores and compares the observed feature at each rank with the distribution of importance scores obtained through

permutations. Existing rank-based permutation implementations tend to evaluate the importance of each feature individually, comparing the observed importance to feature-specific null-distributions. While this is effective in many contexts, this approach may overlook the interactions between correlated features, which is not ideal for genomic datasets. There remains a need for rank-based permutation methods that better account for the interdependencies between features and consider the shared importance of correlated features. In addition to determining feature importance, it is critical to understand how these features interact and contribute to the predictions of the model, in terms of directionality, magnitude, and on an individual and combined basis.

1.2.3.6 SHAP values

SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017) values have emerged as a powerful tool for enhancing the interpretability of ML models. Originating from game theory, SHAP values provide a framework for interpreting ML model output by assigning each feature an importance value for a particular prediction, reflecting both the magnitude and directionality of its effect. In the context of RF models, SHAP values provide global and local interpretability, allowing researchers to understand how each feature contributes to the prediction for individual samples or the sample set. Unlike traditional feature importance metrics, SHAP values offer more detailed insights into the specific influence of each feature which may facilitate the identification of candidate biomarkers in the context of genomics research.

1.2.3.7 Functional Analysis

Understanding the functional roles of identified genes (or candidate biomarkers) is crucial for elucidating their biological significance. Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) is a technique that can assess whether predefined sets of genes show

statistically significant differences between two biological states (e.g. tumour versus normal). This allows researchers to determine whether specific biological pathways or processes are overrepresented among the identified genes.

Gene Ontology (GO) (Ashburner et al., 2000) is another technique for functional annotation, GO is a bioinformatics initiative for the creation of a standardized vocabulary to describe genes and gene products across different species. GO serves as a database resource that provides standardized annotations that classify gene function into three main categories: biological processes, cellular components, and molecular functions. Biological processes describe the pathways and larger processes made up of the activities of multiple gene products, such as cell cycle, signal transduction, or metabolism. Cellular components describe the locations relative to cellular structures in which a gene product performs a function, such as the nucleus, mitochondrion, or cell membrane. Molecular functions are the elemental activities of a gene product at the molecular level, such as binding or catalysis. By integrating GO and GSEA researchers can attribute biological function to a list of candidate biomarkers to attribute them with meaningful biological insights, assess worthiness for future study, and aid in the identification of therapeutic targets and the development of targeted intervention.

1.2.3.8 Biomarker Validation

An ultimate goal of interpreting ML models in a cancer genomics setting is the identification of biomarkers which can inform diagnosis, prognosis, and treatment strategies. Biomarkers are biological indicators that can be objectively measured, such as genes, proteins, or microbes, that can influence specific disease states or outcomes (Strimbu & Tavel, 2010). By leveraging the feature importance scores from RF models, researchers can pinpoint the most influential features that contribute to the prediction of clinical end points (often called targets),

such as tumour location or patient prognosis. Biomarker identification can involve many steps, from identifying features of interest through feature selection, attributing statistical significance, comparisons or rankings with other features, biological validation through cross-referencing with existing literature, functional assays, and experimental studies. The ultimate test of the utility of a biomarker lies in its clinical validation during clinical trials, where its effectiveness and reliability are confirmed in studies on patient populations.

1.3 Overview and Structure of this Thesis

Understanding the complex interplay between host genetics and the gut microbiome is crucial for advancing diagnosis, prognosis, and personalized treatment strategies in CRC. Despite significant advancements in genomic technologies and computational methods, challenges remain in effectively integrating and interpreting multi-omics data. The overarching goal of this thesis is to enhance the analysis and interpretation of genomic and metagenomic data within the context of CRC research by creating interactive applications for exploring results, integrating multi-omics datasets, developing code for efficient machine learning techniques in R, and facilitating biomarker discovery through functional analyses.

As discussed above, while numerous studies have investigated the genomic and microbial aspects of CRC, many have not fully integrated host genomic data with microbiome profiles. Additionally, tumour location, an important factor influencing CRC biology and patient outcomes, was historically overlooked and is still often not adequately considered in existing analyses. Furthermore, there is a lack of user-friendly, integrated tools for exploring complex metagenomic and metatranscriptomic data related to host-microbe interactions in CRC.

An opportunity to address these challenges is presented by our RNA-seq datasets which encompass both host gene expression and microbial profiles from CRC patient samples. By

developing interactive applications for data exploration, and code packages for uncovering candidate biomarkers via explainable RF models, this thesis seeks to facilitate the exploration of such complex datasets and uncover novel biomarkers associated with CRC tumour sidedness.

The main aims of this thesis are to:

Aim 1: Develop an interactive application for genomic and metagenomic data exploration, linking microbial profiles with gene ontology to enhance understanding of host-microbiome interactions in colorectal cancer;

Aim 2: Apply Random Forest Models to identify candidate biomarkers that distinguish right- and left-sided colorectal cancer, and employ rank-based permutation testing to improve model interpretability by quantifying the contributions of individual features;

Aim 3: Develop a code package to facilitate building efficient RF models in R for use in genomic biomarker discovery, while incorporating SHAP analysis and functional gene ontology assessment.

This thesis is structured into five main chapters, each contributing to the objective of enhancing the analysis and interpretation of colorectal cancer genomic and metagenomic data through a combination of software development, discovery and hypothesis generation, and hypothesis testing.

Chapter 2 focuses on software development, detailing the creation of the MetaFunc App, an interactive R Shiny application designed to facilitate the exploration of data generated by the MetaFunc pipeline. This application addresses the need for integrated tools that can analyze microbial abundances, taxonomic profiles, and link them to gene ontology annotations within a single platform. By streamlining the MetaFunc workflow, the MetaFunc App supports researchers in conducting comprehensive studies on host-microbiome interactions.

Chapter 3 involves discovery and hypothesis generation as well as hypothesis testing. By using RF machine learning models, this study aims to identify and explain the contributions of significant genomic and microbial candidate biomarkers that differentiate right-sided and left-sided CRC. This chapter also introduces a rank-based feature permutation test for post-hoc feature selection which is later integrated into the R package introduced in Chapter 4.

Chapter 4 is dedicated to software development and hypothesis testing. This chapter introduces pyRforest, an R package that integrates Python's optimized RF algorithms into the R environment. This code package provides users with an end-to-end workflow for building and analyzing RF models for genomic research. pyRforest features memory-efficient algorithms, advanced feature selection methods, SHAP analysis for model interpretability, and gene ontology analysis for functional assessment.

Chapter 5 summarizes the findings, discusses their implications for genomic research and suggests directions for future work. This includes exploring additional machine learning techniques, integration with additional multi-omics datasets, and further validation of identified candidate biomarkers.

Chapter 2

The MetaFunc App: An R Shiny Application for Joint Exploration of Microbial Diversity, Function, and Host Gene Expression

The MetaFunc App is featured in the following publication (Appendix A):

Sulit, A. K., **Kolisnik, T.**, Frizelle, F. A., Purcell, R., & Schmeier, S. (2023). MetaFunc: taxonomic and functional analyses of high throughput sequencing for microbiomes. *Gut Microbiome, 4*(4), 1-21. https://doi.org/10.1017/gmb.2022.12.

Authors' contributions to the above publication:
Conceptualisation: A.K.S, R.P., S.S.; Formal analysis: A.K.S, F.A.F., R.P., S.S.; Funding Acquisition: R.P.; Investigation: A.K.S., R.P., S.S.; Methodology: A.K.S, **T.K.**, R.P., S.S.; Software: A.K.S, **T.K.**, S.S.; Supervision: R.P., S.S.; Validation: A.K.S, S.S.; Visualisation: A.K.S; Writing – original draft: A.K.S; Writing – review and editing: A.K.S, **T.K.**, F.A.F, R.P., S.S. A.K.S. and S.S. developed and co-wrote the pipeline which ultimately led to MetaFunc, and were involved with the majority of the design. **T.K. developed the shiny application that is integrated in the pipeline.** A.K.S. wrote the manuscript with editorial input from **T.K.**, S.S. and R.P. R.P. further contributed to the design of the pipeline. F.A.F. provided guidance about all clinical aspects of the manuscript.

Author's Note:
While I hold a secondary authorship on the aforementioned manuscript, my contributions to the MetaFunc App are substantial and significant to the MetaFunc project, and include over 2000 lines of code. This chapter focuses on the development of the MetaFunc App, which I led, and presents novel, unpublished content that does not overlap with the Sulit et al. 2023 paper **(Appendix A)**. I have full permissions from all MetaFunc project authors to include this chapter in my thesis work.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Tyler Kolisnik
Name and title of main supervisor:	Dr. Adam Smith, PhD
In which chapter is the manuscript/published work?	Appendix A, with references in Chapter 2

Describe the contribution that the student and members of the supervisory team have made to the manuscript/published work:[1]

Conceptualisation: A.K.S, R.P., S.S.; Formal analysis: A.K.S, F.A.F., R.P., S.S.; Funding Acquisition: R.P.; Investigation: A.K.S., R.P., S.S.; Methodology: A.K.S, T.K., R.P., S.S.; Software: A.K.S, T.K., S.S.; Supervision: R.P., S.S.; Validation: A.K.S, S.S.; Visualisation: A.K.S; Writing – original draft: A.K.S; Writing – review and editing: A.K.S, T.K., F.A.F, R.P., S.S. A.K.S. and S.S. developed and co-wrote the pipeline which ultimately led to MetaFunc, and were involved with the majority of the design. T.K. developed the shiny application that is integrated in the pipeline. A.K.S. wrote the manuscript with editorial input from T.K., S.S. and R.P. R.P. further contributed to the design of the pipeline. F.A.F. provided guidance about all clinical aspects of the manuscript.

Please select one of the following three options:

- [●] **The manuscript/published work is published or in press**

  Please provide the full reference of the research output:

  Sulit, A. K., Kolisnik, T., Frizelle, F. A., Purcell, R., & Schmeier, S. (2023). MetaFunc: taxonomic and functional analyses of high throughput sequencing for microbiomes. Gut Microbiome, 4, e4. https://doi.org/10.1017/gmb.2022.12

- [ ] **The manuscript is currently under review for publication**

  Please provide the name of the journal:

- [ ] **It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:	Date: 2024.09.27	Main supervisor's signature:	Date: 2024.09.27 10:50:38 +12'00'

This form should be placed at the beginning of each relevant thesis chapter.

[1] Refer to the Massey University Publishing and Authorship guidelines (OneMassey for staff, Stream for students) and/ or Contributor Roles Taxonomy (CRediT) guidelines for guidance.

2.1 Introduction

2.1.1 Overview

Metatranscriptomic RNA-sequencing has transformed the study of complex interactions between host organisms and their associated microbiomes and pathogens; yet, the vast scale and complexity of these datasets presents analytical challenges. Extracting meaningful insights from such data requires tools that can integrate microbial and host-gene functions to provide a deeper understanding of biological processes. MetaFunc (Sulit et al., 2023) **(Appendix A)** is a command line-implemented pipeline tool designed to facilitate this process, allowing functional annotation of microbial and host metagenomic and metatranscriptomic data by integrating the processes of protein prediction, functional annotation, and taxonomic classification **(Appendix A)**.

In order to ease the interpretation and exploration of MetaFunc results, we developed a user-friendly R Shiny application (the MetaFunc App). This Shiny app allows interactive exploration of host and microbial abundances, as well as taxonomic and gene ontology tables generated by MetaFunc. The MetaFunc App allows users to pinpoint specific Gene Ontology (GO) (Harris et al., 2004) terms and track the species linked with taxonomic annotations. It also features the ability to invert this process by keying in a species to obtain all associated GO terms. These features are particularly valuable for researchers aiming to understand the functional repertoire of a certain microbial species in the host environment, or compare functional profiles across different species. By linking GO terms to species, researchers can pinpoint specific biological processes within microbial species and track functional traits across them. Conversely, by linking species to GO terms, researchers can explore the functional contributions of individual species across broader phylogenetic contexts.

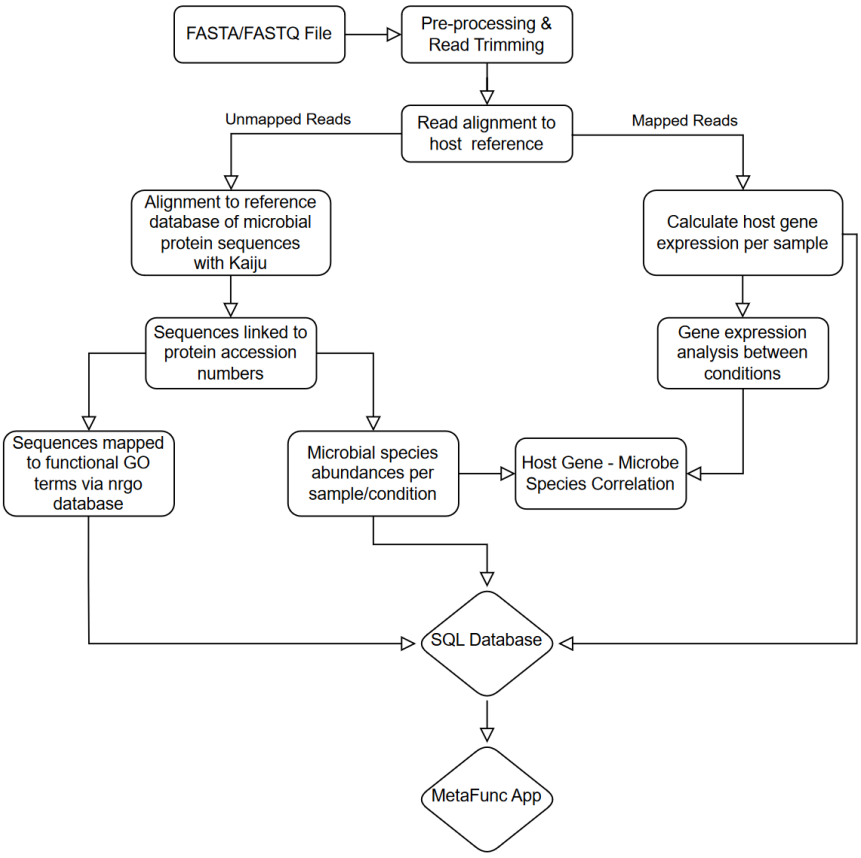
In this chapter, I present the motivation and development of MetaFunc, detail the methodologies used in the development and application of the MetaFunc App, and provide benchmarking results and a case study. Additionally, I discuss the significance of the MetaFunc App, compare it to existing applications, address its scope and limitations, and suggest future directions.

2.1.2 Development History and Specific Contributions

In order to effectively analyze complex metatranscriptomic data, interactive data tools that can integrate multiple data types are crucial. One effective solution is the use of an R shiny application, a web-browser based framework that enables dynamic data displays. Visual inspection of genomic and metagenomic output is useful, including the ability to sort, search, and filter columns (e.g., counts of taxa or functional categories). Previous examples of similar applications that can rank, sort, and rearrange tabular data have proven useful in other contexts, such as the Pavian application (Breitwieser & Salzberg, 2020), which was designed for the exploration of microbial taxa in metagenomic or metabarcoding results. In addition, data exploration via Shiny negates the need for more complicated manipulation steps, for example, sorting or selecting taxa via R.

Figure 2.1.

The MetaFunc Workflow.



Note. Reproduced in part from Sulit et al. 2023 (*includes Kolisnik)* under the Creative Common Attribution License.

Before detailing the design and behaviour of the MetaFunc App, I outline the MetaFunc pipeline itself. Briefly, the MetaFunc pipeline **(Figure 2.1)** begins by aligning reads to the human reference genome using the STAR aligner (Dobin et al., 2013). Human-mapped reads are used for host gene expression and functional analysis, while the remaining human unmapped reads are processed using the metagenomic classifier Kaiju (Menzel et al., 2016) for protein-based taxonomic classification. Kaiju classifies microbial sequences by translating them and matching them with protein sequences and associated accession numbers. These protein accession numbers are subsequently mapped to functional GO (Harris et al., 2004) terms via a custom database

(designated nrgo). The pipeline also calculates microbial species abundances per sample or condition and performs Spearman correlation analysis between host gene expression and microbial species abundances.

The resulting data, including microbial abundances, functional annotations, and host-microbe interactions, are stored in an SQL database for efficient access and retrieval. Users can explore these results interactively with the MetaFunc App, which allows for the visualization of taxonomy and GO tables, facilitating deeper and more detailed investigation into the biological processes, molecular functions, and cellular components present in various species across different datasets and conditions.

2.1.3 Motivation and Research Questions

The motivation for the development of the MetaFunc App was the need for a more intuitive way for the user to sort, search, and interact with the complex taxonomic and functional results generated by the MetaFunc pipeline. The output of this pipeline is more easily interpreted with interactive data displays. The design of the MetaFunc App was centered on fulfilling the following aims:

1. To optimize the accessibility and usability of the complex outputs of MetaFunc for users with varying levels of bioinformatics expertise.
2. To enable users to effectively explore and derive meaningful insights from host and microbial abundances, as well as taxonomic and gene ontology tables produced by the MetaFunc pipeline.
3. To minimise computational overhead and ensure the MetaFunc App remains accessible on platforms with varying compute power and scales to handle large datasets.

In the following section I discuss this development process, including the selection of R Shiny as the development framework, and the methods I implemented to fulfill the above aims.

2.2 Background

2.2.1 Shiny Applications

R Shiny (Chang et al., 2024) was used to create an application to support data display and exploration with integration to the MetaFunc pipeline. Developed by Posit (formerly Rstudio), Shiny was chosen because of its versatility and accessibility, particularly for users without extensive computational backgrounds. It offers an interactive, user-friendly interface, which allows researchers to display and explore their data without needing advanced programming skills in R (R Core Team, 2023) or SQL (Gilmore, 2008). This enables users to focus on deriving meaningful insights from datasets rather than navigating complex computational tasks.

Shiny addresses traditional limitations for data manipulation and analysis steps in R, which typically rely on static tools like ggplot2 and dplyr (Wickham 2016, 2019). Instead, Shiny allows for dynamic interactivity in a browser-based application interface in which users can interact with data in real time using features such as drop-down menus, sliders, and buttons, triggering actions such as recalculating results or refreshing data tables. Shiny integrates a server-side back end with a front-end user interface (UI) that enables interactive data manipulation and display. Additionally, Shiny's server-side processing supports dynamic rendering, enabling the application to display only relevant portions of the data at any given time, improving both speed and performance especially when visualizing large data tables. Shiny apps can be hosted on local or on servers, allowing users to access them remotely without the need for reinstallation or complex configuration.

For more experienced users, Shiny offers a range of advanced functionalities and integration with other tools. For example, it can connect with tools such shinyjs (Attali, 2022), which allows users to store cookies for session management or enhance displays through JavaScript. Shiny also interfaces with R packages such as DBI (R Special Interest Group on Databases (R-SIG-DB) et al., 2024) to manage database connectivity with SQL databases. This versatility has contributed to Shiny's widespread adoption in the development of bioinformatics applications as a powerful framework and useful tool for data analysis.

2.2.2 Shiny Apps in Bioinformatics

Due to its versatility and usefulness, Shiny is used for a number of bioinformatics applications, including the following examples.

shinyGEO (Dumas et al., 2016) imports data directly from the Gene Expression Omnibus (GEO) database to perform differential expression analysis and visualize results. This allows the user to perform complicated biological analyses without having an extensive computational or bioinformatic background.

pcaExplorer (Ludt et al., 2022; Marini & Binder, 2019) helps users perform principal component analysis (PCA) on RNA-seq data to visualize and explain the variance within their dataset, helping to identify patterns and outliers. It provides an interactive graphical interface that allows users to adjust parameters, filter data points, and investigate how specific features contribute to overall variance. explain the variance in their data and identify patterns and outliers. This interactivity enhances data exploration by allowing researchers to assess the relationships between variables and samples, which is particularly useful in the early stages of RNA-seq data analysis. The popularity and utility of this platform is indicated by its number of citations: more than 200 over four years.

shinyGO (Ge et al., 2020) provides users with a tool for gene ontology (GO) enrichment analysis for plants and animals. It provides interactive bar charts of GO terms, enrichment plots for GO-term enrichment, and pathway diagrams that link to KEGG pathways. Gene ontology is useful for associating gene lists to underlying biological functions, cellular components, and molecular components. Again, the utility of the Shiny approach is indicated by the high citation rate of ShinyGO, almost 2,500 since its publication in 2020.

These popular examples illustrate how Shiny apps have enhanced bioinformatics research by making complex analyses more accessible and interactive, and highlight the growing demand for such applications.

2.3 Methodologies

2.3.1 Workflow Integration and Application Launch

The MetaFunc App is integrated with the rest of the MetaFunc pipeline, which allows the app to function as part of a cohesive workflow, rather than as a standalone tool. MetaFunc relies on snakemake (Köster & Rahmann, 2012), a workflow management system that automates the execution of data processing pipelines.

Upon execution, the MetaFunc pipeline automatically generates an SQLite database and launches the MetaFunc App as the final step of workflow, or it can be initiated manually by running the app.R file. This setup accommodates new datasets as they become available; this means that once the pipeline is set up once, any new databases from subsequent pipeline runs will automatically become selectable for viewing within the app, provided they have been correctly placed in the proper directory.

2.3.2 Database Development & Design

The SQL database which serves as the backbone of the MetaFunc App is constructed by the database build script that is integrated into the snakemake workflow. The MetaFunc pipeline results directory consists of a complex list of 290 output files of varying formats. This script locates and processes 8 .tsv files into a unified database structure with the following tables.

Microbial Taxonomy Table: This table catalogs microbial % abundances indexed by taxonomy identifier (TaxID), and is the back end for the Microbiome-Abundances tab in the app. This table contains (1) TaxIDs which are used as a primary key to join with the mapping tables and (2) microbial abundance data for each sample on a per-Tax ID level; (3) the URL for the corresponding NCBI taxonomy browser page for the TaxID.

Gene Ontology Tables: The Microbiome-Gene Ontology tab depends on two separate gene ontology tables, one for storing individual data and one for storing grouped analysis data. The GO individual table stores the following information: (1) The GO ID number; (2) The GO term description; (3) The GO category; (4) the URL of the corresponding AmiGO page for the GO term; (5) the microbial abundance data for each sample on a per-GO term level. The GO grouped table features the same columns, although microbial abundance data is stored for each experimental group.

Mapped Gene Ontology and TaxID Tables: The GO to TaxIDs and TaxIDs to GO tabs depend on two tables, one for individual and one for grouped data. These tables link GO terms to microbial TaxIDs, enabling researchers to trace functional annotations to specific microbial taxa and vice versa. Each table stores the (1) GO ID, (2) the description of the GO term (3) abundances for each sample or group and (4) details on the TaxIDs associated with each GO term. The TaxIDs are mapped to additional information via the mapping tables.

Human Gene Expression Table: This table is the back-end of the Host-Abundances tab, and contains normalized host gene expression data as Transcripts Per Million (TPM) values. The table also includes a URL column that links each gene to its Ensembl entry, allowing users to explore additional gene-specific information.

Mapping Tables: The database includes mapping tables to ensure proper association between taxonomic identifiers, GO identifiers, and sample metadata. The sample mapping table contains metadata for each sample, including (1) sample name, (2) sample type, and (3) internal sample id. The tax id mapping table maps the primary key of TaxIDs to fully parsed taxonomic name data such as Kingdom and lower taxa.

The database is designed using standard database concepts such as referential integrity and normalization, to ensure efficient data management and retrieval. Primary keys such as Tax ID, GO ID and sample names are used to enforce the uniqueness of records, and facilitate rapid lookups and relational joins between tables. With the exception of the mapped GO to Tax ID tables which store some duplicate data to reduce querying time, the database schema adheres to the principles of third normal form, minimizing redundancy and promoting consistency.

2.3.3 Downloading and Installation of the MetaFunc App

The MetaFunc App is distributed as part of the MetaFunc pipeline and can be downloaded directly from the pipeline's GitLab repository at https://gitlab.com/schmeierlab/workflows/metafunc for example using `git clone https://gitlab.com/schmeierlab/workflows/metafunc.git`. The MetaFunc App files can then be found within the /metafunc-shiny/ subdirectory of the repository. This directory contains all necessary files required to run the app.

Users should follow the MetaFunc pipeline installation instructions at: https://metafunc.readthedocs.io/en/latest/usage.html. Additional MetaFunc App instructions can be found at https://metafunc.readthedocs.io/en/latest/rshiny.html#label-shiny.

MetaFunc App dependencies are the following R package: shiny, shinyjs, DT, DBI, formattable, tidyverse, ggplot2, shinythemes, shinydashboard, data.table, dplyr, stringr, reshape2, RSQLite, openxlsx, and splitstackshape. Exact versions are specified in the app .yaml file and are automatically downloaded when the pipeline is installed.

2.3.4 Technical implementation of the MetaFunc App

2.3.4.1 Project Structure

To ensure ease of maintenance and scalability, the MetaFunc App follows best practices for Shiny app development by separating server logic and UI definitions into distinct component scripts. This modular approach supports updates and troubleshooting. The database querying script and initialization variables are organized in separate files, which are integrated into the app through a master script. Databases generated by the MetaFunc pipeline are stored in a subdirectory that is read by the MetaFunc App to populate the internal database selection tool.

2.3.4.2 Front-End Development

The MetaFunc App user-interface (UI) is structured into three main tabs and four subtabs:

The **Home** tab is the initial entry point for the user, and features the database selection and load buttons, usage instructions, and an embedded workflow diagram of the MetaFunc pipeline. A left-justified panel houses the three main tabs and their respective subtabs. The current selected tab is highlighted as a visual indicator for navigation.

The **Microbiome** tab is divided into four subtabs: “Abundances”, “Gene Ontology”, “GO to TaxIDs”, and “TaxIDs to GO”. These subtabs are structured to allow dynamic display of data

tables. Key functionalities include filtering, sorting, and exporting data to Excel, TSV, or CSV formats. The “Abundances” subtab displays a table of microbial abundance data for each sample or sample group. The “Gene Ontology” subtab displays GO data for each sample or sample group. subtabs with GO data feature a toggle to filter by one of three GO categories: Biological Process (BP), Cellular Component (CC), or Molecular Function (MF). The “GO to TaxIDs” and “TaxIDs to GO” subtabs effectively combine the “Abundances” and “Gene Ontology” subtabs, featuring linked tables of abundance and taxonomy data that update dynamically based on user selections. For example, in the “GO to TaxIDs” subtab, the user can select one or more rows of GO information in the top table, and the bottom table will dynamically update to show only the microbial taxa associated with those GO terms. Conversely, in the “TaxIDs to GO” tab users can select one or more rows of taxonomic information in the top table, and the bottom table will update to display the associated GO terms. A reset button is also included to allow users to clear filters and return the tables to their original state.

The **Host** tab includes a single subtab titled “Abundances”, which provides users with data displaying and filtering capabilities for host count (TPM) data. Example screenshots of the user interface can be seen in the `screenshots` folder of the metafunc-shiny gitlab repository.

2.3.4.3 Back-End Development

The back-end development of the MetaFunc App centers around integrating reactive programming and server-side processing to handle the exploration of large datasets. Custom server-side processing functions are responsible for fetching and preprocessing the data from the SQLite database, ensuring the data is formatted consistently for the data tables. Additionally, the app uses shinyjs (Attali, 2022) to store session data for the database path in cookies, allowing the

app to recall which database was previously loaded and automatically reloads it without requiring manual input from the user.

Central to the data table display functionality is the DT R package (Xie et al., 2022), an R wrapper for the JavaScript library DataTables, which facilitates the creation of highly interactive tables. These tables allow the user to sort, search, filter (including setting ranges for numeric values), reorder columns, and paginate the data. Additionally, DT supports embedded HTML URLs, allowing provision of direct links to external resources, such as Ensembl (Martin et al., 2023), NCBI (Sayers et al., 2022), and AmiGO (Gene Ontology Consortium, 2024).

The “GO to TaxIDs” and “TaxIDs to GO” tabs leverage reactive programming to create a dynamic linkage between two interconnected tables. In the “GO to TaxIDs” tab, the upper table displays GO terms, and based on the selected rows, the app triggers a reactive update that fetches and displays the corresponding taxonomic information in the lower table. This functionality is enabled by server-side processing, ensuring that user selections in the first table are instantly passed to the server, which retrieves the relevant taxonomic data and filters and re-renders the second table in real time. Similarly, the “Tax IDs to GO” tab implements the reverse operation.

The MetaFunc App also incorporates CSS-based loading indicators to provide users with visual feedback while data is being processed, allowing users to be aware of background processes to avoid confusion during longer wait times.

Lastly, the app features comprehensive export capabilities, allowing users to download entire tables, or specific rows in CSV, TSV, and XLSX formats. This functionality is achieved through the combination of DT, base R, and openxlsx (Schauberger & Walker, 2024) packages. Custom functions generate a timestamp and append it to the filename to provide users with a unique filename and time record for each export.

2.3.5 Speed and Scalability Optimization

Ensuring scalability without compromising speed was one of the most challenging issues encountered during the development of the MetaFunc App. Initial versions of the app relied on large TSV files for data storage and retrieval, resulting in slow performance as the size of the datasets scaled. Transitioning to an SQLite database considerably improved performance in data joining and retrieval. To handle large datasets more efficiently, table partitioning and pagination were also introduced. Partitioning divided larger tables into smaller, manageable segments allowing only the data necessary for each tab of the app to be loaded, while pagination minimized the memory footprint and client-side rendering time by loading data incrementally.

All calculations and database processing are server-side to offload the computational burden from the client. and dependent only on the hardware of the hosting computer.

Benchmarking tests (in **Results**) were conducted to assess performance with datasets of varying size on a Windows 10 system with 32GB RAM and an AMD Ryzen 9 5900X 12-Core Processor, running at 3.7 GHz with 24 logical processors. The metrics represent an average of five measurements. Response time was recorded using Shiny's showcase mode to track the delay between user input and app reaction. Memory usage was recorded using the profvis R package, and CPU utilization was tracked using the system task manager. CPU usage was monitored using Activity Monitor. Latency, the delay between input and output visualization was recorded using Firefox Developer Tools. Database load times were tracked using a custom wrapper function to measure response times for initial load time of the largest tables in the app, those of the GO to TaxIDs tab. It is important to note that subsequent load times of the tables will always be shorter than the initial load.

2.3.6 Data Availability

Datasets used for the development, validation, and demonstration sets are available at: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA413956 and https://www.ncbi.nlm.nih.gov/bioproject/PRJNA788974.

2.3.7 Testing and Validation

We conducted several structured phases of testing to ensure the functionality and performance of the app. Testing was led by myself and was repeated for validation by the other members of the MetaFunc project. These tests focused on verifying data integrity, app functionality, and performance in the standalone version of the app. Following pipeline integration, additional integration and deployment testing was performed to ensure the app was installable, handled real-world datasets and interacted with other components of the pipeline.

The first phase of testing was installation and deployment testing. In order to ensure the app could reliably be installed on multiple operating system environments, installation was tested on Windows 10 and 11, macOS Mojave and Catalina, and Ubuntu 20.4 LTS systems.

The second phase of testing focused on data integrity checks, with particular attention to identifying errors in data linkage and display. In a worst-case scenario, mislinked or missing samples, GO, or taxonomy information could have led to inaccuracies in the display and skewed functional analysis. This phase involved verifying the accurate linkage and representation of sample ids, GO IDs and Tax IDs, and data values (e.g. abundances, TPM values) within the original TSV files, database, and the app display tables.

The third phase of testing involved functional testing, which aimed to ensure that all user interface elements responded correctly to user inputs. Tests were conducted on every interactive feature, including buttons, row selection, table search, filters, sliders, ability to add or remove

columns. Filtering edge-cases with unusual input values (e.g. negative values, characters) were tested. Functional testing also included checking user experience under different browser environments (Chrome, Firefox) to ensure consistency in app behaviour across platforms.

The final phase of testing focused on performance, specifically measuring how the core functionalities of the MetaFunc App, including database loading, data filtering, and rendering of interactive tables performed under varying database input sizes and user input loads. This was aimed at assessing the ability of the app to handle both typical and extreme data loads (e.g. 2000 samples, 100,000 genes) and inputs. We also tested its resilience by simulating high-intensity user interactions, such as rapidly switching tabs, applying multiple filters, and by pressing multiple buttons at intervals of less than 3 seconds. These tests aimed to identify any bottlenecks or points of failure, including attempts to make the app crash under extreme conditions. We also evaluated stability over extended periods of use (more than one hour) by sending commands at 10-minute intervals, observing CPU usage, memory consumption, and system temperature, to identify any signs of resource exhaustion. This phase provided a comprehensive understanding of the stability of the MetaFunc App during heavy usage scenarios.

These structured testing phases allowed us to systematically identify and resolve issues, thereby allowing us to improve the user experience. Additionally, the MetaFunc App, as part of the broader MetaFunc project, has undergone further validation through peer review processes associated with its publication in Gut Microbiome (Sulit et al., 2023).

2.4 Results

2.4.1 Overview

In this section, we demonstrate the functionality and performance of the MetaFunc App using a step-by-step walkthrough with screenshots. We also present our benchmarking results for performance evaluation of the MetaFunc App.

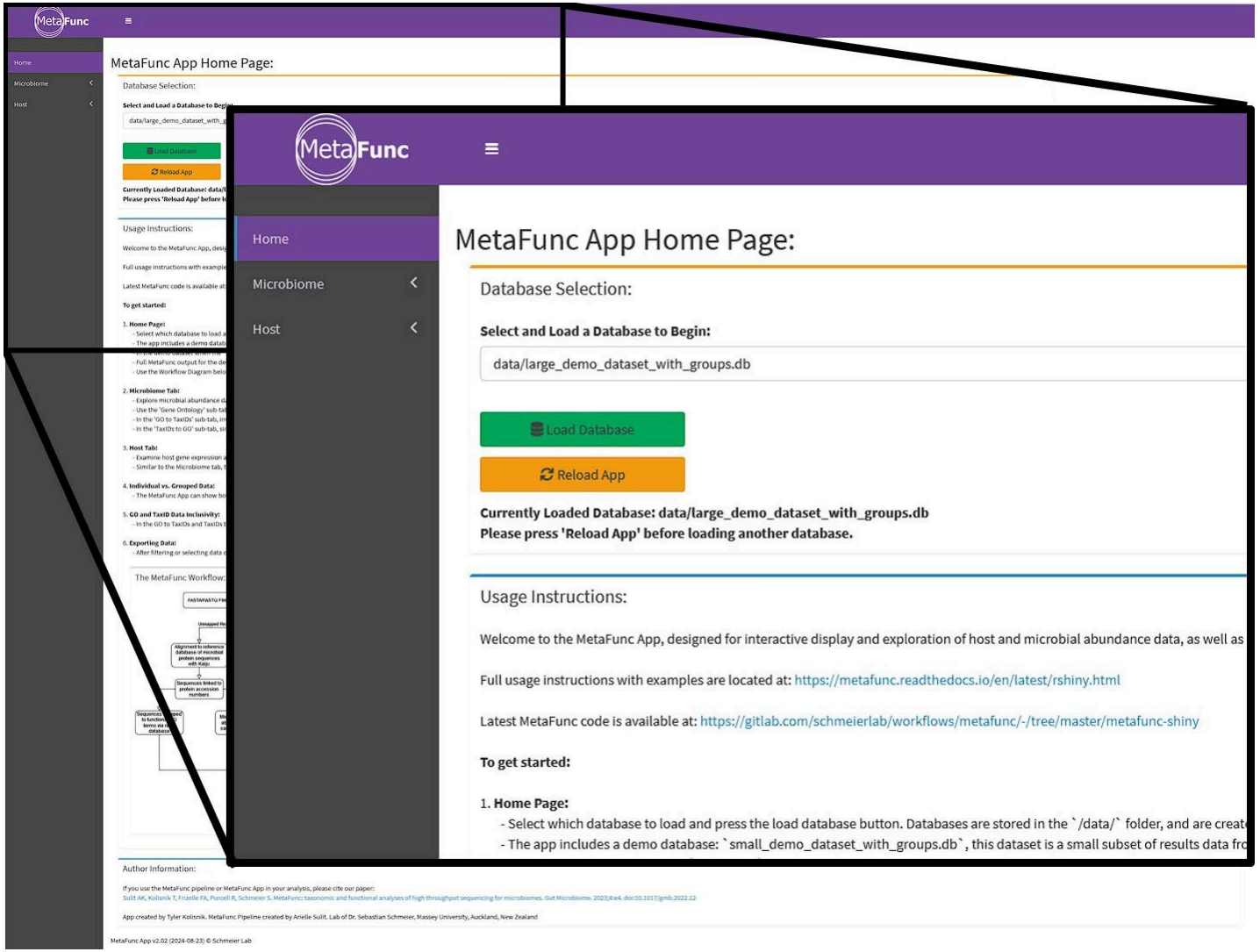
Note on Image Quality: Due to the constraints of the document format, some screenshots included in this section may appear difficult to read. Higher resolution versions of the screenshot images in this chapter can be accessed in the `/screenshots` folder of the metafunc-shiny gitlab repository.

2.4.2 User Interface and Case Study Exploration

2.4.2.1 Home Page

Analysis begins on the Home Page of the MetaFunc App, which provides usage instructions and access to the main functionalities. As shown in **Figure 2.2**, users are directed to select and load a database from the list of databases created by the MetaFunc pipeline.

Figure 2.2 *The Home Page of the MetaFunc App.*

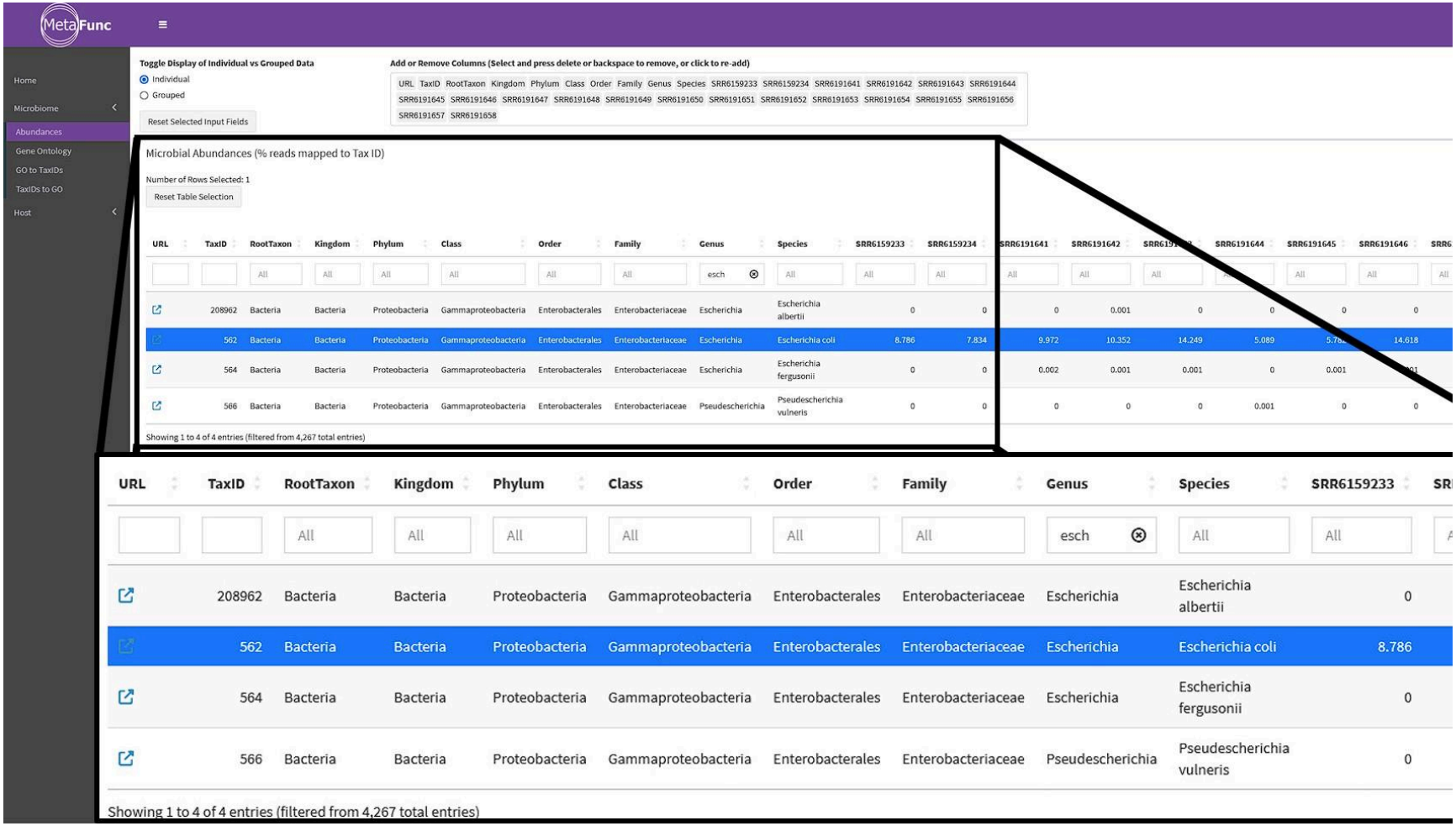


Note. This homepage showcases user navigation options and usage instructions.

2.4.2.2 Microbiome - Abundances

After loading the database, the user is directed by the documentation to choose a tab in the left side panel, “Microbiome” or “Host” depending on what type of data they want to explore. After clicking the “Microbiome” tab, the user is greeted with 4 subtabs; the “Abundances” subtab, illustrated in **Figure 2.3,** indicates the % abundances for each taxon, and can be further filtered, sorted, and searched. In this example the data has been filtered on the ‘Genus’ column by the name ‘*Escherichia*’. In this instance, the user can then identify which *Escherichia* genus microbes are of high or low abundance, and in which samples or sample groups. By highlighting the rows, this data can be exported just for the selected microbes.

Figure 2.3 *The "Microbial - Abundances" subtab of the MetaFunc App.*

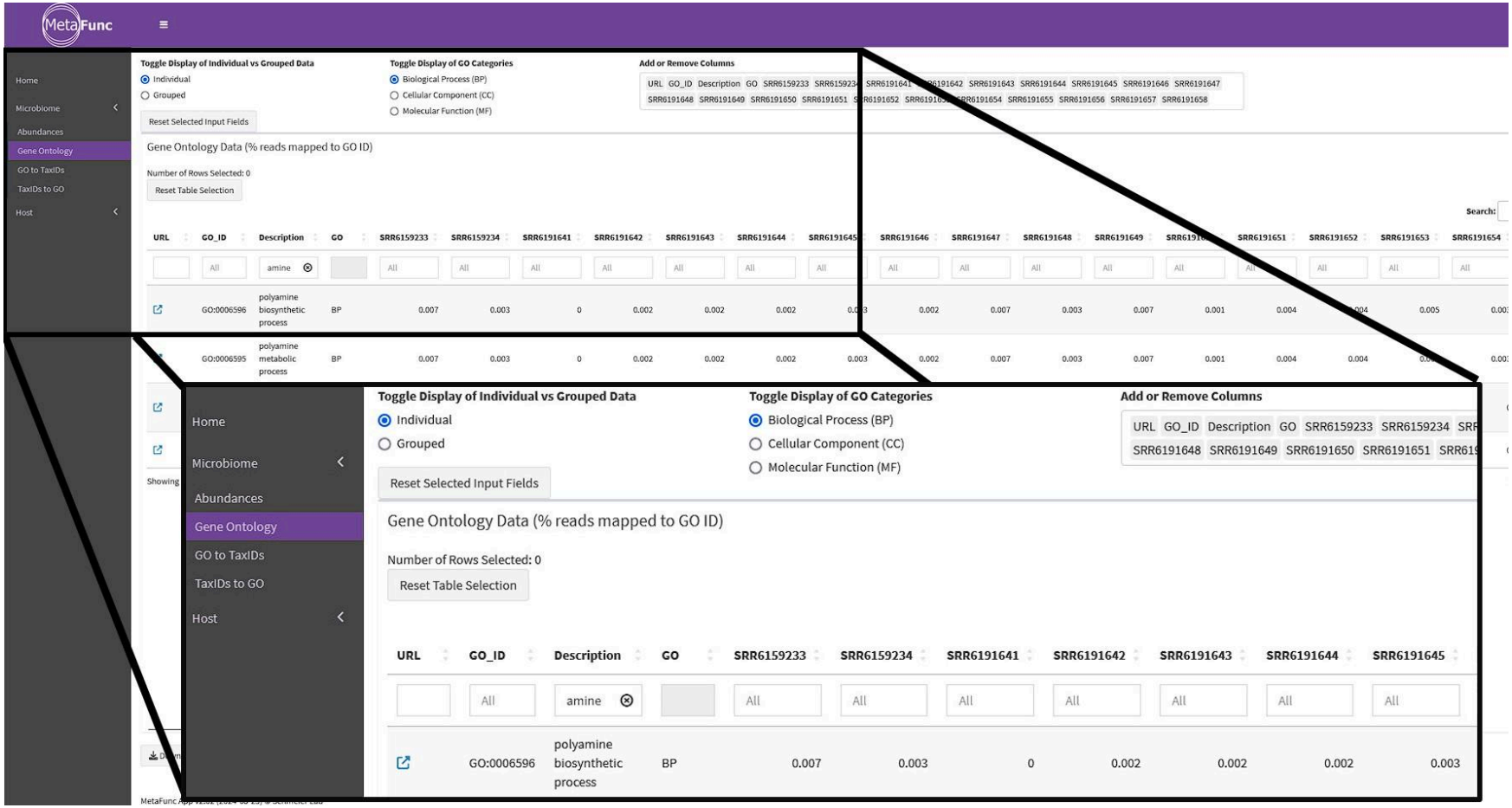


Note. This tab allows users to display and filter microbial abundance data (% read count per taxonomy id) across various samples. The interactive table enables extensive filtering and selection options. The table also includes functionality for resetting filters and exporting the filtered data in .csv or Excel formats.

2.4.2.3 Microbiome - Gene Ontology

If the user wishes to explore GO terms present in their dataset these are accessible via the Gene Ontology subtab, shown in **Figure 2.4**. Users can explore which GO terms have high or low % read counts across one or more samples or sample groups to help with hypothesis generation, or they can quantify abundances associated with predetermined GO terms for hypothesis testing. In this example, the table has been filtered on the 'Description' column by the partial search term 'polyamine', which results in four GO terms. Polyamines are a target of interest in our case study for their known association with tumour progression and cell growth (Sulit et al., 2023).

Figure 2.4 *The "Gene Ontology" subtab of the MetaFunc App.*

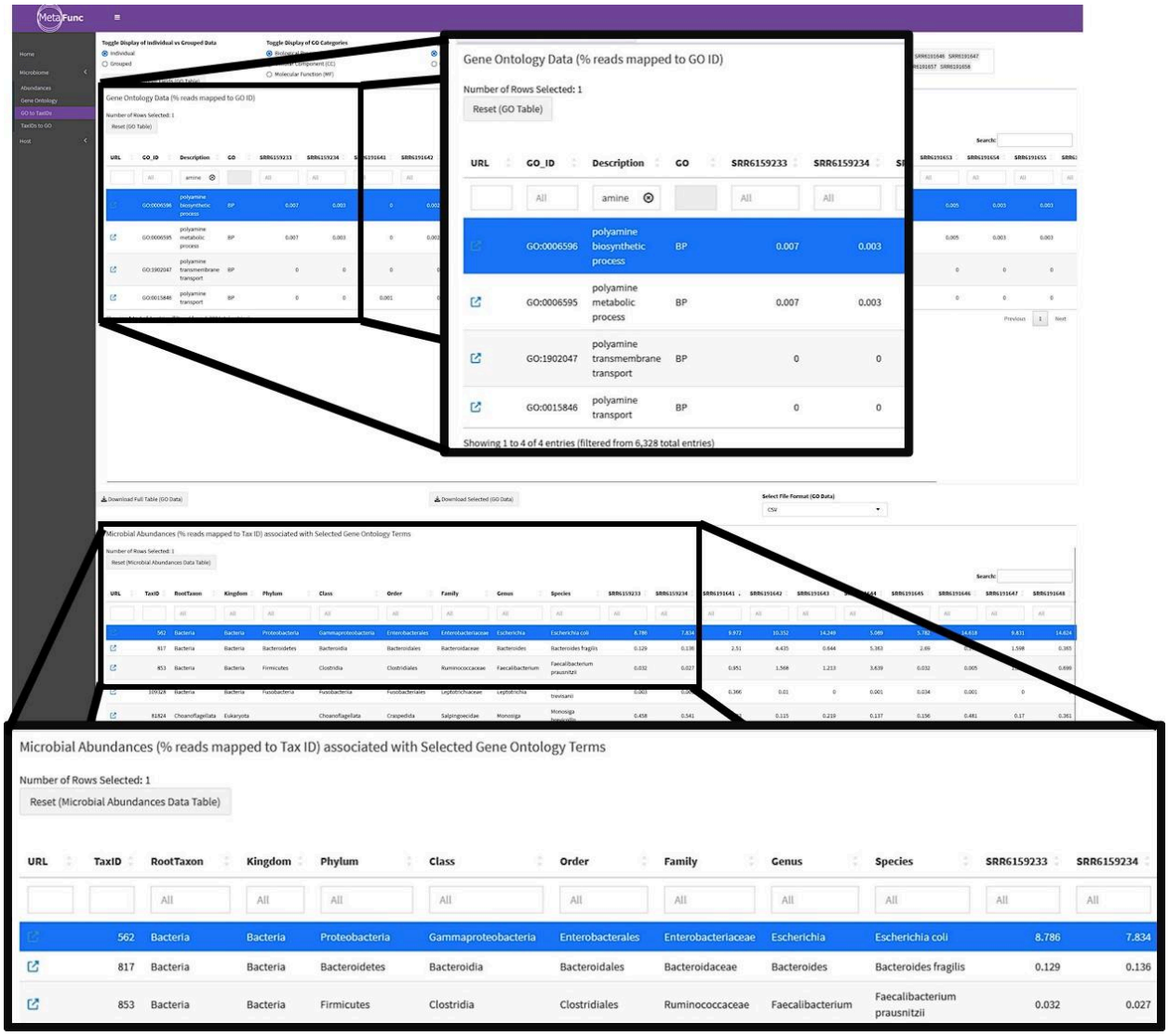


Note. Displayed are GO terms % reads mapped to each GO identifier).

2.4.2.4 Microbiome - GO to TaxIDs

In the next subtab, GO to TaxIDs, **Figure 2.5,** the top table mirrors the gene ontology table shown in the Microbiome - Gene Ontology subtab **(Figure 2.4)**, and the bottom table mirrors the taxonomy table shown in the Microbiome - Abundances subtab **(Figure 2.3)**, except it starts out with no data shown. When the user selects one or more GO identifiers of interest in the top table, the bottom table updates to display corresponding associated taxonomy information and taxonomy abundances. This allows the user to see all microbes associated with a particular GO biological process, cellular component, or molecular function. For example, in **Figure 2.5**, the GO term ‘polyamine biosynthetic process’ has been selected, and the bottom table has populated with the abundance information of 126 microbial taxa associated with this GO term, it has been sorted by highest average abundance to lowest, showing *Escherichia coli* has the highest average abundance of all microbes associated with this GO term.

Figure 2.5 *The "GO to TaxIDs" subtab.*



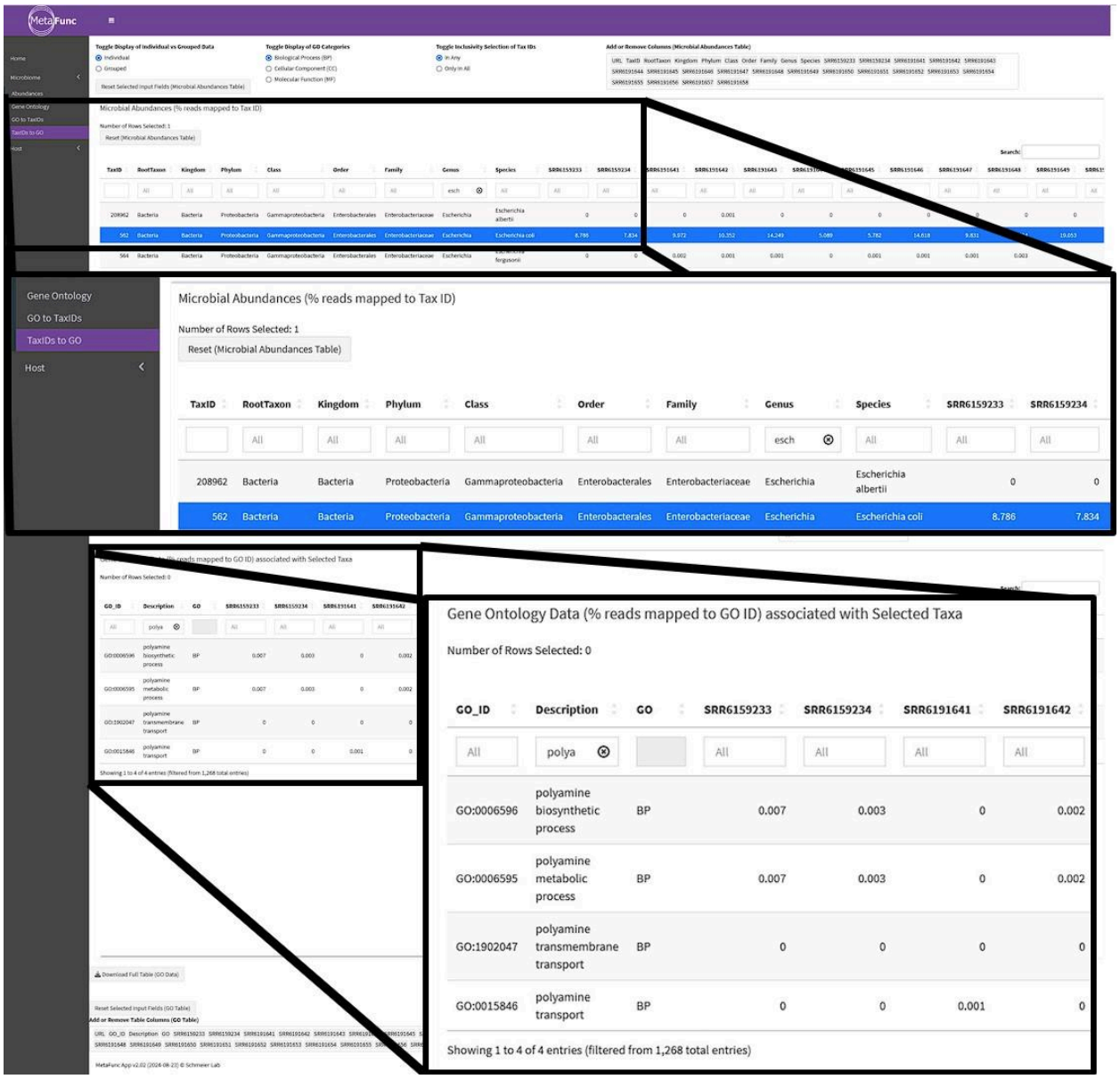
Note. This tab features two interactive tables, where the selection of one or more GO terms in the top table updates the bottom table to display the microbial taxa associated with this GO term.

2.4.2.5 Microbiome - TaxIDs to GO

Similarly, in the TaxIDs to GO tab, shown in **Figure 2.6**, the inverse operation is shown. In this example, the data has been filtered on the genus column for '*Escherichia',* and the microbe *Escherichia coli* has been selected in the top microbial abundances table. The bottom table has rendered the list of 1268 GO biological processes associated with *Escherichia coli*. A filter has

subsequently been applied to the bottom GO table by filtering the description column for ‘polyamine’, showing four related GO terms. The % reads mapped to each GO term for the different polyamine related biological processes associated with *Escherichia coli* are shown.

Figure 2.6 *The “TaxIDs to GO” subtab.*

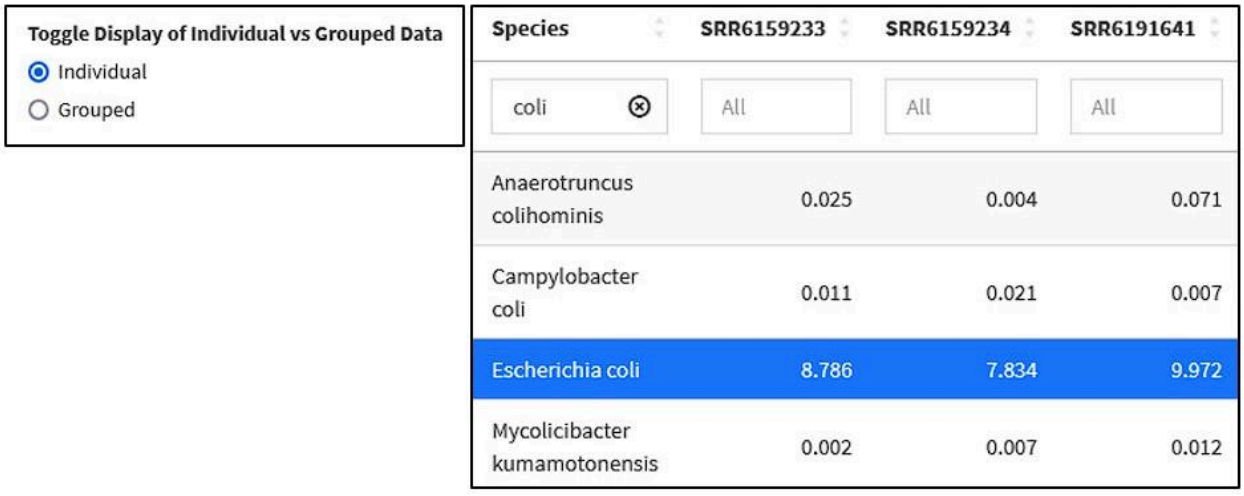


Note. This tab features two interactive tables, where the selection of one or more rows of microbial taxa in the top table updates the bottom table to reflect the GO terms associated with the selected taxa.

2.4.2.6 Single Sample vs Grouped Analysis

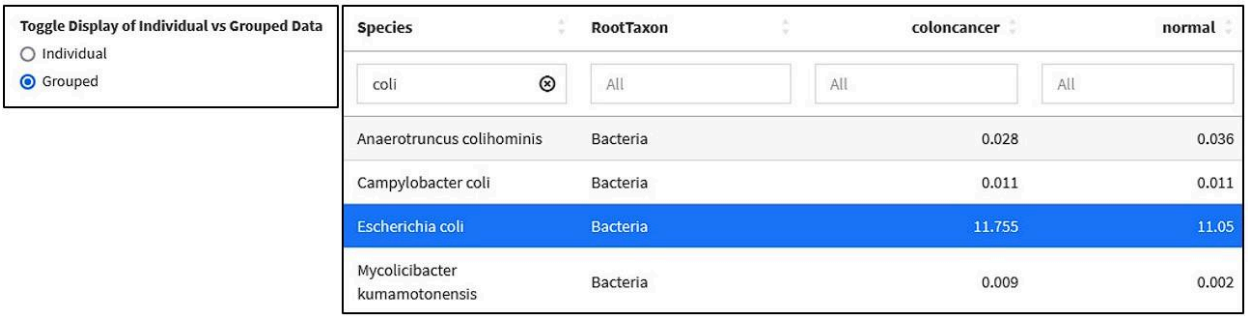
All subtabs under the Microbiome tab of the MetaFunc App can display both individual and grouped analyses, depending on the research objectives. Individual analysis focuses on the exploration of single samples with no additional group context. Grouped analysis allows for the comparison of microbial taxa and their GO terms between predefined groups, such as disease states, treatment types, or molecular subtypes. By comparing **Figures 2.7** and **2.8** below, we can see the differences between the display of individual and grouped analyses on the same table of microbial abundances. Focusing on *Escherichia coli,* when individual analysis is selected this table displays abundances across 3 individual patient identifier columns **(Figure 2.7)** allowing the user to compare the abundance values for each of the 3 patients: (8.876, 7.834, 9.972). When grouped analysis is selected **(Figure 2.8)**, this table displays abundances across the two sample groups, 'coloncancer' and 'normal': (11.755, 11.05).

Figure 2.7 *Individual Microbial Abundances Table.*



Note. Shown is the Microbial Abundances Table, cropped and zoomed featuring 'individual' analysis of 3 patient samples.

Figure 2.8 *Grouped Microbial Abundances Table.*

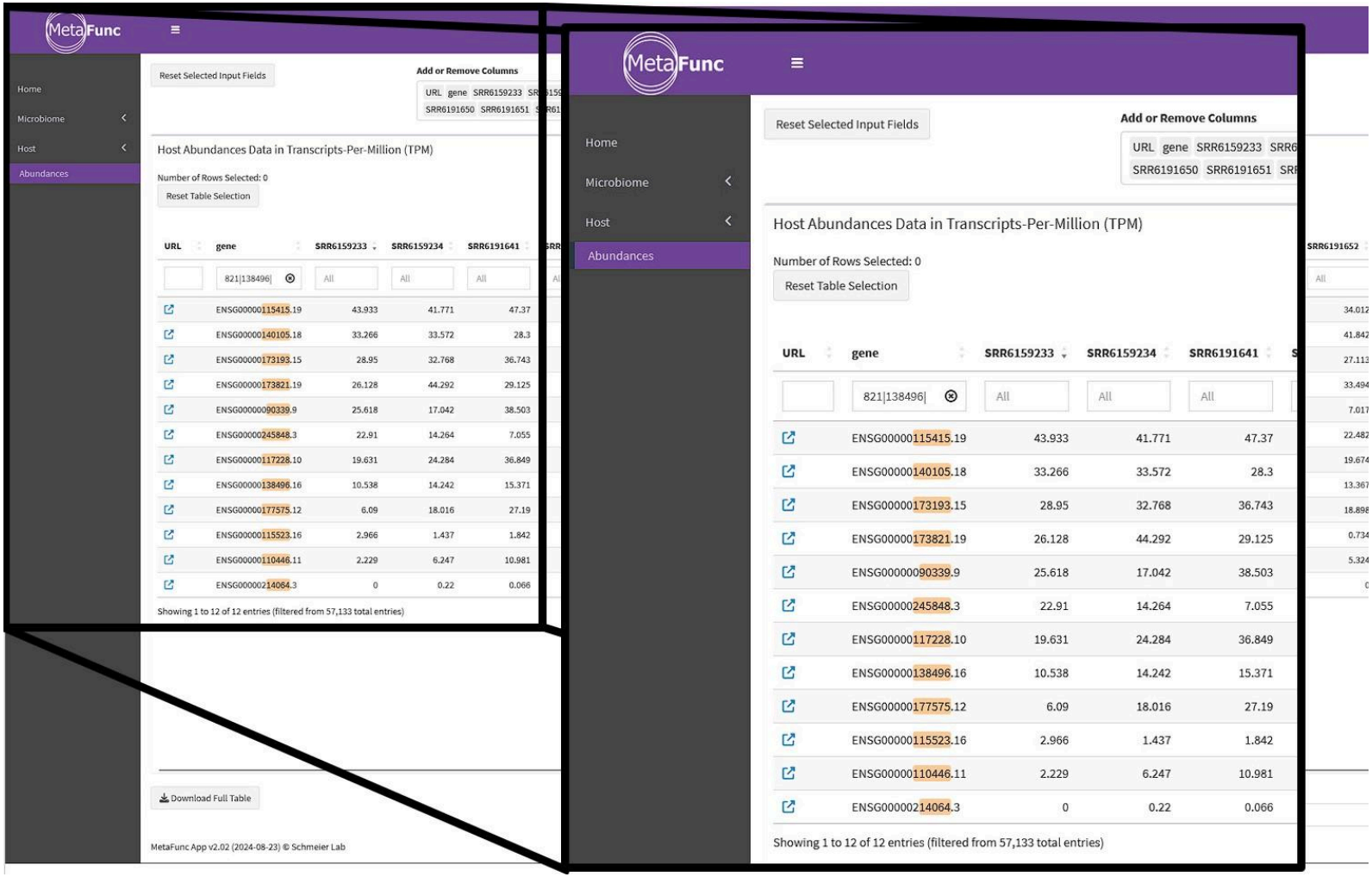


Note. Shown is the Microbial Abundances Table, cropped and zoomed featuring 'grouped' analysis comparing matched colorectal cancer and normal samples.

2.4.2.7 Host - Abundances

Lastly, in **Figure 2.9** the "Host - Abundances" tab is shown which displays TPM normalized human RNA-seq data. In the previous analysis published in Sulit et al. 2023, using the MetaFunc pipeline, we found statistically significant spearman correlations between differentially abundant microbes and differentially expressed host genes for the CMS1 (consensus molecular subtype 1) colorectal cancer subtype for 12 unique genes, these results are presented in **Appendix A Table 2**. In **Figure 2.9**, the "Host - Abundances" tab is shown, where a query limiting the data to 12 genes displays the TPM values for them across 20 patient samples. The user can individually filter for which gene has the highest abundance in each sample, and export the results.

Figure 2.9 *The "Host - Abundances" subtab of the MetaFunc App.*



Note. This figure demonstrates the ability to filter, sort, and search through human RNA-seq TPM data.

2.4.3 Benchmarking Results

Pre-optimization performance was visibly inadequate with long table display times and filtering (greater than 5 minutes), and the app was impractical for real-world use. Detailed benchmarking was only conducted after the optimization efforts described in section 2.3.5. **Table 2.1** summarizes the performance metrics across different dataset sizes, on a computer which meets the minimum requirements for MetaFunc (32GB ram, and a modern high-speed processor).

Table 2.1

Benchmarking results for the MetaFunc App.

Dataset Size (N rows x N samples)	Response Time (s)[1]	Memory Usage[2]	CPU Usage[3]	Latency (ms)[4]	Database Load Time (s)[5]
500 x 20	0.1	205.5 MB	1.20%	45	1
2415 x 308	0.8	1211.7 MB	7.60%	71	3
12,078 x 308	1.4	1645.9 MB	11%	120	19
24,506 x 100	1.1	965.5 MB	8.40%	84	14
24,506 x 308	2.2	2973.1 MB	13%	174	38

Note.
[1] Time to respond to user input
[2] Peak memory usage
[3] CPU usage during interaction across 12 cores
[4] Delay between input and output visualization
[5] Table load time on GO to TaxIDs Tab

These results suggest that increasing the number of samples has a more significant impact on CPU usage, whereas adding more rows primarily increases memory consumption. This observation can be used to guide optimization when scaling the MetaFunc App to accommodate larger datasets. Specifically, we would recommend that users with datasets that have a high number of samples (greater than 300) use a computer with a higher number of CPUs (e.g. 8 or

more), and datasets with many rows (greater than 24,000) to use a higher amount of memory (e.g. 32GB or more). Conversely, for very small datasets (e.g. 500 rows x 20 samples) 1GB of memory and 1 CPU is sufficient.

With adequate computational resources the MetaFunc App can efficiently manage large-scale datasets. The MetaFunc pipeline recommends a minimum of 32GB of RAM and a modern high-speed processor particularly due to the resource-intensive nature of STAR aligner and Kaiju in earlier pipeline steps. These results suggest that given these requirements, the MetaFunc App itself will not introduce additional bottlenecks, and will reliably scale alongside the rest of the pipeline.

2.5 Discussion

2.5.1 Summary

The MetaFunc App, as part of the larger MetaFunc pipeline, provides a comprehensive platform for exploring both taxonomic and functional data in host-microbiome RNA-seq studies. By enabling interactive exploration of metagenomic and metatranscriptomic data, it bridges the gap between complex datasets and actionable insights. The MetaFunc App's ability to link microbial species to functional annotations facilitates the understanding of biological processes, particularly in disease contexts such as cancer.

In the MetaFunc publication (Sulit et al., 2023), which is attached in **Appendix A**, and reiterated in brief in the app walkthrough in our results section, the MetaFunc pipeline along with the MetaFunc App were used to explore and identify candidate microorganisms and their functional contributions, differentiating between CRC samples and normal tissue, as well as among CRC subtypes. The app was used to identify microbial species associated with polyamine

biosynthesis and their contribution to CRC progression, as well as explore the gene ontology terms associated with *Escherichia coli.* The interactive nature of the MetaFunc App eases the computational analysis burden of the end user, and speeds up analysis, and ultimately represents a comprehensive solution to exploring the complex results of the MetaFunc pipeline.

2.5.2 Implications for Data Visualization and Analysis

The MetaFunc App addresses a major challenge in microbiome research, the accessibility and usability of large, complex datasets. By providing a user-friendly interface through the Shiny framework, the MetaFunc App democratizes this complex bioinformatics analysis. The integration of functional annotations and taxonomic data assists researchers with interpreting the biological significance of their findings, particularly in understanding microbial roles in host environments.

While the MetaFunc App is specifically designed to work within the MetaFunc pipeline, its utility in streamlining the exploration of large datasets might serve as an inspiration for the development of similar tools in other domains. Many bioinformatics studies produce large amounts of data that could benefit from interactive app-based exploration, especially as an alternative to static outputs like spreadsheets, which are less effective for managing complex data. By enabling dynamic visual data exploration, the MetaFunc App demonstrates the value of tools that allow researchers to engage more intuitively with their data.

2.5.3 Comparisons with Other Tools

While other bioinformatics tools like Phyloseq (McMurdie & Holmes, 2012) and Animalcules (Zhao et al., 2020) offer interactive analysis of metagenomic data, MetaFunc stands out for its integration of taxonomic and functional analyses in a single platform. Phyloseq for example, is largely focused on phylogenetic analysis and lacks functional annotation capabilities.

Animalcules, however, offers interactive visualizations but does not handle the raw read processing or provide functional insights into gene ontology.

This makes MetaFunc an ideal solution for users interested in both taxonomic and functional analyses, especially for those looking for an all-in-one platform from raw reads to visual app-based dataset exploration. Moreover, MetaFunc's table-based approach, optimized for handling large datasets, provides a practical alternative to Excel-based genomic analyses, which are known to be limited in scalability and prone to errors (Ziemann et al., 2016).

An additional comparison of the MetaFunc pipeline to HUMAnN2 pipeline is provided in the MetaFunc paper **(Appendix A)**.

2.5.4 Limitations and Future Directions

The MetaFunc App inherits some limitations from the MetaFunc pipeline. As the MetaFunc App is not a standalone app, it requires results generated by the MetaFunc pipeline to run, which has a significant user-training overhead and requires significant computational resources. This dependency may limit the accessibility for researchers without infrastructure or expertise to set up and run the full pipeline. Despite optimizations, the app still faces some speed challenges when handling extremely large datasets, particularly for sub-modern systems with low amounts of RAM or CPU cores.

Future enhancements to the MetaFunc App could focus on expanding its functionality and improving scalability. For example, integrating additional external databases for more comprehensive annotations could be useful. Adding plotting and visualization options, such as pathway plots from KEGG (Kanehisa & Goto, 1999) could offer more intuitive ways to visualize abundances. Additionally, optimizing performance by increasing parallelization of backend processes could improve speed, especially when working with large datasets. Further scalability

improvements would ensure that the app remains functional for increasingly larger metagenomic and metatranscriptomic studies.

2.6 Conclusion

The development of the MetaFunc App makes the results of the MetaFunc pipeline more accessible and interpretable. By enabling users to explore metagenomic and metatranscriptomic data in an interactive setting, the app allows researchers to uncover meaningful patterns and relationships that would otherwise be challenging to identify. The MetaFunc pipeline, along with the MetaFunc App, facilitate the analysis of studies that involve complex host-microbe interactions, where taxonomic and functional data must be considered together to fully understand biological context.

The MetaFunc App presents complex results in a clear and customizable manner, making it a valuable tool for both experienced bioinformaticians and bench researchers looking to explore microbial and host data without needing extensive programming skills.

In this chapter, we have discussed the development and methodologies used in the creation of the MetaFunc App, it's importance to the MetaFunc project, the use of Shiny for app development in bioinformatics, and presented a demonstration of the MetaFunc App on a real-world colorectal cancer dataset. In addition, we have also discussed its use in the published study, shown benchmarking results, a comparison with existing tools, and discussed its implications, limitations, and future directions. The MetaFunc App represents a valuable tool for enhancing interpretability of metagenomic and metatranscriptomic data. The MetaFunc App highlights the importance of creating user-friendly bioinformatics applications that help produce meaningful insights from complex data. Chapter 3 continues our exploration of CRC data, shifting the focus to applying machine learning techniques for candidate biomarker discovery.

Chapter 3

Identifying important microbial and genomic biomarkers for differentiating right- versus left-sided colorectal cancer using Random Forest models

The contents of this chapter have been published in BMC Cancer:

Kolisnik, T., Sulit, A. K., Schmeier, S., Frizelle, F., Purcell, R., Smith, A., & Silander, O. (2023). Identifying important microbial and genomic biomarkers for differentiating right- versus left-sided colorectal cancer using random forest models. *BMC Cancer, 23*(647), 1-11. https://doi.org/10.1186/s12885-023-10848-9.

Authors' contributions:
TK conceived the project, and designed and implemented the code for data curation, machine learning, hyperparameter selection, feature reduction, as well as analysis of results and preparation of figures and wrote the first draft and contributed to all versions of the manuscript.
AKS mapped the RNA-seq data to the human genome and mapped the human unmapped reads to produce the microbial counts.
AS guided the development of statistical methods and contributed to editing the manuscript.
OS provided support with idea generation, troubleshooting, statistics, and critical analysis of results, and contributed to editing the manuscript.
RP, FF provided sample collection, processing, RNA sequencing and metadata collection, and minor manuscript revisions.
SS provided initial insight to the formation of the machine learning pipeline, as well as assistance with conceptualization and data normalization.
All authors have read and approved this manuscript.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Tyler Kolisnik
Name and title of main supervisor:	Dr. Adam Smith, PhD
In which chapter is the manuscript/published work?	3

Describe the contribution that the student and members of the supervisory team have made to the manuscript/published work:[1]

TK conceived the project, and designed and implemented the code for data curation, machine learning, hyperparameter selection, feature reduction, as well as analysis of results and preparation of figures and wrote the first draft and contributed to all versions of the manuscript. AKS aligned reads to genome. AS guided statistical methods and contributed to editing the manuscript. OS provided supported conceptualization, troubleshooting, statistics, and analysis of results, and contributed to editing the manuscript. RP, FF provided sample collection, processing, RNA sequencing and metadata collection, and minor manuscript revisions. SS provided conceptualization and data normalization.

Please select one of the following three options:

- [●] **The manuscript/published work is published or in press**

  Please provide the full reference of the research output:

  Kolisnik, T., Sulit, A. K., Schmeier, S., Frizelle, F., Purcell, R., Smith, A., & Silander, O. (2023). Identifying important microbial and genomic biomarkers for differentiating right- versus left-sided colorectal cancer using random forest models. BMC Cancer, 23(1), 647. https://doi.org/10.1186/s12885-023-10848-9

- [ ] **The manuscript is currently under review for publication**

  Please provide the name of the journal:

- [ ] **It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:	Date: 2024.10.26	Main supervisor's signature:	Date: 2024.09.27 10:50:06 +12'00'

This form should be placed at the beginning of each relevant thesis chapter.

[1] Refer to the Massey University Publishing and Authorship guidelines (OneMassey for staff, Stream for students) and/ or Contributor Roles Taxonomy (CRediT) guidelines for guidance.

3.1 Abstract

Background:

Colorectal cancer (CRC) is a heterogeneous disease, with subtypes that have different clinical behaviours and subsequent prognoses. There is a growing body of evidence suggesting that right-sided colorectal cancer (RCC) and left-sided colorectal cancer (LCC) also differ in treatment success and patient outcomes. Biomarkers that differentiate between RCC and LCC are not well-established. Here, we apply Random Forest (RF) machine learning methods to identify genomic or microbial biomarkers that differentiate RCC and LCC.

Methods:

RNA-seq expression data for 58,677 coding and non-coding human genes and count data for 28,557 human unmapped reads were obtained from 308 patient CRC tumour samples. We created three RF models for datasets of human genes-only, microbes-only, and genes-and-microbes combined. We used a permutation test to identify features of significant importance. Finally, we used differential expression (DE) and paired Wilcoxon-rank sum tests to associate features with a particular side.

Results:

RF model accuracy scores were 90%, 70%, and 87% with area under curve (AUC) of 0.9, 0.76, and 0.89 for the human genomic, microbial, and combined feature sets, respectively. 15 features were identified as significant in the model of genes-only, 54 microbes in the model of microbes-only, and 28 genes and 18 microbes in the model with genes-and-microbes combined. *PRAC1* expression was the most important feature for differentiating RCC and LCC in the genes-only model, with *HOXB13, SPAG16, HOXC4,* and *RNLS* also playing a role. *Ruminococcus gnavus* and *Clostridium acetireducens* were the most important in the

microbial-only model. *MYOM3, HOXC4, Coprococcus eutactus, PRAC1, lncRNA AC012531.25, Ruminococcus gnavus, RNLS, HOXC6, SPAG16* and *Fusobacterium nucleatum* were most important in the combined model.

Conclusions:

Many of the identified genes and microbes among all models have previously established associations with CRC. However, the ability of RF models to account for inter-feature relationships within the underlying decision trees may yield a more sensitive and biologically interconnected set of genomic and microbial biomarkers.

3.2 Introduction

3.2.1 Background

Despite being part of the same organ, colorectal cancer tumours can have different pathogenicity, histology and patient outcomes depending on subtype (Fontana et al., 2019) and which side of the splenic flexure they occur (Yang et al., 2016). Left-sided colorectal cancer (LCC, or distal colorectal cancer) affects the rectum, sigmoid colon, descending colon, and distal one-third of the transverse colon. It is generally more common in men, diagnosed at an earlier stage, more responsive to treatment, and patients exhibit a higher rate of survival (Stintzing et al., 2017). Right-sided colorectal cancer (RCC, or proximal colorectal cancer) affects the proximal two-thirds of the transverse colon, ascending colon, and caecum (Stintzing et al., 2017). It is generally more common in women, less responsive to existing treatments, and has poorer outcomes (Yang et al., 2016). Numerous studies have reported vast differences between LCC and RCC in terms of diagnostics, prognostics, histology, epidemiology, pathology, treatment response, and survival (Baran et al., 2018; Bergen et al., 2021; Zhao et al., 2020). Among other

things, these differences suggest that LCC and RCC should be distinguished when developing new treatment regimens and therapeutic drugs (Nagai et al., 2021; Narayanan et al., 2018).

Gut microbiota has been shown to play an influential role in CRC carcinogenesis and progression. However, the mechanisms by which this occurs largely remains unknown (Sánchez-Alcoholado et al., 2020). In addition to cancer progression, it has also been postulated that the gut microbiome may affect gene expression and downstream patient treatment responses (Cercek et al., 2022). To test these hypotheses, there is a need for studies that explore the influence of the gut microbiome on the genomic expression inside colorectal cancer tumour cells.

Machine learning (ML) methods are frequently applied for classification in tasks that rely on high-dimensional genomic data. Here, to query the relationships between the expression of genomic features in CRC and microbial content, we use Random Forest (RF) classification (Breiman, 2001). We selected RF as it can account for interactions and correlations among large numbers of features (Chen & Ishwaran, 2012). Furthermore, RF models do not require normalization or scaling, which makes it possible to combine completely different types of data, for example, microbial count data and RNA-seq datasets.

Here, we explore the ability of RF models to predict CRC sidedness using three different datasets: human genomic feature expression level (RNA-seq), microbial count data (from unmapped human reads), and a combined genomic feature and microbial count dataset. We subsequently use differential expression (DE) analysis on the most important features of the RF model (i.e., biomarkers) to find differential genomic and microbial features and relationships between RCC and LCC. Finally, we discuss the possible biological mechanisms driving differences in these biomarkers.

3.3 Methods & Materials

3.3.1 Patients, Samples and Processing

308 colorectal cancer tumour samples were obtained from patients via surgical resection (partial colectomy). Patients with inherited CRC and those who had received preoperative chemotherapy or radiotherapy were excluded. Patients were over the age of 18 and gave written informed consent. Tumour tissue was obtained between January 2002 and January 2016, with a median tumour tissue date of August 2006. The study was approved by the University of Otago, New Zealand, Human Research Ethics Committee (approval number: H16/037). Patient and clinical data, including anatomical site of tumour, in addition to genomic and microbial data profiles were available for all patients **(Table 3.1)**. Samples were snap frozen in liquid nitrogen at time of surgery and stored at -80°C and transitioned for RNA Extraction using RNAlater®-ICE. RNA was then extracted using the QIAGEN RNAEasy mini kit and sequenced using Illumina HiSeq machines (2x125bp PE v4 sequencing). The samples were machine-randomized to limit any machine-specific noise or calibration bias. Raw Sequence Reads are available at SRA Accession: PRJNA788974.

Sequence reads were first mapped to the human genome (GRCh38) using STAR (v2.73a). The remaining unmapped reads were classified using Kaiju (v1.6.2) to obtain microbial abundances (Dobin et al., 2013; Menzel et al., 2016). Raw genomic reads were TPM (Transcripts Per kilobase Million) normalised prior to data analysis to remove gene length and sequencing depth biases. Microbial abundances were CPM (Counts Per Million) normalized.

3.3.2 Random Forest Model Generation

The RF models were built on the following training datasets: the first contained 58,677 TPM normalized genomic features, the second contained CPM normalized microbial counts for

28,557 taxa, and the third contained a combination of both. A separate validation cohort of 30 samples (15 RCC, 15 LCC) was held out from model generation, leaving 278 patient samples for model development. Genomic and microbial data was available for all 308 patients.

The RF models were parametrized in parallel on high-powered cluster computing nodes with 8,136 cores in 226 × Broadwell nodes, and a total system memory of 31 TB.

Table 3.1

Patient Demographics & Cancer Characteristics.

Characteristic	Value
Patients enrolled - no (%)	**308 (100)**
Median Age - year (range)	**73 (28-91)**
Sex	
Female - no (%)	**163 (53)**
Male - no (%)	**145 (47)**
Cancer Anatomical Side	
Left - no (%)	**172 (56)**
Right - no (%)	**136 (44)**
Metastasis	
Positive - no (%)	**70 (23)**
Negative - no (%)	**238 (77)**
Ethnicity (Self-Reported)	
European - no (%)	**296 (96)**
Māori - no (%)	**9 (3)**
Asian - no (%)	**3 (1)**
Cancer Stage	
T1 - no (%)	**53 (17)**
T2 - no (%)	**128 (42)**

T3 - no (%)	**105 (34)**
T4 - no (%)	**22 (7)**
Nodal Status	
Positive - no (%)	**185 (60)**

RF models were generated using the Python-based scikit-learn Random Forest module (Mölder et al., 2021; Pedregosa et al., 2011). Model hyperparameters were optimized independently for all three training datasets using a grid search with 5-fold cross validation (scikit-learn package GridSearchCV) (Pedregosa et al., 2011). To narrow down suitable hyperparameter sets, the influence of 8 hyperparameters on F1 scores (the weighted average of precision and recall) were each independently observed across a typical range of values for each, while holding the other hyperparameters to their default values **(Supplementary Figures 1-3)**. GridSearchCV with 5-fold cross validation was then used on the smaller set of hyperparameter combinations on the training datasets. A final set of model parameters was selected for each dataset based on highest performing receiver-operator characteristic area-under the curve score (AUROC score), accuracy (total correctly classified cases), and F1 score. For cases in which the performance scores were identical, the model with the fewest features was selected.

Each model was trained using the finalized set of parameters using a 75% train, 25% test split on the dataset of tumour samples from the 278 different patients. The model metrics of accuracy, out-of-bag score, F1 score, ROC AUC score, recall, and precision for each of the three models are reported in **Table 3.2**. Overfitting was assessed by comparing model accuracy with out-of-bag score (number of correct predictions in the out-of-bag sample) and accuracy of the validation cohort. If the model accuracy differed from the out-of-bag score by 0.1 or more, we inferred that there was a strong likelihood overfitting had occurred. A threshold analysis was also

performed for each model, but we found that all optimised thresholds were within 10% of the default value, so we used the default threshold value (0.5). The model was then validated on our validation cohort of 30 samples, this is sometimes referred to as the testing set, and is independent from the testing data used in model training. ROC curves were generated for all three models using the Python package seaborn (Waskom, 2021).

3.3.3 Feature Importance and Retention

The feature importance scores (Gini impurity values) were extracted from each of our three RF models. Given that the models have large numbers of features (greater than 50), it is prudent to perform feature reduction such that only features with high importance (weight) and a high degree of statistical evidence are retained. Using the R package Rf2pval (Kolisnik, 2022), which has since been integrated into the R package pyRforest (Kolisnik et al., 2024), we implemented a rank-based permutation approach to obtain distributions of feature-importance scores at each rank under a null hypothesis where none of the features are associated with the target variable, and assign p-values to the features. We generated 100 randomized datasets in which the target variable ('side') was permuted, retrained the RF models on each, and obtained feature importance scores and scoring metrics for each permuted model. We retained only features with p-values less than 0.05 **(Fig 2 a-c)**. A threshold for feature reduction was identified using the overlap of feature importance scores from the true model with the mean of the permuted feature importance scores **(Tables 1-3)**. Family-wise error rate (FWER) via resampling was used for measuring the probability of making one or more false discoveries during multiple-testing and was calculated using the Rf2pval package for all three models to be FWER<0.05, or less than 5% chance of our features listed above our threshold being incorrectly identified.

3.3.4 Differential Expression, Feature Side-Assignment and Heatmap Generation

DE analysis was performed on each of the feature lists from the three models using edgeR (Robinson et al., 2010). Wilcoxon-rank sum tests were used to calculate p-values to test for DE of each model's features. Heatmaps were generated for assessing feature clustering compared with the clinical labels of cancer stage, metastasis, subtype, gender site and side using the function heatmap.2 in the R package gplots (Warnes et al., 2009). The heatmap implements row-scaled z-scores of the transcripts per million (TPM) read counts, with hierarchical clustering using Pearson distance correlation, and average-linkage distance.

3.4 Results

3.4.1 Random Forest Model Performance

We found that the Random Forest models from all three datasets clearly differentiated between LCC and RCC. Model accuracy on the validation sets ranged from 0.7 to 0.9, with genomic features having an accuracy of 0.94 and 0.9 on the training and validations sets, respectively microbial counts having an accuracy of 0.76 and 0.7, and genomic features with microbial counts having an accuracy of 0.8 and 0.87 **(Figure 3.1)**. Out-of-bag scores were 0.73, 0.74 and 0.74 for the three datasets. The strongest predictors between the LCC and RCC were genomic features, although classifications based on microbial count differences were also consistent. We found 15 statistically significant features in the genomic feature-only model, 54 in the microbial counts-only model, and 46 in the genomic features with microbial counts model **(Fig 3.2 A-C, Tables 3.3-3.5)**.

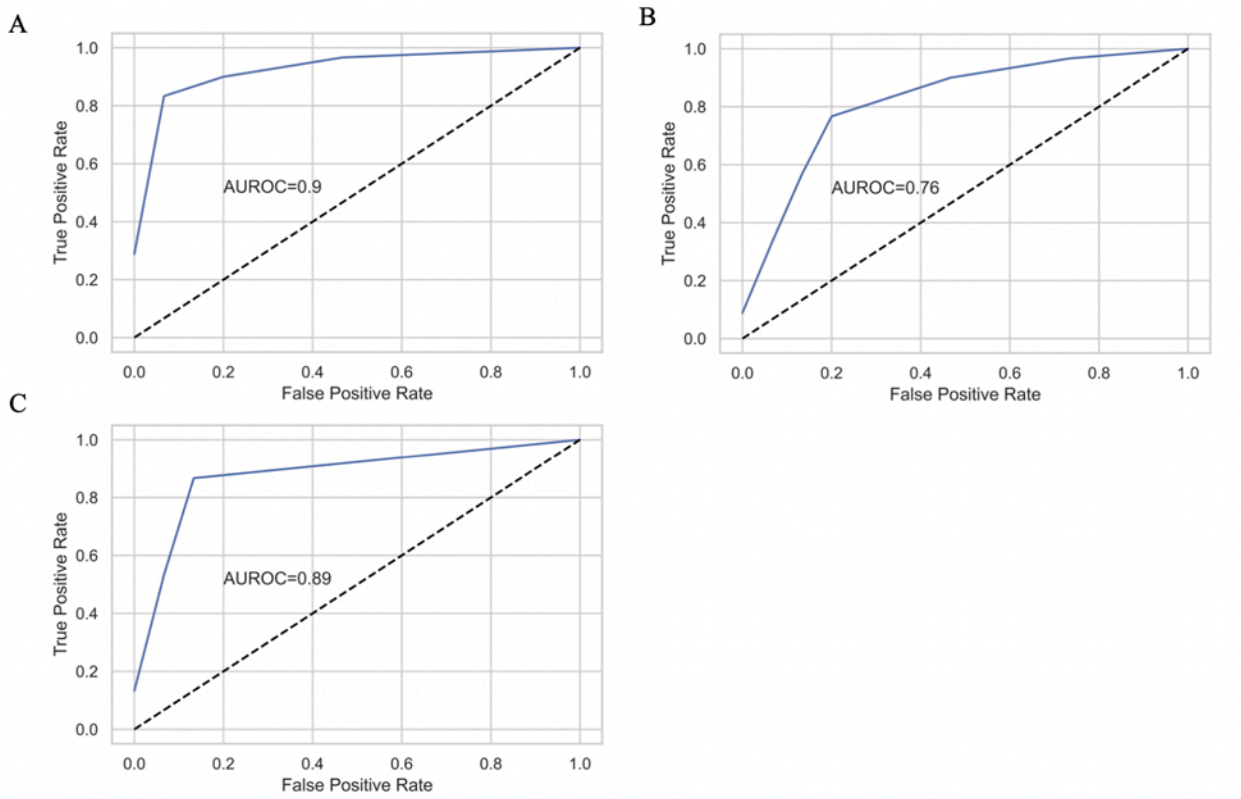
Table 3.2

Random Forest Model Results.

Scoring Metric	**Model**		
Testing Set (5-Fold CV)	**Genes-Only**	**Microbes-Only**	**Genes-and-Microbes**
Accuracy	0.94	0.76	0.8
Out-of-Bag Score	0.73	0.74	0.74
F1 Score	0.95	0.79	0.84
ROC AUC Score	0.94	0.75	0.78
Recall Score	0.95	0.8	0.93
Precision Score	0.95	0.78	0.77
Validation Set (30 held-out samples)			
Accuracy	0.9	0.7	0.87
F1 Score	0.9	0.76	0.88
ROC AUC Score	0.9	0.76	0.89
Recall Score	0.93	0.64	0.79
Precision Score	0.87	0.93	1

Figure 3.1

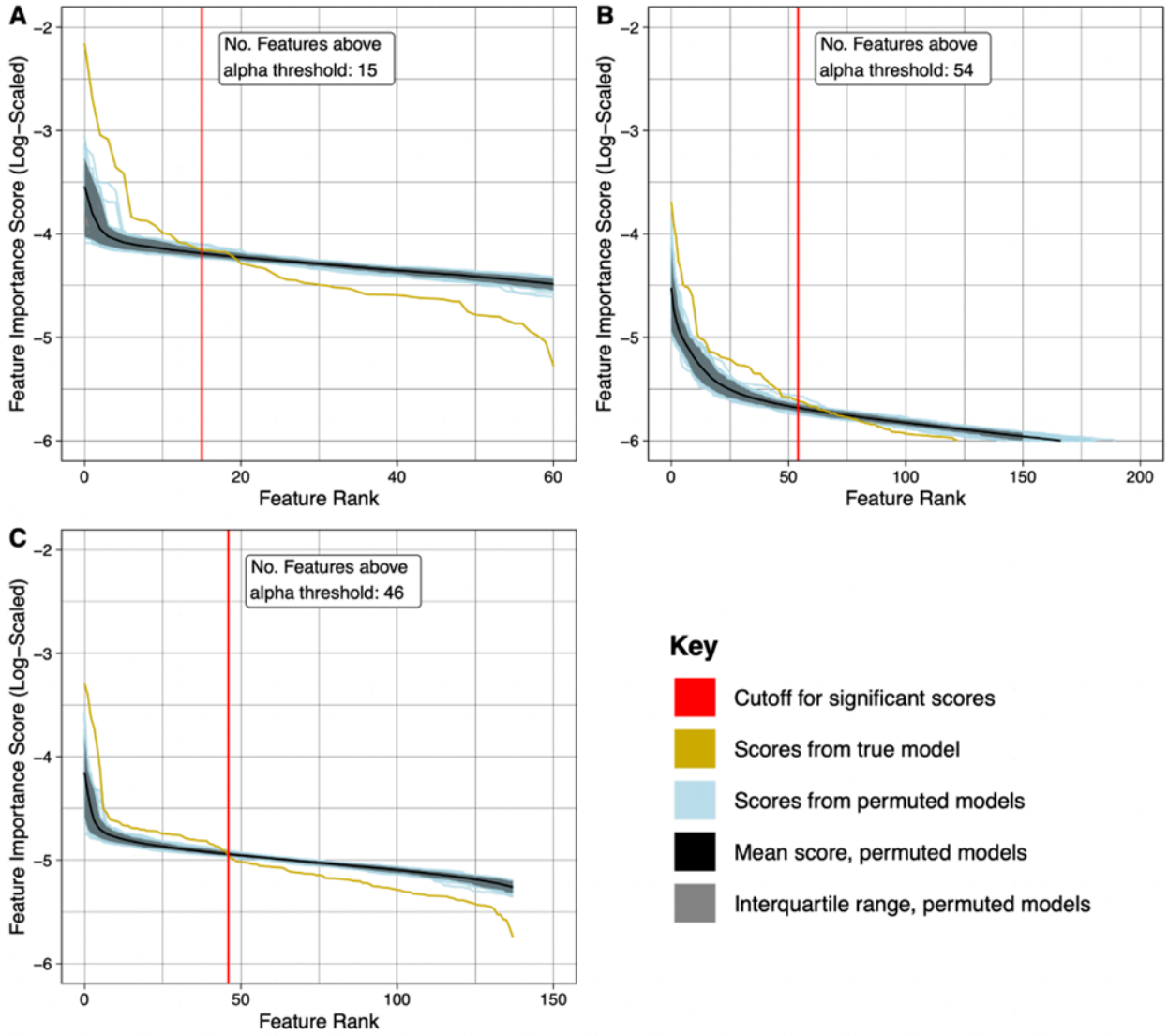
Receiver Operating Characteristic Curves (ROC) as calculated on the held-out validation set.



Note. **a** ROC curve of the genes-only model. **b** ROC curve of the microbial-only model. **c** ROC curve of the genes-and-microbes model.

Figure 3.2

Feature importance plots showing rank-based feature importance scores of the permuted data and the scores of the real (unpermuted) data.



Note. The cutoff for features reported as significant was determined based on an alpha threshold of 0.05, and are to the left of the vertical red line. **a** genes-only model. **b** microbes-only model. **c** shows the genes-and-microbes model.

3.4.2 Significant Model Features

Of the 15 significant features in our gene-only dataset, **(Fig. 3.2A, Table 3.3)** the highest importance score was for the Prostate Cancer Susceptibility Candidate 1 (*PRAC1*) gene, which

has higher expression in LCC. Other significantly important features included those in the *HOX* family of genes, *HOXB13, HOXC4, HOXC6* and *HOXC8.*

Table 3.3

Top ranking features from the RF model trained on the genes-only dataset (Left).

Model Feature Importance Metrics					Differential Expression			
Rank	**Ensemble ID_Gene ID**	**Importance Score**	**Log Importance Score**	**p-value**	**Log2 FC**	**p-value**	**FDR**	**Greater Expr. Side**
1	ENSG00000159182_PRAC1	0.12	-2.16	0	-2.86	5.08E-21	3.10E-19	Left
2	ENSG00000159184_HOXB13	0.07	-2.7	0	-1.78	1.78E-11	2.18E-10	Left
3	ENSG00000144451_SPAG16	0.05	-3.04	0	-0.64	9.00E-09	4.99E-08	Left
4	ENSG00000198353_HOXC4	0.05	-3.08	0	1.88	2.11E-15	6.43E-14	Right
5	ENSG00000184719_RNLS	0.03	-3.36	0	-0.89	2.85E-10	2.48E-09	Left
6	ENSG00000145649_GZMA	0.03	-3.42	0	1.54	4.75E-07	1.93E-06	Right
7	ENSG00000197757_HOXC6	0.02	-3.84	0	1.28	5.02E-12	1.02E-10	Right
8	ENSG00000162409_PRKAA2	0.02	-3.87	0	-1.2	3.90E-10	2.97E-09	Left
9	ENSG00000037965_HOXC8	0.02	-3.87	0	1.32	7.90E-12	1.20E-10	Right
10	ENSG00000147457_CHMP7	0.02	-3.92	0	0.35	8.44E-06	2.71E-05	Right
11	ENSG00000165548_TMEM63C	0.02	-3.99	0	-0.83	5.31E-05	0.000147	Left
12	ENSG00000203880_PCMTD2	0.02	-4.01	0	-0.43	5.16E-08	2.62E-07	Left
13	ENSG00000119397_CNTRL	0.02	-4.08	0.01	0.27	1.84E-06	7.00E-06	Right
14	ENSG00000103485_QPRT	0.02	-4.09	0.01	-1.01	2.20E-10	2.24E-09	Left
15	ENSG00000170677_SOCS6	0.02	-4.13	0.03	0.39	4.14E-06	1.40E-05	Right

Note. Top ranking features with p-values less than 0.05 and their importance scores discovered by our genes model (Left). Side-paired differential expression (fold change) analysis results of TPM values for the same features (Right) Wilcoxon-rank sum test was used to calculate p-values and FDR (Benjamini & Hochberg, 1995).

In the microbes-only dataset, 54 features were identified by Rf2pval as significantly important **(Fig. 3.2B, Table 3.4)**. The taxon with the highest importance score was *Ruminococcus gnavus*, which shows higher counts in RCC **(Table 3.4)**. *Clostridium acetireducens* ranked second and was more abundant in RCC.

Table 3.4.

Top ranking features with p-values less than 0.05 and their importance scores discovered by our microbes-only model (Left).

Model Feature Importance Metrics					Differential Expression			
Rank	**Tax ID_Name**	**Importance Score**	**Log Importance Score**	**p-value**	**Log2 FC**	**p-value**	**FDR**	**Greater Expr. Side**
1	33038_[Ruminococcus] gnavus	0.025	-3.69	0	2.07E+00	1.31E-15	3.54E-14	Right
2	76489_Clostridium acetireducens	0.020	-3.91	0	1.88E+00	1.69E-13	1.52E-12	Right
3	1701326_uncultured bacterium 5G4	0.018	-4.04	0	1.73E+00	6.55E-11	2.52E-10	Right
4	397291_Lachnospiraceae bacterium A4	0.014	-4.27	0	2.30E+00	5.29E-16	2.85E-14	Right
5	2293240_Ruminococcus sp. TF10-6	0.013	-4.34	0	2.41E+00	2.86E-15	5.14E-14	Right
6	239935_Akkermansia muciniphila	0.011	-4.51	0	-5.30E-01	2.27E-05	3.96E-05	Left
7	1531_[Clostridium] clostridioforme	0.011	-4.55	0	1.47E+00	1.73E-12	1.03E-11	Right
8	936381_Selenomonas sp. CM52	0.010	-4.61	0	4.03E+00	8.68E-11	3.13E-10	Right
9	46228_Ruminococcus lactaris	0.009	-4.76	0	1.86E+00	2.96E-12	1.45E-11	Right
10	43064_Trichococcus pasteurii	0.007	-4.97	0	2.18E+00	1.67E-08	4.10E-08	Right
11	1262831_Clostridium sp. CAG:678	0.007	-5.01	0	1.3	0.0322	0.0464	Right
12	1824_Nocardia asteroides	0.007	-5.02	0	-0.19	0.00112	0.00172	Left
13	208479_[Clostridium] bolteae	0.007	-5.03	0	1.76E+00	2.18E-13	1.68E-12	Right
14	2026799_Verrucomicrobia bacterium	0.006	-5.14	0.01	-0.17	2.02E-06	3.76E-06	Left
15	1262706_Azospirillum sp. CAG:260	0.006	-5.14	0	1.15	1.86E-09	5.57E-09	Right

Note. As per **Table 3.3**, with microbes-only model data. Side-paired differential expression (fold change) analysis results of CPM values for the same features (Right) Wilcoxon-rank sum test was used to calculate p-values and FDR (Benjamini & Hochberg, 1995).

59 features were deemed significant in the genes-and-microbes model **(Figure 3.2C, Table 3.5)**: 28 genomic features and 18 microbes. Notable features include *MYOM3*, *HOXC4*, *Coprococcus eutatus*, *PRAC1*, *lncRNA AC012531.3*, *Ruminococcus gnavus*, *RNLS*, *HOXC6*, *SPAG16*, and *Fusobacterium nucleatum*.

Table 3.5

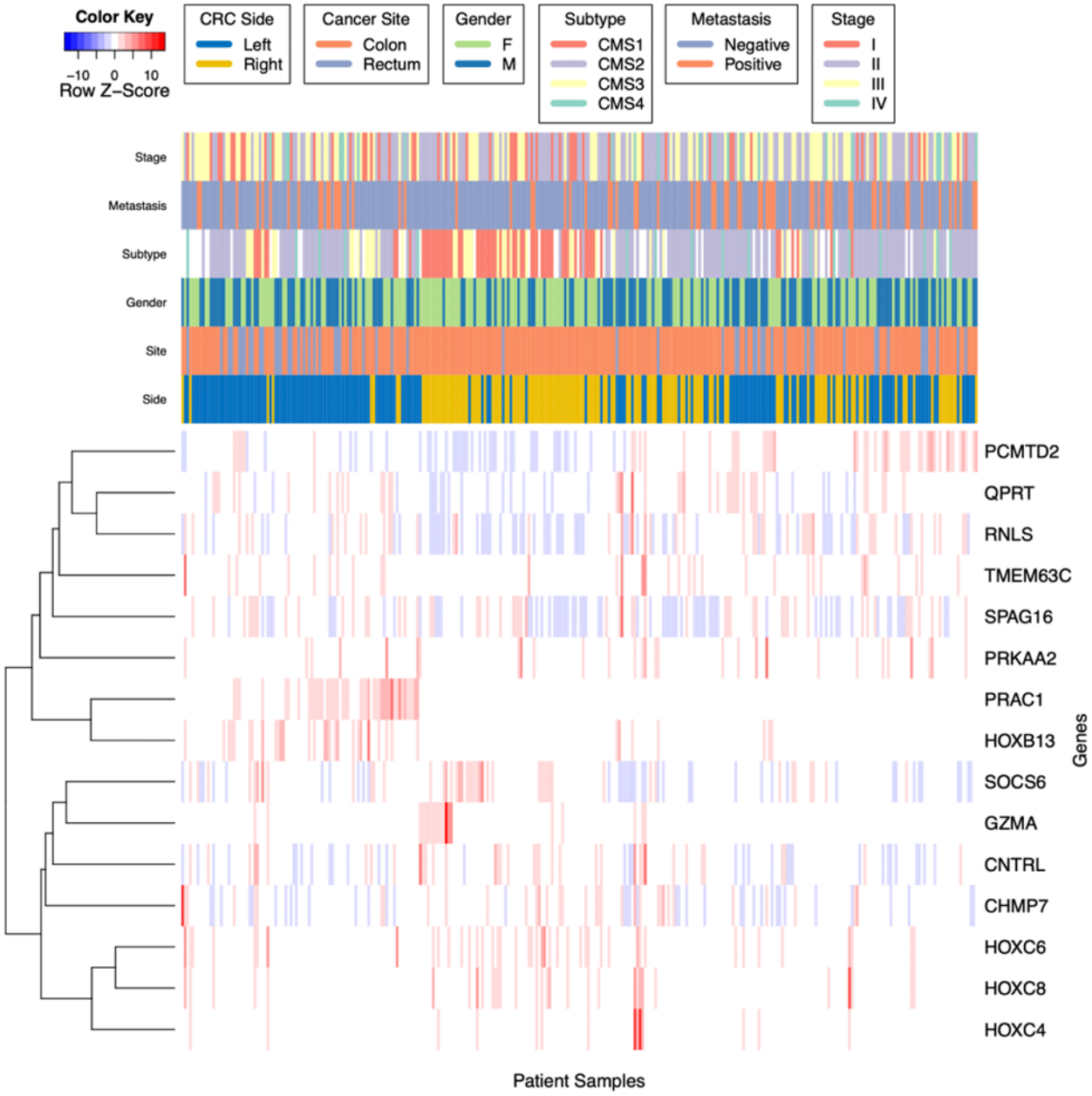
Top ranking features with p-values less than 0.05 and their importance scores discovered by our genes-and-microbes model (Left).

Model Feature Importance Metrics					Differential Expression			
Rank	**ENSG ID_Gene/Tax ID_Name**	**Importance Score**	**Log Importance Score**	**p-value**	**Log2 FC**	**p-value**	**FDR**	**Associated Side**
1	ENSG00000142661_MYOM3	0.037	-3.29	0	-0.62	1.36E-08	2.71E-08	Left
2	ENSG00000198353_HOXC4	0.033	-3.40	0	1.88	2.11E-15	3.23E-14	Right
3	33043_Coprococcus eutactus	0.027	-3.62	0	2.08	2.00E-14	1.54E-13	Right
4	ENSG00000159182_PRAC1	0.024	-3.72	0	-2.86	5.08E-21	2.34E-19	Left
5	ENSG00000260597_AC012531.25	0.020	-3.90	0	1.19	4.26E-12	2.17E-11	Right
6	33038_[Ruminococcus] gnavus	0.016	-4.13	0	2.07	1.31E-15	3.02E-14	Right
7	ENSG00000184719_RNLS	0.011	-4.50	0.01	-0.89	2.85E-10	9.37E-10	Left
8	ENSG00000197757_HOXC6	0.011	-4.54	0	1.28	5.02E-12	2.31E-11	Right
9	ENSG00000144451_SPAG16	0.010	-4.61	0	-0.64	9.00E-09	1.97E-08	Left
10	851_Fusobacterium nucleatum	0.010	-4.62	0	1.67	2.73E-06	3.49E-06	Right
11	ENSG00000273374_RP11-383I23.2	0.010	-4.62	0	-1.02	2.86E-08	5.47E-08	Left
12	446043_uncultured Lachnospira sp.	0.010	-4.63	0	1.46	5.97E-09	1.37E-08	Right
13	165179_Prevotella copri	0.009	-4.66	0	1.6	1.50E-09	3.85E-09	Right
14	154288_Turicibacter sanguinis	0.009	-4.67	0	1.9	5.00E-08	8.51E-08	Right
15	59620_uncultured Clostridium sp.	0.009	-4.67	0	1.09	6.86E-12	2.87E-11	Right

Note. As per **Tables 3** and **4**, with genes-and-microbes model data. Side-paired differential expression (fold change) analysis results of TPM and CPM values for the same features (Right) Wilcoxon-rank sum test was used to calculate p-values and FDR (Benjamini & Hochberg, 1995).

Figure 3.3

A heatmap of scaled gene expression values of the top-scoring genomic features discovered by the genes-only RF model and clinical characteristics.



Note. Hierarchical clustering of both genes and patients is via Pearson correlation, based on average linkage distance. The colors indicate row-scaled z-scores of TPM RNA-seq gene expression ratios.

We used hierarchical clustering to ascertain connections between gene expression profiles and clinical characteristics and consensus subtyping scores **(Figure 3.3)**. As expected, of the six

clinical characteristics that we considered (cancer stage, post-operative metastasis, consensus molecular subtype (CMS), gender, and site), side is most closely linked to the gene expression levels of our top genomic features. There is a clear cluster of left-sided CRC samples that show higher expression levels of *PRAC1* and *HOXB13* (left side of heatmap). There is also a subset of RCC that show higher expression of *HOXC4, HOXC6,* and *HOXC8* (middle of heatmap), although not all RCC exhibit this pattern. Heatmap for microbes-only model is shown in **Supplementary Figure 3.4**, and the heatmap for genes-and-microbes model is shown in **Supplementary Figure 3.5**.

In total, our models discovered 107 unique genomic and microbial features which played a significant role in the differentiating between CRC anatomical sides. Only six genomic features were common to both the genomic and genomic-plus-microbial models: *PRAC1, SPAG16, HOXC4, RNLS, HOXC6* and *PRKAA2*; and six microbes which were common to both our microbes-only, and genes-and-microbes models: *Ruminococcus gnavus, Ruminococcus sp. TF10-6, Selenomonas sp. CM52, Verrucomicrobia bacterium, Anaerostipes caccae* and *Turicibacter sanguinis.*

3.5 Discussion

Many of the previous studies on RCC vs LCC and gene expression have used the publicly available TCGA data. Here, we used a novel dataset of 308 patients, with microbial data from human unmapped reads, which adds to the growing body of evidence of the genomic and microbial differences between the sites (Jiang et al., 2020, Liang et al., 2018).

One difficulty in characterizing the roles of the microbiome and the genome in RCC vs LCC is that there is genomic and microbial heterogeneity both between and within the two anatomical locations (Liu et al., 1999). A primary reason for this heterogeneity is that the

proximal and distal areas of the colon have different embryonic origins and physiological functions: the right-side of the colon is derived from the embryonic midgut and is involved in digestion, and the left side of the colon is derived from the embryonic hindgut and is involved primarily in the storage of fecal matter and water absorption. Despite these different functions, the microbial content is similar in these two parts of the colon because they are attached, and peristaltic movement allows stool matter to pass both forwards and backwards (Martin 1914). Numerous studies have shown a strong correlation between gut dysbiosis and CRC, but less is known about the microbial taxa that differentiate RCC and LCC and, perhaps, play a role in carcinogenesis (Liu et al., 2021). *Fusobacterium, Prevotella, Clostridium, Akkermansia,* and *Ruminococcus* are among the most frequently reported bacteria in studies on CRC-related microbial dysbiosis (Liu et al., 2021; Lucas et al., 2017). All were deemed significantly important microbial taxa in the RF models presented here.

All three RF models showed strong predictive accuracy. The microbes only model showed the poorest predictive capability while the genes-only model was the highest performing. It is perhaps surprising that the combined model was not the most predictive. We postulate that this may be due to the fact that while microbes and genes may both affect CRC, microbial taxa are in fact indirect players, with effects that are reflected as altered genomic expression within the tumour, leading to cancer growth.

Finally, there were a number of highly important genes that differed between the genomic features only and combined RF models. One other point of interest is that there are some different top genes in the genes-only model when compared with the genes-and-microbes model. This may suggest that these genes and microbes act in consort and our genes-and-microbes RF model may have identified some underlying biological interactions.

3.5.1 Patterns in RCC

The RF models showed that increased expression of the *HOX* family of genes was characteristic of RCC. Specifically, we observed an upregulation of *HOXC4, HOXC6, HOXC8, and HOX-related lncRNA AC012531.3,* and a downregulation of *HOXB13* **(Tables 3.3, 3.5)**. The *HOX* (homeobox) gene family is most well-known for guiding embryonic development (Luo et al., 2019). *HOX* mutations that cause either increased or decreased expression have been associated with several types of cancer (Li et al., 2019) as tumour suppressors and proto-oncogenes. However, their role in CRC is not well understood (Li et al., 2019).

The top microbes identified by the microbes-only model include *Ruminococcus gnavus, Clostridium acetereducens, Lachnospiraceae,* and, *Ruminococcus sp. TF10-6* **(Table 3.4)**. *R. gnavus* causes inflammation in Crohn's disease models, and influences immunotherapy responses in CRC (Henke et al., 2019; Rebersek, 2021). *C. acetireducens* is an anaerobic bacterium that has no previously known associations to CRC. However, it is known to oxidize alanine to produce butyrate, and butyrate has been associated with CRC tumourigenesis (Okumura et al., 2021). *Lachnospiraceae spp* are also known to produce short-chain fatty acids which are known to have increased abundance in CRC patients (Yang et al., 2019); *Ruminococcus sp. TF10-6*, also more abundant in RCC; and *Akkermansia muciniphila*, more abundant in LCC. There is some evidence that the largely uncharacterized *lncRNA AC012531.3* which is located in one of the *HOX* gene loci, plays a role in colorectal cancer carcinogenesis (Wang et al., 2018).

For the genes-and-microbes model the top features include *Coprococcus eutactus, Ruminococcus gnavus, Fusobacterium nucleatum, Lachnospira sp.,* and *Prevotella copri. Coprococcus eutactus* has a very high feature importance score and is the microbe with the

highest association to RCC in our genes-and-microbes model **(Table 3.5)**. *C. eutactus* has previously been associated with longer cancer progression-free survival, and was not found in the microbes-only model, which could hint at a genomic-microbial interaction between *C. eutactus* and CRC side (Peters et al., 2019). *Ruminococcus gnavus, Ruminococcus sp.* and *Lachnospira* were previously identified as being important to CRC and associated with structural segregation of the mucosa (Chen et al., 2012). *Fusobacterium nucleatum* was found to be important in the genes-and-microbes model, but was not discovered by the microbes-only model. *F. nucleatum* is one of the most commonly associated species with CRC, and it is believed to act as a pathobiont (Han, 2015; Wang et al., 2022). It is also known to cause periodontal disease and is currently being explored as a biomarker for high-risk CRC. Given that *F. nucleatum* was only significant in the genes-and-microbes model, and that it is known to be only situationally pathogenic, this suggests this taxon may become pathogenic under specific gene co-activation (Han, 2015; Wang et al., 2022). *Prevotella copri*, also identified uniquely by our genes-and-microbes model, has been shown to be significantly enriched in the gut microbiome of CRC patients compared with normal patients (He et al., 2021).

3.5.2 Patterns in LCC

One recurring pattern in LCC is the expression of genes known to be associated with the prostate or prostate cancer. This is of interest given the heightened prevalence of LCC in men, and the left-sided colon's close proximity to the prostate (Lee et al., 2015). Prostate cancers and LCCs can be challenging to distinguish from biopsy samples, due to similarities in morphology and immunohistochemistry (Owens et al., 2007). Genes that are of high importance in our RF models that are associated with both LCC and prostate cancer include *PRAC1, HOXB13, SPAG16* **(Tables 3.3, 3.5)** (Luo et al., 2019). *PRAC1* has been previously associated with LCC as

well as prostate cancer (Jiang et al., 2020; Hu et al., 2018). *HOXB13* has a protective effect against tumour proliferation in RCC (Xie et al., 2019), and a reduction in the expression of *HOXB13* via hypermethylation of the *DNMT3B-HOXB13-C-myc* signaling axis has been associated with tumour proliferation and metastasis in RCC (Xie et al., 2019). Our results indicated that *HOXB13* is under-expressed in RCC relative to LCC, which adds evidence to the hypothesis that decreased *HOXB13* expression is specifically associated with RCC (Xie et al., 2019). Elevated *HOXC6* has been linked to poor overall survival in LCC patients, but not RCC patients (Xie et al., 2019). *MYOM3* was the top-ranking feature in our genes-and-microbes model and has a higher expression in LCC as determined using differential gene expression analysis. *MYOM3* has not been studied in CRC but it has been linked to clinical outcomes in renal and lung cancer (Human Protein Atlas, 2020; Yang et al., 2021).

While our microbes-only model mostly identified microbes associated with RCC, *Akkermansia muciphila* was more common in LCC **(Table 3.4)**. *A. muciniphila* degrades mucin in the gut, and has previously been shown to exacerbate colitis-associated CRC development in mice (Wang et al., 2022), and is associated with total pathological response in treatment of non-small cell lung cancer (Cascone et al., 2021). *Akkermansia* has been noted as one of three microbes most likely to have a causal association with differential CRC treatment effectiveness (Cercek et al., 2022).

The genes-and-microbes model also identified microbes that for the most part were enriched in RCC **(Table 3.5)**. Only two microbes in this model are present at higher levels in LCC, namely, *Verrucomicrobia bacterium* and *Fimbriiglobus ruber*. *Verrucomicrobia* has been studied as a biomarker for the early detection of CRC (Wu et al., 2021). However, *Fimbriiglobus* is largely uncharacterized.

3.6 Conclusions

Understanding microbial-genomic interactions may be important for informing treatment regimens in colorectal cancer. This study uses machine learning Random Forest (RF) models and differential gene expression (DE) to discover and associate genetic and microbial biomarkers with LCC and RCC. Three RF models with accuracy scores of 0.9, 0.7, and 0.87 were created and these yielded 15, 54 and 46 significantly important features. DE analysis was used to quantify changes in expression between CRC side. Our genes plus microbe model identified microbes that did not appear in our microbes-only model, including *C. eutactus, F. nucleatum* and *P. copri*, and this may indicate that the Random Forest model is uncovering interactive effects between genes and microbes. RCC was most associated with the *HOX* family of genes, including *HOX*-associated *lncRNA AC012531.25*. LCC was highly associated with prostate cancer related genes, which is of interest as LCC is more common in men. The future of CRC research lies in personalized genomics, and the biomarkers identified by these three classification models may play an important role in the observed variability in clinicopathological and treatment outcomes of CRC patients.

The use of Random Forest models for candidate biomarker discovery in this chapter directly inspired Chapter 4, which focuses on the development and application of pyRforest, an R package, designed to enhance the efficiency and interpretability of Random Forest models in genomic data analysis.

3.7 Declarations

Ethics approval and consent to participate

The study was approved by the University of Otago, New Zealand, Human Research Ethics Committee (approval number: H16/037). Informed consent was obtained from all subjects

and/or their legal guardians. All experiments were performed in accordance with relevant ethics guidelines and the Declaration of Helsinki.

Competing Interests

The authors declare that they have no competing interests.

Funding

The funding for the data analysis in this work was provided by Massey University School of Natural and Computational Sciences. Funding for data collection was provided by the lab of Dr. Rachel Purcell at University of Otago. Computational resources were provided by New Zealand eScience Infrastructure (NeSI).

Availability of data and materials

Raw Sequence Reads available at SRA Accession: PRJNA788974. Additional datasets and code are available from the corresponding author on request.

The R package Rf2pval used in these analyses is available at: www.github.com/tkolisnik/Rf2pval

Rf2pval has since been integrated into the R package pyRforest: www.github.com/tkolisnik/pyRforest

SciKit Learn can be imported from: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Supplementary Code is available online at BMC Cancer.

Acknowledgements

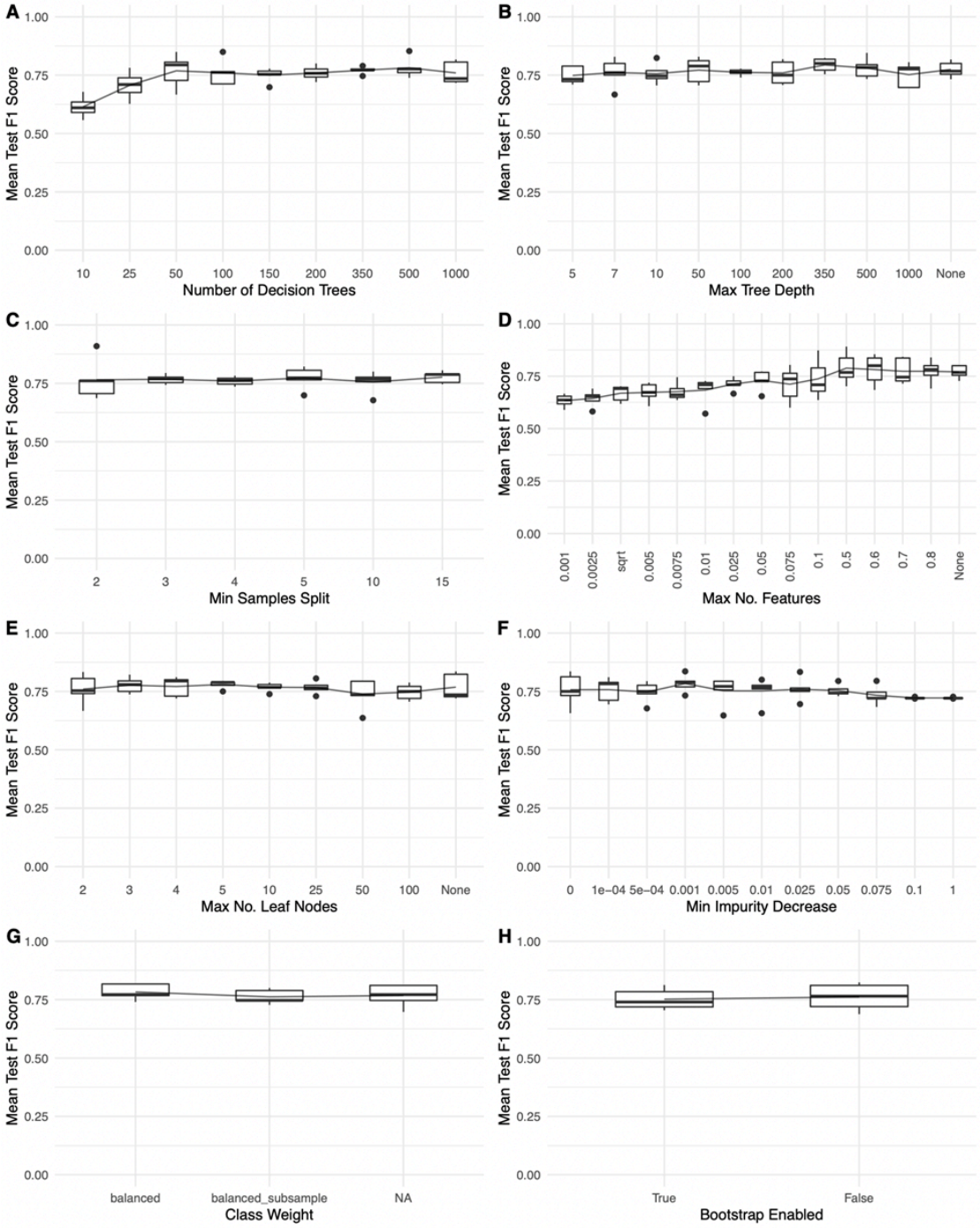
We would like to acknowledge Dr. Rachel Purcell, Dr. Frank Frizelle, and all members of the Purcell Lab team at University of Otago for providing the RNA-seq data that was used in this study.

The authors wish to acknowledge the use of New Zealand eScience Infrastructure (NeSI) high performance computing facilities, consulting support and/or training services as part of this research. New Zealand's national facilities are provided by NeSI and funded jointly by NeSI's collaborator institutions and through the Ministry of Business, Innovation & Employment's Research Infrastructure programme. URL https://www.nesi.org.nz.

3.8 Supplementary Material

Supplementary Figure 3.1

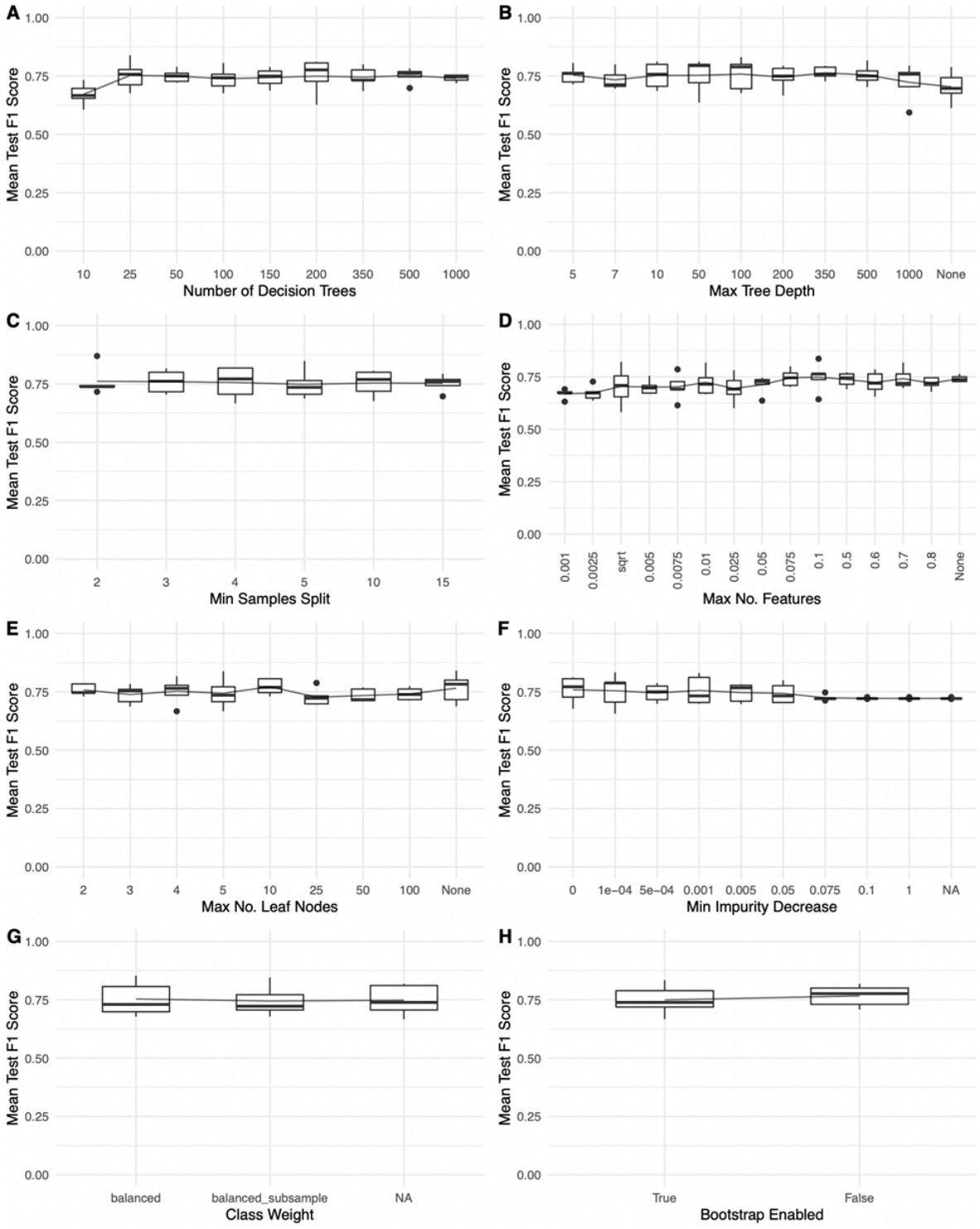
Genes-Only model mean test F1 scores over 8 hyperparameters set to a varying range of intervals.



Note. Used in building the RF genes-only model for purposes of narrowing down the hyperparameters searched during GridSearchCV to speed up computing time.

Supplementary Figure 3.2

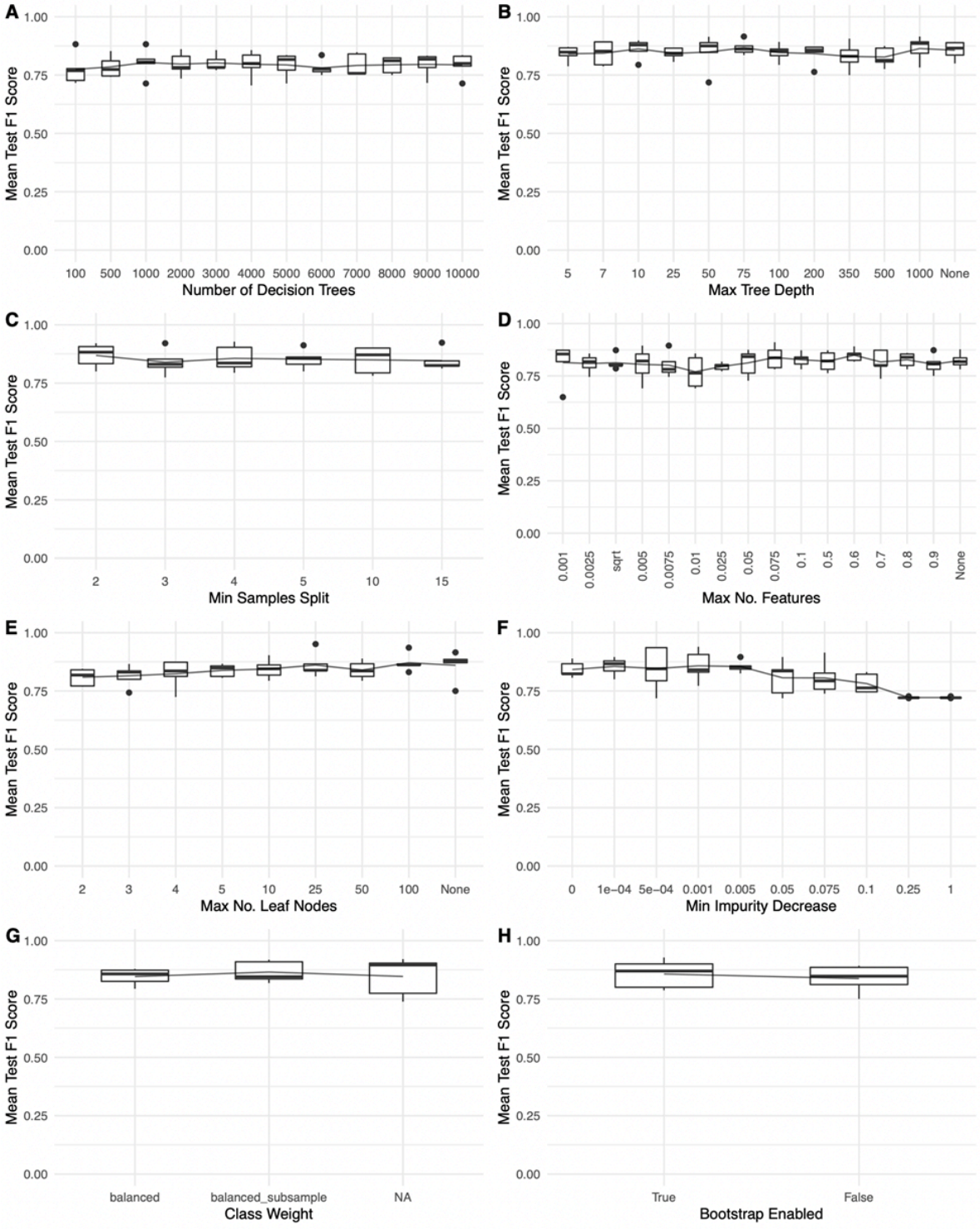
Microbes-only model mean test F1 scores over 8 hyperparameters set to a varying range of intervals.



Note. Used in building the RF microbial-only model for purposes of narrowing down the hyperparameters searched during GridSearchCV to speed up computing time.

Supplementary Figure 3.3

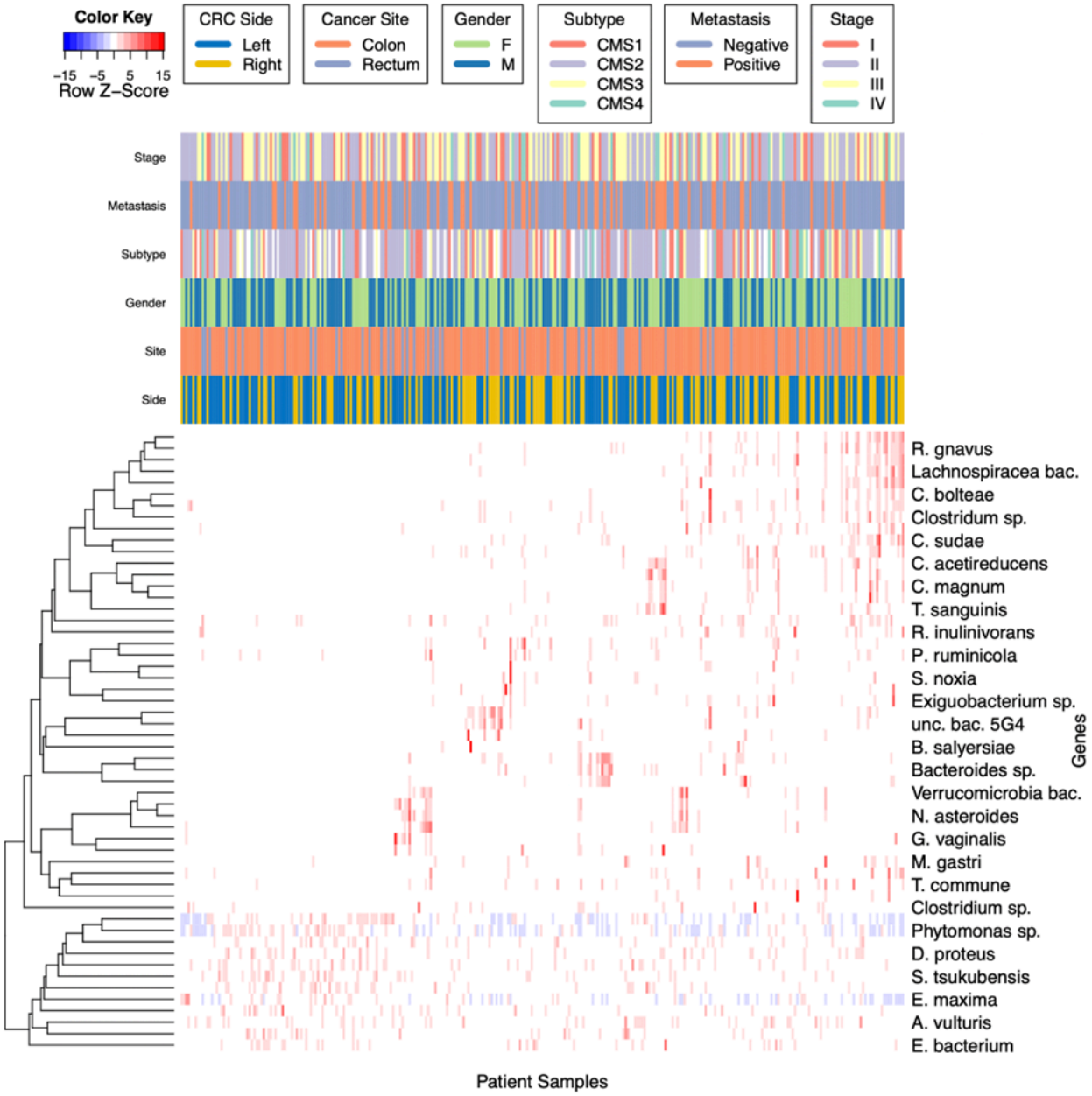
Genes-and-microbes model mean test F1 scores over eight hyperparameters set to a varying range of intervals.



Note. Used in building the RF genes-and-microbes model for purposes of narrowing down the hyperparameters searched during GridSearchCV to speed up computing time.

Supplementary Figure 3.4

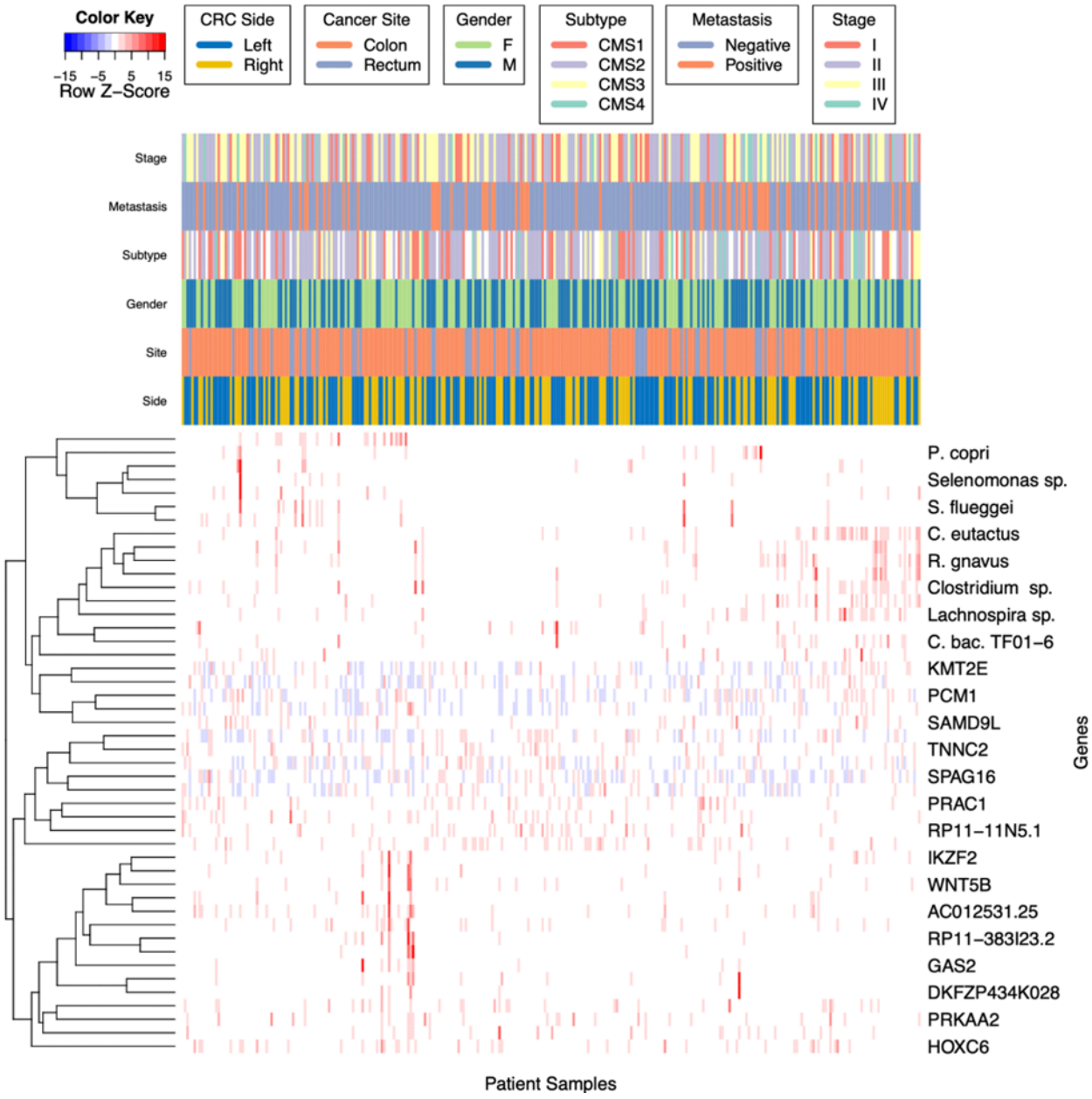
A heatmap of scaled gene expression values of the top-scoring microbial features discovered by the microbes-only RF model and clinical characteristics.



Note. Hierarchical clustering of both genes and patients is via Pearson correlation, based on average linkage distance. The colors indicate row-scaled z-scores of TPM RNA-seq gene expression ratios.

Supplementary Figure 3.5

A heatmap of scaled genomic and microbial expression values of the top-scoring features discovered by the genes-and-microbes RF model and clinical characteristics.



Note. Hierarchical clustering of both genes and patients is via Pearson correlation, based on average linkage distance. The colors indicate row-scaled z-scores of TPM RNA-seq gene expression ratios, as well as CPM normalized RNA-seq human unmapped microbial reads.

Supplementary Table 3.1

Full version of Table 3.4.

Model Feature Importance Metrics					Differential Expression			
Rank	**Tax ID_Name**	**Importance Score**	**Log Importance Score**	**p-value**	**Log2 FC**	**p-value**	**FDR**	**Greater Expr. Side**
1	33038_[Ruminococcus] gnavus	0.025	-3.69	0	2.07E+00	1.31E-15	3.54E-14	Right
2	76489_Clostridium acetireducens	0.020	-3.91	0	1.88E+00	1.69E-13	1.52E-12	Right
3	1701326_uncultured bacterium 5G4	0.018	-4.04	0	1.73E+00	6.55E-11	2.52E-10	Right
4	397291_Lachnospiraceae bacterium A4	0.014	-4.27	0	2.30E+00	5.29E-16	2.85E-14	Right
5	2293240_Ruminococcus sp. TF10-6	0.013	-4.34	0	2.41E+00	2.86E-15	5.14E-14	Right
6	239935_Akkermansia muciniphila	0.011	-4.51	0	-5.30E-01	2.27E-05	3.96E-05	Left
7	1531_[Clostridium] clostridioforme	0.011	-4.55	0	1.47E+00	1.73E-12	1.03E-11	Right
8	936381_Selenomonas sp. CM52	0.010	-4.61	0	4.03E+00	8.68E-11	3.13E-10	Right
9	46228_Ruminococcus lactaris	0.009	-4.76	0	1.86E+00	2.96E-12	1.45E-11	Right
10	43064_Trichococcus pasteurii	0.007	-4.97	0	2.18E+00	1.67E-08	4.10E-08	Right
11	1262831_Clostridium sp. CAG:678	0.007	-5.01	0	1.3	0.0322	0.0464	Right
12	1824_Nocardia asteroides	0.007	-5.02	0	-0.19	0.00112	0.00172	Left
13	208479_[Clostridium] bolteae	0.007	-5.03	0	1.76E+00	2.18E-13	1.68E-12	Right
14	2026799_Verrucomicrobia bacterium	0.006	-5.14	0.01	-0.17	2.02E-06	3.76E-06	Left

15	1262706_Azospirillum sp. CAG:260	0.006	-5.14	0	1.15	1.86E-09	5.57E-09	Right
16	360807_Roseburia inulinivorans	0.006	-5.15	0	1.9	1.49E-12	1.01E-11	Right
17	83915_Distigma proteus	0.006	-5.17	0	0.29	0.109	0.137	Right
18	1328_Streptococcus anginosus	0.006	-5.19	0	-0.36	0.106	0.136	Left
19	747377_Clostridium sp. DMHC 10	0.006	-5.20	0	1.99	2.12E-09	6.04E-09	Right
20	91626_Mucor ambiguus	0.006	-5.20	0	0.83	1.19E-09	3.78E-09	Right
21	1741_Thermodesulfobacterium commune	0.005	-5.21	0	-0.73	0.54	0.583	Left
22	88431_Dorea longicatena	0.005	-5.23	0	2.01	2.12E-14	2.86E-13	Right
23	2702_Gardnerella vaginalis	0.005	-5.25	0	-0.49	0.624	0.661	Left
24	33954_Clostridium magnum	0.005	-5.26	0	2.33	2.69E-08	6.32E-08	Right
25	291644_Bacteroides salyersiae	0.005	-5.27	0	2.17	0.00102	0.00161	Right
26	105841_Anaerostipes caccae	0.005	-5.27	0	1.67	1.90E-12	1.03E-11	Right
27	839_Prevotella ruminicola	0.005	-5.27	0	0.95	0.172	0.197	Right
28	134013_Phytomonas sp. isolate Hart1	0.005	-5.28	0	-0.29	0.0379	0.0525	Left
29	1262910_Oscillibacter sp. CAG:155	0.005	-5.29	0	-0.26	0.833	0.849	Left
30	83656_Streptomyces tsukubensis	0.005	-5.31	0	-0.68	0.000729	0.00119	Left
31	135083_Selenomonas noxia	0.005	-5.35	0	3.44	8.73E-06	1.57E-05	Right
32	162156_uncultured Bacteroides sp.	0.005	-5.35	0	0.39	1.87E-07	3.60E-07	Right
33	39486_Dorea formicigenerans	0.005	-5.35	0	1.23	4.61E-08	9.96E-08	Right
34	872327_Lactobacillus pasteurii	0.005	-5.35	0	1.21	1.31E-08	3.54E-08	Right
35	1736298_Exiguobacterium sp. Leaf196	0.005	-5.38	0	2.89	3.47E-12	1.56E-11	Right
36	310300_Bacteroides pyogenes	0.004	-5.41	0	1.76	3.29E-08	7.41E-08	Right
37	260710_Olavius algarvensis spirochete endosymbiont	0.004	-5.42	0	1.8	1.66E-08	4.10E-08	Right

38	1522369_Cyanophora sudae	0.004	-5.46	0	1.61	1.19E-13	1.28E-12	Right
39	857335_Candidatus Dactylopiibacterium carminicum	0.004	-5.46	0	0.15	0.14	0.168	Right
40	261299_Intestinibacter bartlettii	0.004	-5.48	0	2	2.23E-11	9.24E-11	Right
41	564198_Mycolicibacterium bacteremicum	0.004	-5.50	0	-0.82	0.0132	0.0198	Left
42	411484_Clostridium sp. SS2/1	0.004	-5.51	0	1.99	1.44E-10	4.87E-10	Right
43	1777_Mycobacterium gastri	0.004	-5.57	0.01	0.17	0.418	0.471	Right
44	1857645_Actinomyces vulturis	0.004	-5.58	0.01	0.47	0.161	0.189	Right
45	11768_Feline leukemia virus	0.004	-5.58	0.01	-0.49	6.37E-05	0.000107	Left
46	1736483_Microbacterium sp. Root180	0.004	-5.58	0.01	-0.7	0.129	0.159	Left
47	1768115_bacterium F082	0.004	-5.58	0.01	1.44	9.65E-08	1.93E-07	Right
48	877415_Erysipelotrichaceae bacterium NK3D112	0.004	-5.60	0.01	0.64	0.0327	0.0464	Right
49	1632858_Massilibacterium senegalense	0.004	-5.61	0.01	-0.09	0.444	0.489	Left
50	5804_Eimeria maxima	0.004	-5.61	0.01	-0.19	0.0681	0.0919	Left
51	154288_Turicibacter sanguinis	0.004	-5.62	0.01	1.9	5.00E-08	1.04E-07	Right
52	490622_Trichoderma arundinaceum	0.004	-5.63	0.01	-0.04	0.791	0.821	Left
53	653385_Actinomyces sp. oral taxon 849	0.004	-5.63	0.01	-0.05	0.938	0.938	Left
54	467085_Candidatus Symbiothrix dinenymphae	0.004	-5.65	0.01	-1.95	0.0923	0.122	Left

Note. Top ranking features with p-values less than 0.05 and their importance scores discovered by our microbes-only model (Left). Side-paired differential expression (fold change) analysis results of CPM values for the same features (Right) Wilcoxon-rank sum test was used to calculate p-values and FDR (Benjamini & Hochberg, 1995).

Supplementary Table 3.2

Full version of Table 3.5.

Model Feature Importance Metrics					Differential Expression			
Rank	**ENSG ID_Gene/Tax ID_Name**	**Importance Score**	**Log Importance Score**	**p-value**	**Log2 FC**	**p-value**	**FDR**	**Associated Side**
1	ENSG00000142661_MYOM3	0.037	-3.29	0	-0.62	1.36E-08	2.71E-08	Left
2	ENSG00000198353_HOXC4	0.033	-3.40	0	1.88	2.11E-15	3.23E-14	Right
3	33043_Coprococcus eutactus	0.027	-3.62	0	2.08	2.00E-14	1.54E-13	Right
4	ENSG00000159182_PRAC1	0.024	-3.72	0	-2.86	5.08E-21	2.34E-19	Left
5	ENSG00000260597_AC012531.25	0.020	-3.90	0	1.19	4.26E-12	2.17E-11	Right
6	33038_[Ruminococcus] gnavus	0.016	-4.13	0	2.07	1.31E-15	3.02E-14	Right
7	ENSG00000184719_RNLS	0.011	-4.50	0.01	-0.89	2.85E-10	9.37E-10	Left
8	ENSG00000197757_HOXC6	0.011	-4.54	0	1.28	5.02E-12	2.31E-11	Right
9	ENSG00000144451_SPAG16	0.010	-4.61	0	-0.64	9.00E-09	1.97E-08	Left
10	851_Fusobacterium nucleatum	0.010	-4.62	0	1.67	2.73E-06	3.49E-06	Right
11	ENSG00000273374_RP11-383I23.2	0.010	-4.62	0	-1.02	2.86E-08	5.47E-08	Left
12	446043_uncultured Lachnospira sp.	0.010	-4.63	0	1.46	5.97E-09	1.37E-08	Right
13	165179_Prevotella copri	0.009	-4.66	0	1.6	1.50E-09	3.85E-09	Right

14	154288_Turicibacter sanguinis	0.009	-4.67	0	1.9	5.00E-08	8.51E-08	Right
15	59620_uncultured Clostridium sp.	0.009	-4.67	0	1.09	6.86E-12	2.87E-11	Right
16	ENSG00000101470_TNNC2	0.009	-4.69	0	-1.29	3.34E-10	1.02E-09	Left
17	936381_Selenomonas sp. CM52	0.009	-4.69	0	4.03	8.68E-11	3.07E-10	Right
18	ENSG00000162409_PRKAA2	0.009	-4.69	0	-1.2	3.90E-10	1.12E-09	Left
19	ENSG00000124915_DKFZP434K028	0.009	-4.70	0	0.91	1.67E-05	1.97E-05	Right
20	712991_Lachnospiraceae bacterium oral taxon 500	0.009	-4.71	0	2.33	9.20E-07	1.28E-06	Right
21	ENSG00000101076_HNF4A	0.009	-4.72	0	-0.48	4.73E-08	8.37E-08	Left
22	ENSG00000177409_SAMD9L	0.009	-4.72	0	1.01	7.55E-08	1.20E-07	Right
23	2293240_Ruminococcus sp. TF10-6	0.009	-4.72	0	2.41	2.86E-15	3.28E-14	Right
24	ENSG00000146386_ABRACL	0.009	-4.72	0	-0.29	1.71E-04	0.000192	Left
25	2305245_Clostridiaceae bacterium TF01-6	0.009	-4.74	0	1.79	1.96E-14	1.54E-13	Right
26	ENSG00000111186_WNT5B	0.009	-4.75	0	-0.53	2.65E-04	0.00029	Left
27	ENSG00000188610_FAM72B	0.009	-4.75	0	0.67	2.58E-06	3.39E-06	Right
28	ENSG00000250829_RP11-11N5.1	0.009	-4.75	0	-3.79	1.14E-08	2.39E-08	Left
29	ENSG00000126218_F10	0.009	-4.75	0	-0.55	4.46E-10	1.21E-09	Left

30	28133_Prevotella nigrescens	0.009	-4.75	0	1.27	5.70E-08	9.37E-08	Right
31	ENSG00000103534_TMC5	0.009	-4.76	0	0.29	7.29E-03	0.0078	Right
32	113574_Hyphomicrobium sp. GJ21	0.009	-4.77	0	2.24	2.10E-09	5.09E-09	Right
33	ENSG00000078674_PCM1	0.008	-4.78	0	0.39	3.38E-08	6.22E-08	Right
34	2026799_Verrucomicrobia bacterium	0.008	-4.79	0	-0.17	2.02E-06	2.73E-06	Left
35	105841_Anaerostipes caccae	0.008	-4.80	0	1.67	1.90E-12	1.09E-11	Right
36	2320113_bacterium 1xD42-87	0.008	-4.80	0	1.4	2.73E-07	4.05E-07	Right
37	ENSG00000005483_KMT2E	0.008	-4.81	0	0.15	2.08E-02	0.0213	Right
38	ENSG00000197217_ENTPD4	0.008	-4.81	0	0.3	9.00E-07	1.28E-06	Right
39	ENSG00000030419_IKZF2	0.008	-4.81	0	0.48	1.36E-05	1.64E-05	Right
40	135080_Selenomonas flueggei	0.008	-4.81	0	3.21	9.96E-13	6.54E-12	Right
41	ENSG00000166173_LARP6	0.008	-4.84	0	-0.8	1.21E-05	1.51E-05	Left
42	ENSG00000267506_RP11-13K12.1	0.008	-4.85	0	-1.79	1.16E-11	4.43E-11	Left
43	ENSG00000148935_GAS2	0.008	-4.87	0	0.05	2.34E-07	3.59E-07	Right
44	ENSG00000026559_KCNG1	0.008	-4.87	0	-0.22	5.56E-02	0.0556	Left
45	1908690_Fimbriiglobus ruber	0.007	-4.90	0.04	-0.46	2.04E-02	0.0213	Left

46	ENSG00000144891_AGTR1	0.007	-4.91	0.03	-1.83	4.72E-05	5.42E-05	Left

Note. Top ranking features with p-values less than 0.05 and their importance scores discovered by our genes-and-microbes model (Left). Side-paired differential expression (fold change) analysis results of TPM and CPM values for the same features (Right) Wilcoxon-rank sum test was used to calculate p-values and FDR (Benjamini & Hochberg, 1995).

Supplementary Table 3.3

Patient Demographics & Cancer Characteristics Stratified by Cancer Side.

Characteristic	LCC Patients	RCC Patients
Patients enrolled - no (%)	172 (56)	136 (44)
Median Age - year (range)	72.7 (28.7, 89.8)	75.3 (36.9, 91.5)
Sex		
Female - no (%)	79 (26)	84 (27)
Male - no (%)	93 (30)	52 (17)
Metastasis		
Positive - no (%)	44 (14)	26 (8)
Negative - no (%)	128 (42)	110 (36)
Cancer Stage		
T1 - no (%)	34 (11)	19 (6)
T2 - no (%)	63 (20.5)	65 (21)
T3 - no (%)	63 (20.5)	42 (14)
T4 - no (%)	12 (4)	10 (3)
Nodal Status		
Positive - no (%)	98 (32)	87 (28)
Negative - no (%)	74 (24)	49 (16)

Note. An alternative version of **Table 3.1**, stratified by LCC and RCC patient groups.

Chapter 4

pyRforest: A comprehensive R package for genomic data analysis featuring scikit-learn Random Forests in R

The contents of this chapter have been published in *Briefings in Functional Genomics*.

Kolisnik, T., Keshavarz-Rahaghi, F., Purcell, R., Smith, A., & Silander, O. (2024).

pyRforest: A comprehensive R package for genomic data analysis featuring scikit-learn Random Forests in R. *Briefings in Functional Genomics, 2024*(38), 1-9.

https://doi.org/10.1093/bfgp/elae038.

Authors' contributions:
TK conceived the project, wrote the code and documentation for the R package, performed testing, analysis of results and figure generation, and wrote the first draft and contributed to all versions of the manuscript. FK assisted with testing of the pyRforest package and manuscript proof-reading. RP provided data and input on methodology and analysis. AS guided the development of statistical methods and validation, manuscript proof-reading, and guided presentation of results. OKS provided support with development, troubleshooting, statistics, and critical analysis of results, and contributed to the manuscript. All authors have read and approved this manuscript.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Tyler Kolisnik
Name and title of main supervisor:	Dr. Adam Smith, PhD
In which chapter is the manuscript/published work?	4

Describe the contribution that the student and members of the supervisory team have made to the manuscript/published work:[1]

TK conceived the project, wrote the code and documentation for the R package, performed testing, analysis of results and figure generation, and wrote the first draft and contributed to all versions of the manuscript. FK assisted with testing of the pyRforest package and manuscript proof-reading. RP provided data and input on methodology and analysis. AS guided the development of statistical methods and validation, manuscript proof-reading, and guided presentation of results. OKS provided support with development, troubleshooting, statistics, and critical analysis of results, and contributed to the manuscript.

Please select one of the following three options:

- (●) **The manuscript/published work is published or in press**

  Please provide the full reference of the research output:
  Kolisnik, T., Keshavarz-Rahaghi, F., Purcell, R., Smith, A., & Silander, O. (2024). pyRforest: A comprehensive R package for genomic data analysis featuring scikit-learn Random Forests in R. Briefings in Functional Genomics, 2024(38), 1-9. https://doi.org/10.1093/bfgp/elae038.

- ( ) **The manuscript is currently under review for publication**

  Please provide the name of the journal:

- ( ) **It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:	Date 2024.10.22	Main supervisor's signature:	Date: 2024.09.27 10:51:04 +12'00'

This form should be placed at the beginning of each relevant thesis chapter.

[1] Refer to the Massey University Publishing and Authorship guidelines (OneMassey for staff, Stream for students) and/ or Contributor Roles Taxonomy (CRediT) guidelines for guidance.

4.1 Abstract

Random Forest models are widely used in genomic data analysis and can offer insights into complex biological mechanisms, particularly when features influence the target in interactive, non-linear, or non-additive ways. Currently, some of the most efficient Random Forest methods in terms of computational speed are implemented in Python. However, many biologists use R for genomic data analysis, as R offers a unified platform for performing additional statistical analysis and visualization. Here we present an R package, pyRforest, which integrates Python scikit-learn `RandomForestClassifier` algorithms into the R environment. pyRforest inherits the efficient memory management and parallelization of Python, and is optimized for classification tasks on large genomic datasets, such as those from RNA-seq. pyRforest offers several additional capabilities, including a novel rank-based permutation method for biomarker identification. This method can be used to estimate and visualize p-values for individual features, allowing the researcher to identify a subset of features for which there is robust statistical evidence of an effect. In addition, pyRforest includes methods for the calculation and visualization of SHapley ADditive Explanations (SHAP) values. Finally, pyRforest includes support for comprehensive downstream analysis for gene ontology and pathway enrichment. pyRforest thus improves the implementation and interpretability of Random Forest models for genomic data analysis by merging the strengths of Python with R. pyRforest can be downloaded at: https://www.github.com/tkolisnik/pyRforest with an associated vignette at https://github.com/tkolisnik/pyRforest/blob/main/vignettes/pyRforest-vignette.pdf.

4.2 Introduction

The field of genomics has evolved over the past two decades, from studies primarily focused on individual genes to comprehensive analyses of entire genomes (Satam et al., 2023). This shift has been facilitated by advances in sequencing technologies and bioinformatics tools, allowing researchers to generate and analyze vast quantities of genomic data (Talukder et al., 2021). To derive meaningful biological insights from these large datasets, advanced statistical methods, such as machine learning, are increasingly required.

Decision trees are a supervised learning method frequently used for classification. The values of different features are used to split data along branches to make classification predictions. Random Forest (RF) models (Breiman, 2001) are ensembles of random trees, i.e., decision trees where only a random subset of features are used at each split, known for their high predictive power and robustness to dataset noise for handling large datasets, such as those of genomic research. RFs can mitigate overfitting through mechanisms specifically designed to introduce variability among the individual trees, such as bootstrapping and feature bagging. These properties make RF models particularly competitive for binary classification problems (Montesinos López et al., 2022).

Python (Python Language, 2019) is often used for implementing RF models due to having libraries designed with efficient memory handling and parallelization in mind. However, R (R Core Team, 2023) is commonly used as a computing environment by biologists, though implementations of RF models in R are slower and less optimized compared to Python (Kotthaus et al., 2015). Many RF implementations in R suffer from inefficient memory handling and lack parallelization strategies, which are crucial when processing large genomic datasets (Kotthaus et al., 2015). This creates a barrier for researchers who must often choose between the advanced

machine learning capabilities of Python and the comprehensive genomic analysis tools available in R.

To address these limitations, we present pyRforest, an R package that integrates the scikit-learn `RandomForestClassifier` algorithm (Pedregosa et al., 2011) implemented in Python into the R environment using the reticulate package (Ushey et al., 2024). pyRforest enables users familiar with R to leverage the machine learning strengths of Python without requiring any Python coding knowledge. This integration improves memory management and parallel processing, allowing the user to create RF models on larger genomic datasets with less RAM usage when compared to R, while enhancing the interpretability and biological relevance of RF models in genomic studies.

In addition, pyRforest offers several innovative features, including a novel rank-based permutation method for identifying significantly important features, by estimating and visualizing p-values for individual features. This allows researchers to prioritize a reduced list of biomarkers for further analysis, while speeding up computation. Additionally, pyRforest includes methods for calculating and visualizing SHapley ADditive Explanations (SHAP) values (Lundberg & Lee, 2017; Lundberg et al., 2020), while also supporting comprehensive downstream analysis for gene ontology and pathway enrichment using clusterProfiler (Yu et al., 2012) and g:Profiler (Reimand et al., 2007). By merging the computational strengths of Python with the statistical and visualization capabilities of R, pyRforest addresses current limitations in genomic research workflows, contributing to the ongoing evolution towards more versatile and integrated bioinformatics tools and the need for more explainable artificial intelligence (Hassija et al., 2024).

This paper introduces and outlines the capabilities of the pyRforest package for R. We first describe the development, use, and functionality of pyRforest. Next, we compare pyRforest to existing RF implementations, highlighting its computational advantages and innovative features. We then present a case study on colorectal cancer (CRC), demonstrating the utility of pyRforest in identifying key biomarkers and providing insights into their biological implications. Finally, we discuss the advantages and limitations of the package, as well as potential future directions for its development and application.

4.3 Methods and Design

R Package Development

pyRforest was created using the Rstudio IDE for R (RStudio Team, 2020). The devtools package (Wickham et al., 2022) was used to streamline the code development, testing and documentation of the package. Github was used for version control and collaboration.

4.3.1 Integration of R and Python

The pyRforest setup process supports the built-in virtualenv environments of reticulate (Ushey et al. 2024) as well as conda (Anaconda Software Distribution, 2020) environments for Python package management. In our experience, virtualenv works best on Linux systems, and conda is most suitable for Windows or Apple Arm64/M series Mac users.

For detailed setup instructions, including a step-by-step guide for configuring virtualenv or Conda with reticulate, we refer users to a vignette included with pyRforest. The vignette includes a demonstration of the capabilities of pyRforest with an example dataset. The development of pyRforest and all case studies were carried out on an M1 iMac with 16GB RAM. The package was also tested successfully on a Windows 10 x64 system with an Intel i5 processor, and Linux Ubuntu 20.04 LTS systems.

4.3.2 Dataset Preparation and Hyperparameter Optimization

pyRforest includes functions for formatting and partitioning data into training, validation, and testing sets. pyRforest optionally supports multiple class weighting strategies to ensure balance of the target variable across the data partitions, which can be particularly important if the sample size is unbalanced with respect to the target class.

pyRforest optionally uses scikit-learn `BayesSearchCV` or `GridSearchCV` (Pedregosa et al., 2011) to optimize RF hyperparameters, with default settings optimized for the genomic datasets presented in this paper. However, pyRforest allows users to customize hyperparameters, such as ‘n_estimators’ (number of decision trees), ‘max_features’ (maximum number of criteria for node splits), and ‘max_depth’ (maximum parent nodes on a decision tree), and others (see documentation for a full list) (Pedregosa et al., 2011).

4.3.3 Model Tuning and Evaluation

The model is initially trained using the training set, and tuned using the validation set, with final assessment scores calculated on the testing set. The default metric for assessing model performance is the Area under the Receiver Operating Characteristic (ROC) curve (AUC) score, which is a combined measure of sensitivity and specificity. However, the user can choose from a range of other metrics, including precision and accuracy. The validation test set allows the user to test for model overfitting via the training set. For example, if training results in overfitting, such models will often underperform on the validation set; this poorer performance can be used as an indication of overfitting, requiring further tuning or pruning of hyperparameters.

After the hyperparameter-tuning phase, the testing set is used to assess model performance. Importantly, the testing set is not used in model fitting or in tests for overfitting.

pyRforest provides the user with a range of scoring metrics, such as accuracy, ROC-AUC score, sensitivity, specificity, and F1 score.

4.3.4 Post-Hoc Feature Importance Significance Testing

pyRforest includes a procedure for identifying which features are significantly important to the predictive performance of the final model. The importance of each feature within the dataset is measured using the Gini importance score, and features are ranked from most to least important.

The statistical significance of each feature at each rank is determined using permutation to generate a null distribution of importance score profiles at each rank, to which the true importance scores of features at that rank are compared. This approach allows the calculation of p-values for each ranked feature by comparing the observed importance score to the null distribution of importance scores obtained under permutation for the corresponding rank. By default, pyRforest performs 1,000 permutations (although this is customizable). This permutation test allows users to identify a subset of features for which there is statistical evidence of their importance.

Generating a Null Distribution of Importance Scores at Each Rank

The null distribution of importance scores for each rank is generated using the following steps:

1. Fit RF model; calculate importance scores for each feature; rank features from highest to lowest importance score.
2. For each permutation (repeat 1000 times (default)):
    i. Randomly permute the values of the target variable across samples.
    ii. Fit RF model.

iii. Calculate importance scores and rank them from highest to lowest.

3. For each feature according to its importance rank:

i. Obtain the null distribution of importance scores under permutation for that rank.

ii. Determine the p-value: the proportion of null importance scores that are greater than the observed importance score.

The p-values for each feature capture the probability of obtaining an importance score as great or greater than the one observed at its rank under a null hypothesis of no association between features and target variable.

In comparison, in the most common approach to RF analysis, variable importance is assessed by the change in accuracy after the permutation of the value of a variable is obtained in an out-of-bag sample (Breiman, 2001; Liaw & Wiener, 2002). Other permutation-based methods generally fall into two types: in one, class labels are swapped, and for each feature, the distribution of importance scores (frequently, Gini) is tested relative to the unpermuted dataset. For example, the rfPermute package (Archer, 2023) uses the mean decrease in Gini score for all permutations relative to the non-permuted data. In the other type, class labels are swapped, and for each feature, the distribution of the ranks is tested relative to the unpermuted dataset (Altmann et al., 2010). This contrasts to the case of pyRforest, in which the feature importance (Gini) at each rank is compared to the importance of the feature at that rank in the unpermuted dataset. More succinctly, pyRforest uses rank-based feature importance rather than feature-based importance, or feature-based *rank* importance.

The result is a list of features for which, given a chosen significance threshold (alpha), there is statistical evidence for non-zero importance to the predictive outcome of the final model.

In a biological context, these features may be further studied as a list of potentially important biomarkers, or to gain insight into biological mechanisms.

pyRforest also offers plotting functions to visualize these features, leveraging the popular R package ggplot2 (Wickham, 2016). These plots can aid in the interpretation of the significance assessment. These permutation and visualization steps allow the researcher to determine which features (for example, genes) found by the RF model are deemed significant and worthy of further research.

4.3.5 SHAPley Additive exPlanations (SHAP)

SHAP values help explain the predictions of RF models, offering insight into when, why, and how specific features are important in determining class membership (Lundberg & Lee, 2017; Lundberg et al., 2020). To facilitate this, pyRforest offers built-in functions for calculating and plotting SHAP values from the `shap.TreeExplainer` class within the SHAP package. SHAP analysis can complement biological contextualization by providing a means to interpret the effects of individual features on model predictions.

4.3.6 Biological Interpretation

Finally, when using pyRforest on genomic data, pyRforest facilitates biological interpretation of the produced ranked lists of significant genes, which are automatically prioritized and formatted for compatibility with downstream analytical tools. As shown in the examples within the vignette, the resultant data format is suitable for direct integration with clusterProfiler (Yu et al., 2012) and g:Profiler (Reimand et al., 2007), powerful platforms for biological annotation and analysis. Specifically, these tools enable users to contextualize the statistically significant features reduced by pyRforest within biological pathways, functions, and processes.

4.4 Comparisons with Alternative Implementations

We first benchmarked memory usage and compute time for three different RF methods: R randomForest; scikit-learn `RandomForestClassifier` directly in Python; and our pyRforest implementation of scikit-learn `RandomForestClassifier` in R which leverages reticulate. **Table 4.1** shows the results of this benchmarking. As expected, pyRforest performed similarly to a direct implementation of scikit-learn `RandomForestClassifier` with some additional overhead memory and run time inefficiencies due to the integration layer between R and Python. Both pyRforest and scikit-learn `RandomForestClassifier` vastly outperformed R randomForest on large datasets. While R randomForest excelled with very small datasets, the inefficiencies in memory and run time rapidly scaled with dataset size. On our genomic training dataset of size 58,678 features x 248 samples, average memory usage of R randomForest was over 5x that of pyRforest, and the run time average was over 10x as long.

Table 4.1

A benchmarking analysis comparing memory usage and run time of Random Forest model fitting across different datasets, using R's randomForest, Python's scikit-learn `RandomForestClassifier`, and the pyRforest implementation of scikit-learn `RandomForestClassifier` in R.

Dataset	**Trees *†**	**Memory Usage (MiB)‡**			**Execution Time (seconds)**		
		randomForest	scikit-learn	pyRforest	randomForest	scikit-learn	pyRforest
Iris	50	25.2	0.39	996	0.10	0.58	0.39
Iris	100	43.1	0.41	1000	0.12	0.62	0.42
Iris	150	62.5	0.53	1010	0.14	0.64	0.58
Income	50	4310	26.8	732	24.9	2.69	124
Income	100	8040	38.3	837	49.8	4.61	254
Income	150	11760	50.3	1260	75.7	6.58	377
Demo RNAseq	50	14.5	0.38	1180	0.15	0.65	0.54

Demo RNAseq	100	19.8	0.42	1180	0.18	0.67	0.377
Demo RNAseq	150	25.2	0.52	1178	0.213	0.84	0.51
CRC RNAseq	50	13160	38.8	2290	308	3.70	34.3
CRC RNAseq	100	13190	40.7	2400	566	4.33	63.4
CRC RNAseq	150	13218	42.5	2454	822	5.12	92.5

* All parameters were held constant where possible

† Number of estimators in Python is equivalent to Number of trees in R

‡ Average Times and Average Memory Usage represent an average over 5 repetitions

Note. In the case of RNA-seq data, the target metric was ROC-AUC; for the Iris and Income datasets, the metric was accuracy. Execution time and memory usage are the average of five runs. In all cases, five-fold cross validation was used. Dataset sizes and classification types are: Iris, 150 samples, 4 features, multiclass; Income, 30,162 samples, 14 features, two-class; Demo RNA-seq, 40 samples, 101 features, binary; CRC RNA-seq, 248 samples, 58,678 features, binary.

We also compared the benchmarking and results of different feature identification approaches in **Supplementary Table 4.1** using the CRC RNA-seq dataset. This table illustrates a comparative analysis of feature identification approaches, evaluating the number of important features with non-zero importance scores found, run time, and peak memory usage during benchmarking. Feature importance results were assessed on the default scikit-learn `RandomForestClassifier`, pyRforest, scikit-learn's `inspection` module with 1000 permutations, the randomForest R package, and the rfPermute R package. As shown in this table, pyRforest's approach to rank-based permutation feature importance testing offers a considerably faster approach to feature identification relative to scikit-learn's `inspection`, a Python package that also employs permutation testing. In multiple attempts at testing the R packages that utilize permutation, randomForest (Liaw & Wiener, 2002) and rfPermute (Archer, 2023) (which relies

on R randomForest), the memory usage became prohibitive, ultimately causing the process to crash. This comparative analysis demonstrated the nuanced capability of pyRforest for feature selection. Of the 58,678 features available in the CRC RNA-seq dataset, pyRforest identified 83 genomic features with significant importance. This is a considerably reduced set compared to the 1008 features that had non-zero importance in the default scikit-learn `RandomForestClassifier` model without any feature prioritization applied. In contrast, we found the scikit-learn `inspection` module to be overly conservative in identifying features, as it identified only a single feature and missed several features that are well-established to affect CRC location (Jiang et al., 2020; Kolisnik et al., 2023).

Several other advantages of pyRforest relative to the two other RF implementations in R are shown in **Supplementary Table 4.2**. For example, in contrast to other tools, pyRforest innately supports direct class weighting, k-fold cross-validation, and simple integration with cross-validation tools, streamlining the analysis process.

In summary, pyRforest bridges the computational and methodological divide between Python and R, providing users with advantages of both: the advanced machine learning capabilities and efficiencies of Python scikit-learn and the comprehensive analysis strengths of R. It introduces an innovative rank-based permutation method tailored for biomarker prioritization in genomic data analysis. Additionally, pyRforest simplifies GO analysis with easy integration into clusterProfiler and g:Profiler (Reimand et al., 2007; Yu et al., 2012). This unique combination of features positions pyRforest as a strong tool for end-to-end genomic studies.

4.5 Case Study

4.5.1 RNA-seq Data Analysis in Colorectal Cancer

In this case study, we employed pyRforest to analyze a dataset comprising RNA-seq data from colorectal cancer samples obtained from the University of Otago, Christchurch, New Zealand analyzed in our previous study (Kolisnik et al., 2023). This dataset includes 308 patient samples, and encompasses a comprehensive range of 58,678 genomic features, including genes and long non-coding RNAs (lncRNAs). The primary focus of this analysis was to distinguish the cancer location within the colorectum based on the binary outcome variable 'side' (left vs. right).

4.5.2 Data Preparation and Model Training

The raw genomic data was mapped to the human genome (GRCh38) using STAR (v2.73a) (Dobin et al., 2013) and TPM normalized to remove sequencing depth and gene length biases. It was then split into training (248 samples), validation (30 samples), and testing (30 samples) sets, ensuring a robust evaluation for the RF model developed using pyRforest. The validation and testing sets were balanced for outcome, and the training set was minimally unbalanced with 57% left-sided samples and 43% right-sided samples. The RF model was trained and tuned using the pyRforest exhaustive grid search, yielding optimal hyperparameters noted in **Supplementary Table 4.3**. The initial training of the model on the training set identified 1008 features with non-zero Gini importance scores, indicating a significant level of complexity and feature interaction within the data.

4.5.3 Model Performance

In our case study, the RF model exhibited strong performance across a range of scoring metrics, including accuracy, F1 score, precision, recall and ROC-AUC score, as seen in **Table**

4.2. Notably, the model achieved high accuracy rates of 0.87 and 0.80 for the validation and testing sets respectively, with both F1 and ROC-AUC scores closely mirroring these values at 0.88 and 0.80. The model demonstrated strong generalization capabilities on the unseen testing dataset, and consistently high performance at discriminating between the outcome classes.

Table 4.2

K-fold cross-validation scoring metrics for the RF model, stratified by dataset splits (Validation, Testing) on the CRC dataset from our case study.

Scoring Metric	Validation Set	Testing Set
Accuracy	0.87	0.8
F1	0.88	0.8
Precision	0.78	0.8
Recall	1.0	0.8
ROC-AUC	0.88	0.8

Note. The table presents key performance metrics including Accuracy, F1 score, Precision, Recall, and ROC-AUC.

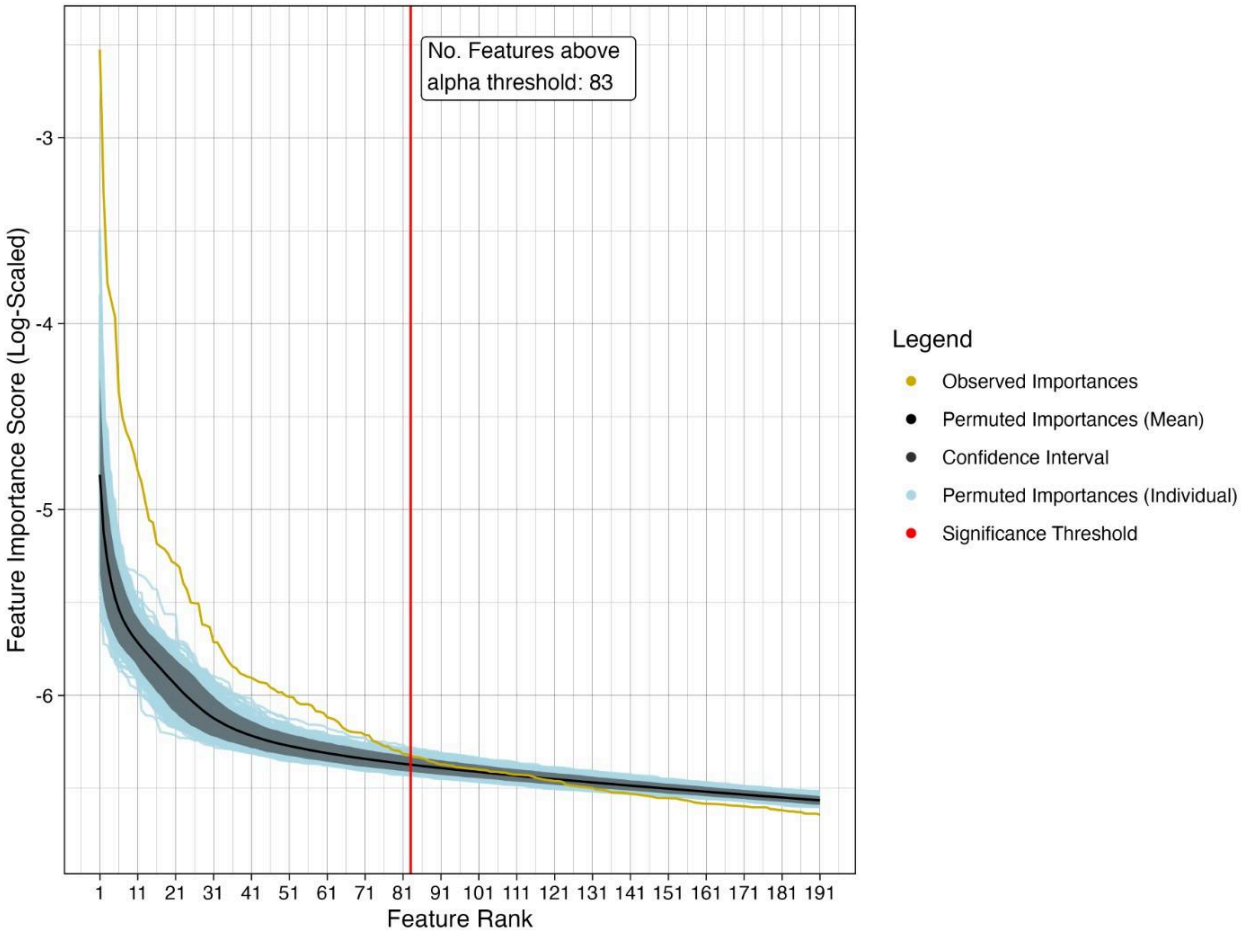
4.5.4 Identifying Significantly Important Features

The rank-based permutation approach of pyRforest identified 83 significantly important biomarkers of the total available set of 58,678 **(Figure 4.1)**. Each feature was assigned a p-value, providing a statistical basis for their inclusion in the downstream analyses. Notably, the small nuclear protein *PRAC1* emerged as the top feature, indicating its potential role in the lateralization of CRC, which is a result consistent with many previous studies (Jiang et al., 2020; Kolisnik et al., 2023). *HOXB13*, and the *lncRNA ENSG00000242407* were also identified as important features. *HOXB13* and this lncRNA colocalize with *PRAC1*: all are between 48.721 and 48.734 Mbp on chromosome 17. Their functional relationship to *PRAC1* is not clearcut. *HOXC4* and *HOXC6* were also identified; as developmental genes, these are functionally related to *HOXB13* and overlap on chromosome 12 between 53.991 and 54.017 Mbp. In addition, *QPRT*

(tryptophan metabolism) and *WASF3* (cell movement and adhesion), were identified, both of known importance to CRC (Yu et al., 2024; Zhang et al., 2022). Finally, pyRforest identified a second lncRNA, *ENSG0000250829* and this has no identified function, suggesting new avenues for future research. *PRAC1* was the sole feature identified by the scikit-learn `inspection` feature identification approach**.**

Figure 4.1

Feature importance plot showing individual and mean rank-based feature importance scores of the permuted data.



Note. Feature ranks are shown on the x-axis, with the importance score on the y-axis. The yellow line indicates the importance scores of the features at each rank in the unpermuted data. The blue lines indicate the importance scores at each rank for the permuted data sets, and the shaded gray region contains 95% of the ranked importance scores of the permuted data. The significance threshold was determined based on an alpha of 0.05 (i.e. the importance score of the ranked feature in the unpermuted is larger than the score in 95% of permuted data), and is indicated by the red vertical line.

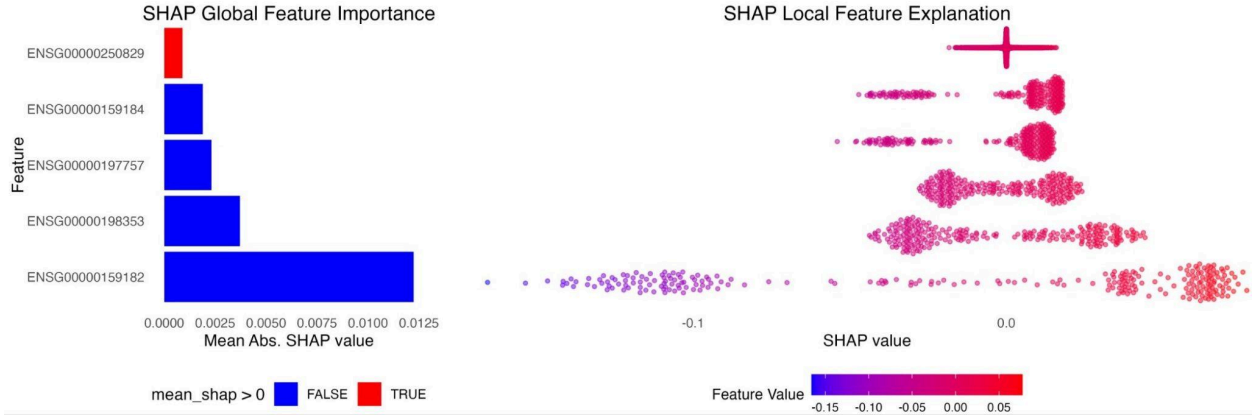
4.5.5 SHAP Analysis

Calculation of SHAP values provides additional insights into the direction of the contribution (positive or negative) of these features to the target variable's outcome within the

RF model, here defined as associated with left-sidedness **(Figure 4.2)**. SHAP importance rankings are independent from pyRforest rankings. Specifically, SHAP analysis indicated that *PRAC1* and *HOXB13* contribute to the predisposition towards left-sided CRC. This observation is consistent with outcomes from other studies, highlighting the role of these biomarkers as being differentially expressed depending on spatial location of the disease (Jiang et al., 2020; Sulit et al., 2023). *HOXC4* and *HOXC6* SHAP directionality deviated from expected values from our previous study: the bimodal feature explanations of these genes indicate the relationship of these genes with outcome may be complex, and that their contribution is context dependent. In other words, depending on the expression level of other genes, these can contribute either positively or negatively toward left-sided CRC. SHAP analysis indicated that *lncRNA ENSG0000250829* is the top feature contributing towards right-sided CRC, although the cross-shaped spread of the local feature explanation also indicates a complex, non-linear relationship with outcome. The inclusion of SHAP analysis enhances our understanding of the decisions of the predictive model, offering a clearer picture of how individual features can influence the output of the model.

Figure 4.2

SHAPley Additive Explanation plots which illustrate the impact of individual features on model prediction.



Note. The left panel displays the global feature importance with the mean absolute SHAP values, where blue signifies features driving predictions towards 'left-sided CRC' and red signifies

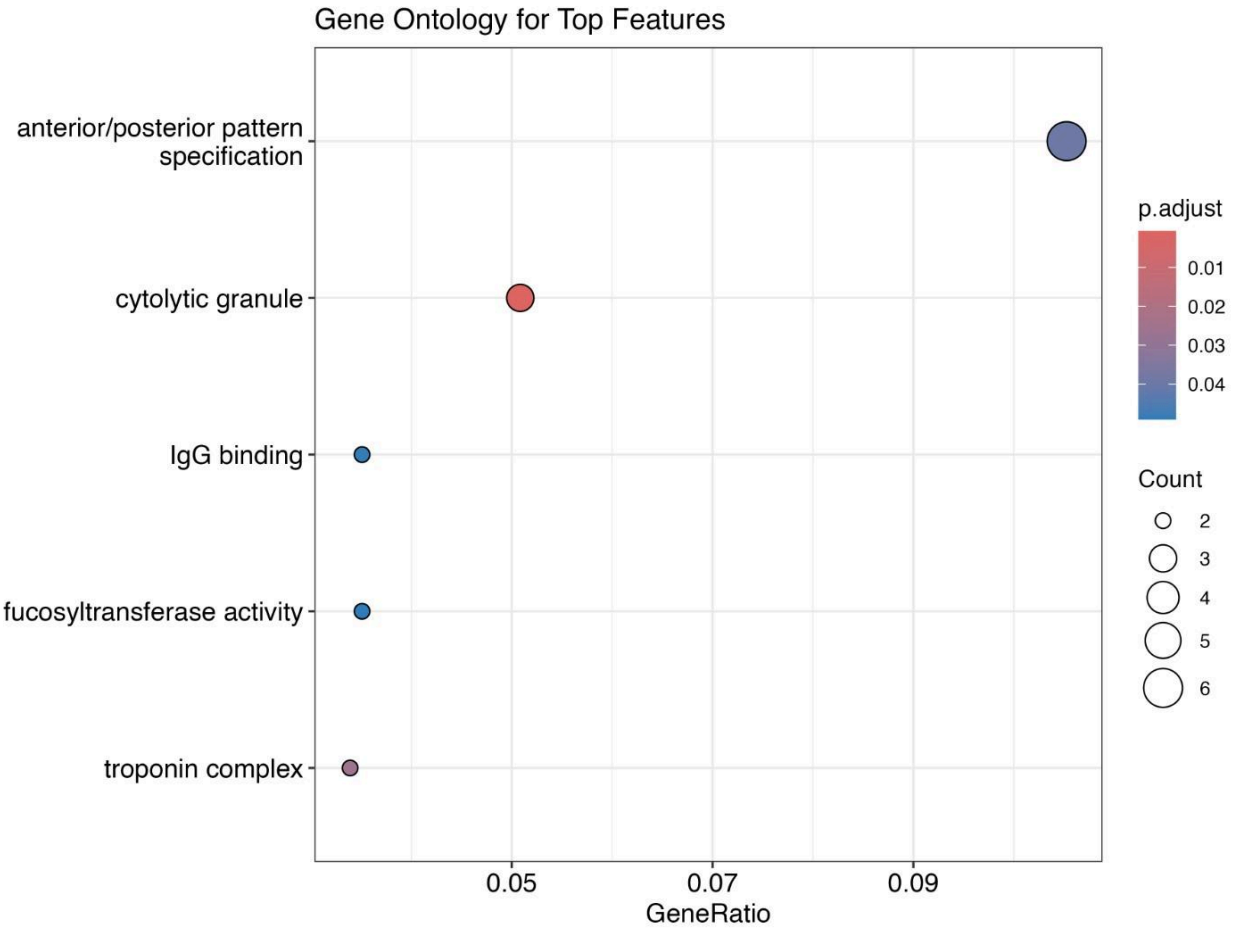
features driving predictions towards 'right-sided CRC'. The right panel shows the distribution of individual SHAP values for each feature, reflecting their contribution to each prediction - negative SHAP values suggest a push towards 'left-sided CRC', whereas positive values indicate a push towards 'right-sided CRC'. The color intensity corresponds to the feature value magnitude, with cooler colors representing lower values and warmer colors indicating higher values. From top to bottom, the ENSEMBL Gene IDs in the figure above correspond to: the novel lncRNA transcript *AC108865.1*; *HOXB13*; *HOXC6*; *HOXC4*; and *PRAC1*.

4.5.6 clusterProfiler and g:Profiler Analysis

pyRforest returns a list of genes that can easily be used as input for the R packages clusterProfiler and g:Profiler. The clusterProfiler (Yu et al., 2012) gene ontology (GO) enrichment analysis using the 83 identified features, revealed involvement in significant biological processes such as ‘anterior/posterior pattern specification’ and ‘cytolytic granule’ **(Figure 4.3)**. The top results from g:Profiler (Reimand et al., 2007) gene enrichment analysis also reinforce the findings that there is a potential link between prostate development processes and CRC anatomical side **(Figure 4.4)**. The enriched GO terms provide biological plausibility to the findings of the case study, indicating that the significant features identified by pyRforest are not only statistically relevant but also biologically grounded in the context of CRC lateralization.

Figure 4.3

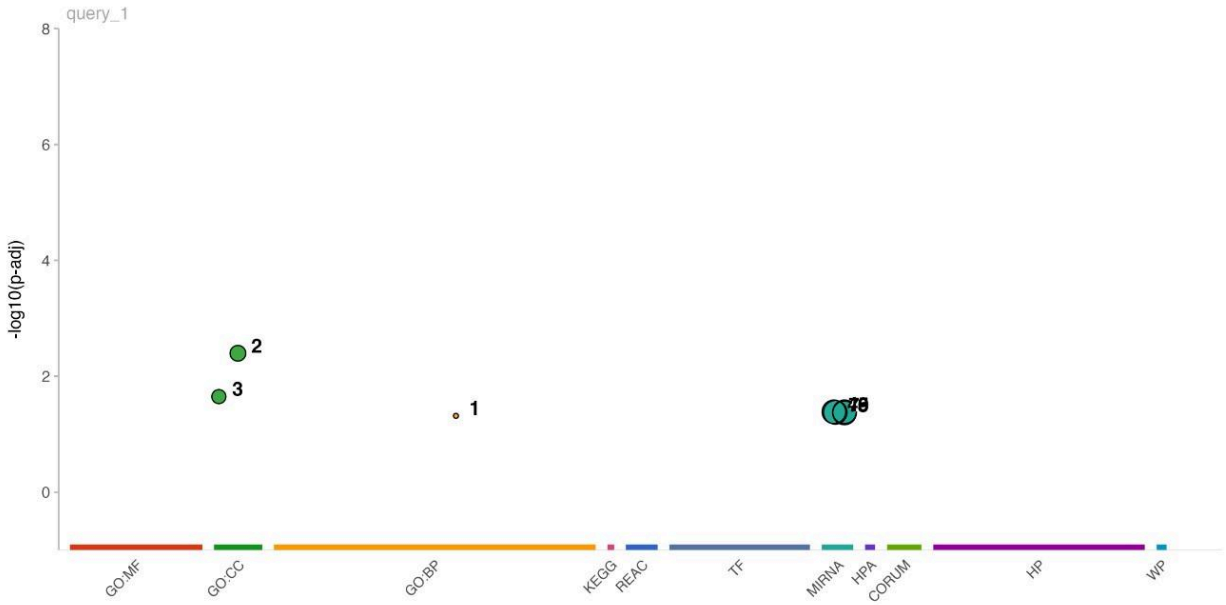
Gene Ontology enrichment analysis using clusterProfiler on the significant features identified by pyRforest.



Note. The x-axis represents the GeneRatio, the proportion of genes involved in the GO term relative to the total number of genes studied. The y-axis lists the GO terms associated with the identified features. Circle size indicates the gene count associated with each term, while the color gradient represents the adjusted p-value (p.adjust), with cooler (bluer) colors indicating less significance.

Figure 4.4

Enrichment analysis Manhattan plot from g:Profiler calculated on the 83 significant features found in our CRC case study.



id	source	term_id	term_name	term_size	p_value
1	GO:BP	GO:0060743	epithelial cell maturation involved in prostate gland development	3	4.8e-02
2	GO:CC	GO:0044194	cytolytic granule	14	4.0e-03
3	GO:CC	GO:0005861	troponin complex	9	2.2e-02
4	MIRNA	MIRNA:hsa-miR-4419a	hsa-miR-4419a	298	4.1e-02
5	MIRNA	MIRNA:hsa-miR-6130	hsa-miR-6130	298	4.1e-02
6	MIRNA	MIRNA:hsa-miR-6129	hsa-miR-6129	299	4.2e-02
7	MIRNA	MIRNA:hsa-miR-4510	hsa-miR-4510	300	4.2e-02
8	MIRNA	MIRNA:hsa-miR-6127	hsa-miR-6127	302	4.3e-02
9	MIRNA	MIRNA:hsa-miR-6133	hsa-miR-6133	302	4.3e-02

g:Profiler (biit.cs.ut.ee/gprofiler)

Note. The x-axis categorizes enriched terms from various databases, and the y-axis shows their significance (-log10 p-value). Points represent terms, with notable significantly enriched terms named in the table below the plot.

4.6 Discussion

pyRforest combines the computational power of Python and the statistical and genomic analysis capabilities of R into one package, addressing a crucial need in modern genomic research. pyRforest offers computational speed and memory efficiency advantages over common R-based RF implementations **(Table 4.1)**. Moreover, pyRforest offers a range of extra utility

functions for biomarker selection and analysis. These include (1) a rank-based permutation approach which allows the user to identify the features that have statistically significant importance scores, (2) integration SHAP for feature interpretation and (3) compatibility with clusterProfiler and g:Profiler for producing gene enrichment results which allow the user to explore the list of significantly important features in the context of external databases of known biological functions **(Figures 4.3** and **4.4)**.

A primary objective of analyzing genomic data with RF models is to discern important biomarkers. We present a comparison of pyRforest with two other packages which each differ in their approach to using permutation to identify important features. The 'permutation_importance' function in the scikit-learn `inspection` module (Pedregosa et al., 2011) independently permutes features, which is computationally intensive and destroys any associations among features. In contrast, pyRforest permutes only the response variable, thereby leaving the relationships among features intact. This approach offers advantages in computational speed, and we consider this approach to yield a more appropriate null model, given the extent of the known correlations and gene-gene interactions present within real genomic data. Another R package, rfPermute (Archer, 2023), acts as a wrapper for the randomForest R package and, like pyRforest, permutes only the target variable rather than the features. However, it compares the importance of each feature to its own distribution of importances scores under a permuted target variable. In contrast, pyRforest compares each observed feature importance to the distribution of permuted importance scores at its rank, rather than to the distribution of permuted importance scores of the feature itself. The null distribution for the most important feature (rank 1) is the scores obtained at rank 1 under permutation, regardless of the possibility of the identity of the specific feature changing with each

permutation. We consider that these two properties, permuting the target variable and not the features, and using rank-based importance distributions to offer a more appropriate null model and basis for comparing observed importance scores, and, ultimately, identifying a reduced set of potentially important features.

The approach of pyRforest also leads to two key effects that differentiate it from other rank-based methods, such as that of the previously mentioned rfPermute (Archer, 2023), and the approach implemented in Altmann et al., 2010, which determines feature significance by comparing the rank of a feature in the test data to the distribution of ranks in permuted data. First, pyRforest can more effectively select important variables in datasets where a large proportion of the variables are predictive. For example, in a dataset with 100 variables, a variable's importance rank in permuted data will typically center around 50 (with a standard deviation of approximately 7 under a Poisson distribution). However, if 50 or more variables are predictive, traditional methods may not detect any variable as significantly important compared to its average rank in the permuted data. In contrast, pyRforest may still identify the 50^{th} variable as important if its observed importance is significantly greater than in the permuted data. Second, pyRforest is more attuned to feature importances taking extreme values, especially in permuted datasets. Such values may be uncommon, but they are often crucial for identifying important features in an RF model. Feature-rank based methods can obscure these outliers, as they rely on non-parametric comparisons against a null distribution, potentially missing important features with high importance.

In addition, the pyRforest rank-based approach forgoes the need to permute and evaluate each feature individually, and this generally offers improved speed compared to feature-based approaches.

Despite its innovative contributions, the pyRforest R package comes with limitations worth noting. The primary challenge is the computational time associated with the permutation process, although benchmarking results indicate that pyRforest is significantly faster than the widely used package scikit-learn `inspection`. One potential enhancement could be to reduce computation time by fitting a theoretical probability distribution to an empirical distribution derived from a smaller subset of permutations (Altmann et al., 2010). Additionally, pyRforest is currently tailored specifically for classification problems, but extending its capabilities to support regression analysis would further enhance its versatility and allow it to be applied to additional machine learning problems.

In summary, despite these limitations, we have demonstrated the utility of pyRforest as a powerful analytical tool. In our case study on a CRC dataset, pyRforest identified 83 significant biomarkers and facilitated the exploration of gene ontology, SHAP values, and gene set enrichment, far surpassing the 15 biomarkers reported in the original study (Kolisnik et al., 2023). The integration with SHAP values allowed us to uncover new insights into how each feature influences cancer location, highlighting pyRforest's ability to uncover deeper biological insights. This package offers researchers a robust framework for biomarker discovery. Furthermore, our rank-based permutation approach was also successfully applied in a study by Keshavarz-Rahaghi et al., 2022, *(includes Kolisnik),* identifying key features associated with p53 activity, which enhances our understanding of TP53-related transcriptional signatures across various cancer types.

4.7 Conclusion

By integrating Python's efficient `RandomForestClassifier` algorithm, pyRforest enables researchers to leverage computationally efficient machine learning approaches while staying in

the R ecosystem. pyRforest offers a suite of tools for fitting, interpreting, and contextualizing RF models specifically for genomic studies. Its rank-based permutation approach for biomarker identification, alongside SHAP analysis, and integration with gene ontology tools aims to improve the interpretability and biological interpretation of RF models. These features make pyRforest a valuable resource for conducting and interpreting genome-scale RF studies. Furthermore, performance improvements in the scikit-learn package can be easily integrated into pyRforest ensuring that the package remains up to date without requiring extensive code modifications. We encourage the research community to utilize, contribute to, and build upon pyRforest.

4.8 Author Statements

Acknowledgements

We would like to acknowledge the Department of Surgery at the University of Otago for providing the patient samples and RNA-seq data used in the case study.

Data Availability

Raw Sequence Reads are available at SRA BioProject ID: PRJNA788974 (NCBI 2021).

Code Availability

The R package pyRforest used in these analyses is available at: www.github.com/tkolisnik/pyRforest.

Funding

The funding for the data analysis in this work was partially provided by the Massey University School of Natural Sciences.

Ethics Approval and consent to participate

This study was approved by the University of Otago, New Zealand, Human Research Ethics Committee (approval number: H16/037). Informed consent was obtained from all subjects and/or their legal guardians. All experiments were performed in accordance with relevant ethics guidelines and the Declaration of Helsinki.

4.9 Supplementary Material

Supplementary Table 4.1

A comparative analysis of feature identification approaches, evaluating the number of features with non-zero importance, run time, and peak memory usage during benchmarking.

Method	N Features with Non-Zero Importance	Run Time (hours)	Peak Memory Usage (16GB Max)
Default scikit-learn `RandomForestClassifier`	1008	8 h	10GB
pyRforest	83	6 h	8GB
scikit-learn `inspection` 'permutation_importance' (1000 feature-based permutations)	1	95 h	16GB
randomForest R package	NA	NA	16GB
rfPermute R package	NA	NA	16GB

Note. Feature importance results were assessed on the default s*cikit-learn* `RandomForestClassifier`, pyRforest, scikit-learn `inspection` 'permutation_importance' with 1000 permutations, the randomForest R package, and the rfPermute R package.

Supplementary Table 4.2

Feature and Capability Comparison of RandomForest Implementations for Genomic Data Analysis.

Feature/ Capability	Scikit-learn `RandomForestClassifier`	pyRforest	R randomForest	Rfpermute
Direct Class Weighting	Yes (via class_weight parameter)	Yes (via class_weight parameter)	No	Yes, via a preprocessing function
Parallel Processing Support	Yes (via n_jobs parameter)	Yes (via n_jobs parameter)	Requires manual setup or additional packages	Yes but memory inefficient, poor batching
Automatic Out-of-Bag (OOB) Error Estimate	Yes (via oob_score parameter)	Yes (via oob_score parameter)	Yes	Yes
Handling Missing Values	No (requires preprocessing)	Yes - preprocesses missing values	Yes (automatic handling with na.action)	Yes

Cross-Validation Integration	Easy integration with *scikit-learn* based cross-validation tools	k-fold cross-validation integrated	Requires manual setup or additional packages	Requires manual setup or additional packages
Grid Search for Hyperparameter Tuning	Direct integration with *scikit-learn `GridSearchCV`*	Direct integration with *scikit-learn `GridSearchCV`*	No. Requires extensive manual setup or additional packages	No. Requires extensive manual setup or additional packages
Built-in Feature Importance Evaluation	Yes (attribute feature_importances_)	Yes provides base model feature importance scores and permuted feature importance scores	Yes, with options for different importance types	Yes, with options for different importance types
Feature identification	Requires additional modules such as sklearn.inspection	Yes: null-distribution permutation rank based feature identification	Requires manual setup or additional packages	Yes: null-distribution permutation feature based feature identification
Extensive Ecosystem Compatibility	High (Python data science stack)	Extremely High - Compatible with both Python ecosystems for machine learning and R ecosystems for bioinformatics	Moderate (R ecosystem, less extensive for machine learning, but more extensive for bioinformatics)	Moderate (R ecosystem, less extensive for machine learning, but more extensive for bioinformatics)

Supplementary Table 4.3

The top-scoring model hyperparameters as optimized by the pyRforest `tune_and_train_rf_model` function which makes use of scikit-learn `GridSearchCV`.

Parameter	**Setting**
Bootstrap	TRUE
class_weight	balanced
criterion	gini
max_depth	8
max_features	0.2
min_samples_leaf	4
min_samples_split	2
n_estimators (# trees)	100
warm_start	FALSE

Chapter 5 Conclusions

5.1. Overview

The field of genomics has been transformed by technological innovations that enable the high-throughput sequencing of genetic material, resulting in huge amounts of biological data. To extract meaningful insights from these datasets we need advanced computational tools and methods. This thesis aims to develop useful bioinformatics tools and machine learning techniques to enhance the analysis and interpretation of genomic and metagenomic data, and apply these tools in the context of colorectal cancer research.

In the preceding chapters, I have presented user-friendly bioinformatics applications for data exploration, the application of Random Forest models for biomarker discovery, and the integration of efficient machine learning algorithms into R for genomic data analysis. This concluding chapter summarizes key findings from each chapter, discusses their implications, and outlines potential future directions for research in this field.

5.2 Summary of Findings

5.2.1 Development of the MetaFunc App (Chapter 2)

In Chapter 2, I presented the development of the MetaFunc App, an interactive Shiny application designed to facilitate the exploration of metagenomic and metatranscriptomic data generated by the MetaFunc pipeline (Sulit et al., 2023). The MetaFunc pipeline processes raw RNA-seq data to produce comprehensive taxonomic and functional profiles of microbial communities alongside host expression data.

The MetaFunc App addresses the challenges of interpreting the complex output from MetaFunc by providing a user-friendly interface that integrates data exploration and analysis

tools. Key features of the app include dynamic data tables that allow users to filter, sort, and export data related to microbial abundances, gene ontology terms, and host gene expression. The app enables users to link functional annotations to specific microbial taxa and vice versa, facilitating a deeper understanding of microbial roles in host environments. Automated database generation and integration with the MetaFunc pipeline ensure new datasets are available for loading within the app after each run of the MetaFunc pipeline.

A usage case of the MetaFunc App was demonstrated using a colorectal cancer dataset, which showcased its ability to identify microbial species associated with polyamine biosynthesis, a process linked to tumour progression. Benchmarking tests indicate that the app scales efficiently with large datasets when adequate computational resources are available.

5.2.2 Identifying Biomarkers for Colorectal Cancer Sidedness Using Random Forests (Chapter 3)

Chapter 3 focuses on the application of Random Forest machine learning models to identify genomic and microbial biomarkers that differentiate right-side colorectal cancers from left-side colorectal cancers. Using RNA-seq TPM normalized expression data for 58,677 human genes and CPM normalized data for 28,557 microbial taxa from 308 patient tumour samples, I developed three Random Forest models: (1) A genes-only model which achieved a predictive accuracy of 90% and identified 15 significant genomic features; (2) A microbes-only model which achieved a predictive accuracy of 70% and identified 54 significant microbial features; and (3) a combined genes-and-microbes model which achieved a predictive accuracy of 87% and identified 46 significant features comprising 28 genes and 18 microbes.

Key findings from these models included the identification of important genomic biomarkers such as *PRAC1*, *HOXB13*, *HOXC4*, *HOXC6*, and *RNLS*, which are associated with

developmental processes and have known links to cancer biology (Hu et al., 2018; Liang et al., 2018; Luo et al., 2019). Significant microbial biomarkers identified included *Ruminococcus gnavus*, *Clostridium acetireducens*, and *Fusobacterium nucleatum,* which have also been identified in other literature studies (Liu et al., 2021; Lucas et al., 2017; Wang et al., 2022). *Fusobacterium nucleatum* was also identified as a significant microbe of interest in the MetaFunc paper, which utilized a different CRC dataset and relied on Spearman correlation (Sulit et al., 2023). These findings suggest a role for these microbes and genes in colorectal cancer pathogenesis, highlighting the association of distinct microbial communities with specific cancer locations. While this does not imply that the microbes cause cancers to occur on the left or right side, the findings show that sidedness can be predicted based on microbial composition alone, and that left-sided and right-sided CRCs harbour different microbial communities, potentially interacting with host gene expression.

The use of Random Forest models can allow the identification of features that may interact in complex, non-linear ways (Chen & Ishwaran, 2012), providing a more nuanced understanding of these factors contributing to colorectal cancer sidedness. The benefits of using Python's scikit-learn Random Forest module coupled with the challenges of using Python for downstream genomic data analysis in this chapter directly inspired the creation of the pyRforest package presented in Chapter 4.

5.2.3 Development of the pyRforest R package (Chapter 4)

In Chapter 4, I introduced pyRforest, an R package that integrates Python's scikit-learn RandomForestClassifier into the R environment via the reticulate package. This integration addresses the limitations of existing Random Forest implementations in R regarding

computational efficiency and memory management, and the limitations of using Python for downstream genomic and functional analyses.

Key features of pyRforest include efficient machine learning capabilities that leverage Python's optimized algorithms for faster computation and reduced memory usage when handling large genomic datasets. This package introduces a novel rank-based permutation method to estimate p-values for individual features, first demonstrated in Chapter 3, aiding in the identification of significant biomarkers. It also incorporates SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017) to interpret the contribution of each feature to model predictions. Furthermore, pyRforest facilitates downstream analysis with the gene ontology and pathway enrichment tools clusterProfiler (Yu et al., 2012) and g:Profiler (Reimand et al., 2007) for biological contextualization.

I demonstrate the utility of pyRforest through a case study. The package successfully identified candidate biomarkers previously identified in Chapter 3, that are also known in wider literature for their importance in colorectal cancer. Furthermore, it provided additional insights into their biological implications, showcasing its potential for enhancing genomic analyses.

5.3 Future Directions

5.3.1 Enhancing the MetaFunc App

Further development of the MetaFunc App could involve integration of additional databases, such as KEGG for pathway analysis (Kanehisa & Goto, 1999), to provide users with richer visualizations of functional insights. Adding interactive plotting capabilities and visuals on other results of the MetaFunc pipeline, such as for DEGs and Spearman correlation could expand the utility of the app and serve to aggregate results in one place. Performance optimization of the

app's backend processes, possibly through increased parallelization could enhance scalability and speed, allowing for further efficiency improvements on handling large datasets.

5.3.2 Expanding the use of Machine Learning for Biomarker Identification

Further research could explore the use of other machine learning algorithms, such as gradient boosting machines or deep learning models to improve prediction performance. Integrating additional multi-omics data, including metabolomic and proteomic datasets could provide a more holistic understanding of colorectal cancer disease mechanisms. Re-training and testing the models on additional multi-site colorectal cancer datasets could increase robustness and reduce overfitting making the model more suitable for real-world applications. Expanding into analysis of different data types such as single-cell data and subsequently applying the identified biomarkers in prospective clinical studies could validate their utility in diagnosis, prognosis, or treatment stratification.

5.3.3 Further Development of pyRforest

Extending pyRforest to handle regression problems would broaden its applicability to other types of genomic studies. Porting additional scikit-learn modules from Python into R such as GradientBoostingClassifier (Pedregosa et al., 2011), would provide users with a wider variety of machine learning models to choose from. Another potential enhancement to the rank-based feature importance permutation test could involve fitting a theoretical probability distribution to an empirical distribution derived from a smaller set of permutations, which would reduce computation time while maintaining statistical rigor. Further enhancements could also include more advanced hyperparameter tuning algorithms, and expanding the SHAP analysis to better interpret complex models.

5.4 Concluding Remarks

This thesis demonstrates the potential of integrating bioinformatics tools and machine learning techniques to advance the analysis of genomic and metagenomic colorectal cancer data. The development of the MetaFunc App provides researchers with an accessible platform to explore the complex results of the MetaFunc pipeline. The application of Random Forest models highlights the power of machine learning for biomarker discovery, revealing genomic and microbial factors associated with disease characteristics. The creation of pyRforest merges the computational efficiency of Python with the analytical strengths of R. This integration enhances the explainability of the RF models through methods such as rank-based permutation, SHAP analysis, gene ontology, and gene set enrichment.

The findings and tools presented in this thesis contribute to a deeper understanding of colorectal cancer biology and offer valuable resources for the broader genomics community. By enabling more accessible, efficient, and interpretable analyses, these contributions support ongoing efforts to translate complex biological data into meaningful clinical applications.

In conclusion, the intersection of bioinformatics, machine learning, and genomics holds promise for advancing biomedical research. Continued development and collaboration in these areas will be essential for unlocking new insights into disease mechanisms, improving diagnostic and prognostic tools, and ultimately enhancing patient outcomes.

References

Alboaneen, D., Alqarni, R., Alqahtani, S., Alrashidi, M., Alhuda, R., Alyahyan, E., & Alshammari, T. (2023). Predicting colorectal cancer using machine and deep learning algorithms: Challenges and opportunities. *Big Data and Cognitive Computing, 7*(2), 74.

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics, 26*(10), 1340–1347.

Anaconda Software Distribution. (2020). In Anaconda Documentation (Version 2-2.4.0). Anaconda Inc. https://docs.anaconda.com/.

Archer, E. (2023). rfPermute: Estimate permutation p-values for random forest importance metrics. https://github.com/EricArcher/rfPermute.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: Tool for the unification of biology. *Nature Genetics, 25*(1), 25–29.

Attali, D. (2022). shinyjs: Easily improve the user experience of your shiny apps in seconds. https://deanattali.com/shinyjs/.

Ballinger, A. B., & Anggiansah, C. (2007). Colorectal cancer. *BMJ, 335*(7622), 715–718.

Baran, B., Mert Ozupek, N., Yerli Tetik, N., Acar, E., Bekcioglu, O., & Baskin, Y. (2018). Difference between left-sided and right-sided colorectal cancer: A focused review of literature. *Gastroenterology Research and Practice, 11*(4), 264–273.

Baratloo, A., Hosseini, M., Negida, A., & El Ashal, G. (2015). Part 1: Simple definition and calculation of accuracy, sensitivity, and specificity. *Emergency (Tehran, Iran), 3*(2), 48-49.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, 57*(1), 289-300.

Bergen, E. S., Scherleitner, P., Ferreira, P., Kiesel, B., Müller, C., Widhalm, G., Dieckmann, K., Prager, G., Preusser, M., & Berghoff, A. S. (2021). Primary tumor side is associated with prognosis of colorectal cancer patients with brain metastases. *ESMO Open, 6*(3), 100168.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). Classification and regression trees (1st ed.). Routledge.

Breitwieser, F. P., & Salzberg, S. L. (2020). Pavian: Interactive analysis of metagenomics data for microbiome studies and pathogen identification. *Bioinformatics (Oxford, England), 36*(4), 1303–1304.

Brennan, C. A., & Garrett, W. S. (2016). Gut microbiota, inflammation, and colorectal cancer. *Annual Review of Microbiology, 70*, 395–411.

Cascone, T., William, W. N., Jr, Weissferdt, A., Leung, C. H., Lin, H. Y., Pataer, A., Godoy, M. C. B., Carter, B. W., Federico, L., Reuben, A., Khan, M. A. W., Dejima, H., Francisco-Cruz, A., Parra, E. R., Solis, L. M., Fujimoto, J., Tran, H. T., Kalhor, N., Fossella, F. V., Sepesi, B. (2021). Neoadjuvant nivolumab or nivolumab plus ipilimumab in operable non-small cell lung cancer: The phase 2 randomized NEOSTAR trial. *Nature Medicine, 27*(3), 504–514.

Cercek, A., Lumish, M., Sinopoli, J., Weiss, J., Shia, J., Lamendola-Essel, M., El Dika, I. H., Segal, N., Shcherba, M., Sugarman, R., Stadler, Z., Yaeger, R., Smith, J. J., Rousseau, B.,

Argiles, G., Patel, M., Desai, A., Saltz, L. B., Widmar, M., Diaz, L. A., Jr. (2022). PD-blockade in mismatch repair-deficient locally advanced rectal cancer. *The New England Journal of Medicine.* https://doi.org/10.1056/NEJMoa2201445.

Chang, W., Cheng, J., Allaire, J. J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2024). shiny: Web application framework for R. https://shiny.posit.co/.

Chen, W., Liu, F., Ling, Z., Tong, X., & Xiang, C. (2012). Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer. *PloS One, 7*(6), e39743.

Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics, 99*(6), 323–329.

Coleman, O. I., & Haller, D. (2021). Microbe-mucus interface in the pathogenesis of colorectal cancer. *Cancers, 13*(4), 616.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology, 17*(1), 13.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21–27.

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal, 6*(2), 94–98.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner.

Bioinformatics, 29(1), 15–21.

Dumas, J., Gargano, M. A., & Dancik, G. M. (2016). shinyGEO: A web-based application for analyzing gene expression omnibus datasets. *Bioinformatics, 32*(23), 3679–3681.

Eide, P. W., Bruun, J., Lothe, R. A., & Sveen, A. (2017). CMScaller: An R package for consensus molecular subtyping of colorectal cancer pre-clinical models. *Scientific Reports, 7*(1), 16618.

Eide, P. W., Moosavi, S. H., Eilertsen, I. A., Brunsell, T. H., Langerud, J., Berg, K. C. G., Røsok, B. I., Bjørnbeth, B. A., Nesbakken, A., Lothe, R. A., & Sveen, A. (2021). Metastatic heterogeneity of the consensus molecular subtypes of colorectal cancer. *Genomic Medicine, 6*(1), 59.

Expression of MYOM3 in renal cancer - The Human Protein Atlas. (2020). https://www.proteinatlas.org/ENSG00000142661-MYOM3/pathology/renal+cancer.

Fontana, E., Eason, K., Cervantes, A., Salazar, R., & Sadanandam, A. (2019). Context matters—Consensus molecular subtypes of colorectal cancer as biomarkers for clinical trials. *Annals of Oncology, 30*(4), 520–527.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232.

Gao, Z., Guo, B., Gao, R., Zhu, Q., & Qin, H. (2015). Microbiota disbiosis is associated with colorectal cancer. *Frontiers in Microbiology, 6.* https://doi.org/10.3389/fmicb.2015.00020.

Garza, D. R., Taddese, R., Wirbel, J., Zeller, G., Boleij, A., Huynen, M. A., & Dutilh, B. E. (2020). Metabolic models predict bacterial passengers in colorectal cancer. *Cancer & Metabolism, 8*(1), 3.

Gene Ontology Consortium. (2024). AmiGO 2: Welcome. Retrieved August 28, 2024, from

https://amigo.geneontology.org/amigo.

Ge, S. X., Jung, D., & Yao, R. (2020). ShinyGO: A graphical gene-set enrichment tool for animals and plants. *Bioinformatics, 36*(8), 2628–2629.

Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., & Knight, R. (2018). Current understanding of the human microbiome. *Nature Medicine, 24*(4), 392–400.

Gilmore, W. J. (Ed.). (2008). SQLite. In Beginning PHP and MySQL: From Novice to Professional. *Apress*. (pp. 567–590).

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics, 17*(6), 333–351.

Grivennikov, S. I., Greten, F. R., & Karin, M. (2010). Immunity, inflammation, and cancer. *Cell, 140*(6), 883–899.

Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., Bot, B. M., Morris, J. S., Simon, I. M., Gerster, S., Fessler, E., De Sousa E Melo, F., Missiaglia, E., Ramay, H., Barras, D., Tejpar, S. (2015). The consensus molecular subtypes of colorectal cancer. *Nature Medicine, 21*(11), 1350–1356.

Häfner, M. F., & Debus, J. (2016). Radiotherapy for colorectal cancer: Current standards and future perspectives. *Visceral Medicine, 32*(3), 172–177.

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell, 144*(5), 646–674.

Han, Y. W. (2015). Fusobacterium nucleatum: A commensal-turned pathogen. *Current Opinion in Microbiology, 23*, 141–147.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis,

S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Gene Ontology Consortium. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research, 32* (Database issue), D258–D261.

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., & Hussain, A. (2024). Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation, 16*(1), 45–74.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.) *Springer.*

Henke, M. T., Kenny, D. J., Cassilly, C. D., Vlamakis, H., Xavier, R. J., & Clardy, J. (2019). Ruminococcus gnavus, a member of the human gut microbiome associated with Crohn's disease, produces an inflammatory polysaccharide. *Proceedings of the National Academy of Sciences, 116*(26), 12672–12677.

He, T., Cheng, X., & Xing, C. (2021). The gut microbial diversity of colon cancer patients and the clinical significance. *Bioengineered, 12*(1), 7046–7060.

Homola, D. (2017). Integration and visualisation of clinical-omics datasets for medical knowledge discovery. Imperial College London. https://danielhomola.com/assets/DanielHomola_PhD.pdf.

Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. *Nature, 486*(7402), 207–214.

Hu, W., Yang, Y., Li, X., Huang, M., Xu, F., Ge, W., Zhang, S., & Zheng, S. (2018). Multi-omics approach reveals distinct differences in left- and right-sided colon cancer. *Molecular Cancer Research: MCR, 16*(3), 476–485.

IARC. (2019). Colorectal cancer. In *Colorectal cancer screening*. International Agency for Research on Cancer.

Inamura, K. (2018). Colorectal cancers: An update on their molecular pathology. *Cancers, 10*(1). https://doi.org/10.3390/cancers10010026.

Jiang, Y., Yan, X., Liu, K., Shi, Y., Wang, C., Hu, J., Li, Y., Wu, Q., Xiang, M., & Zhao, R. (2020). Discovering the molecular differences between right- and left-sided colon cancer using machine learning methods. *BMC Cancer, 20*(1), 1012.

Johnson, S.C. (1967). Hierarchical clustering schemes. *Psychometrika, 32*(3), 241-254. https://doi.org/10.1007/BF02289588.

Jolliffe, I. T. (2013). Principal component analysis (1986th ed.) [PDF]. Springer. https://doi.org/10.1007/978-1-4757-1904-8.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255–260.

Kanehisa, M., & Goto, S. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes.

Keshavarz-Rahaghi, F., Pleasance, E., Kolisnik, T., & Jones, S. J. M. (2022). A p53 transcriptional signature in primary and metastatic cancers derived using machine learning. *Frontiers in Genetics, 13*, 987238.

Kolisnik, T. (2022). Rf2pval: R package for obtaining p-values and cutoffs for features in random forest models. Github. https://github.com/tkolisnik/Rf2pval.

Kolisnik, T., Sulit, A. K., Schmeier, S., Frizelle, F., Purcell, R., Smith, A., & Silander, O. (2023). Identifying important microbial and genomic biomarkers for differentiating right- versus left-sided colorectal cancer using random forest models. *BMC Cancer, 23*(647), 1-11.

Kolisnik, T., Keshavarz-Rahaghi, F., Purcell, R., Smith, A., & Silander, O. (2024).

pyRforest: A comprehensive R package for genomic data analysis featuring scikit-learn Random Forests in R. *Briefings in Functional Genomics, 2024*(38), 1-9. https://doi.org/10.1093/bfgp/elae038.

Köster, J., & Rahmann, S. (2012). Snakemake a scalable bioinformatics workflow engine. *Bioinformatics, 28*(19), 2520–2522.

Kostic, A. D., Chun, E., Meyerson, M., & Garrett, W. S. (2013). Microbes and inflammation in colorectal cancer. *Cancer Immunology Research, 1*(3), 150–157.

Kostouros, A., Koliarakis, I., Natsis, K., Spandidos, D. A., Tsatsakis, A., & Tsiaoussis, J. (2020). Large intestine embryogenesis: Molecular pathways and related disorders (Review). *International Journal of Molecular Medicine, 46*(1), 27–57.

Kotthaus, H., Korb, I., Lang, M., Bischl, B., Rahnenführer, J., & Marwedel, P. (2015). Runtime and memory consumption analyses for machine learning R programs. *Journal of Statistical Computation and Simulation, 85*(1), 14–29.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal, 13*, 8–17.

Kuipers, E. J., Grady, W. M., Lieberman, D., Seufferlein, T., Sung, J. J., Boelens, P. G., van de Velde, C. J. H., & Watanabe, T. (2015). Colorectal cancer. *Nature Reviews. Disease Primers, 1*, 15065.

Kumar, A., Gautam, V., Sandhu, A., Rawat, K., Sharma, A., & Saha, L. (2023). Current and emerging therapeutic approaches for colorectal cancer: A comprehensive review. *World Journal of Gastrointestinal Surgery, 15*(4), 495–519.

Lapidus, A. L., & Korobeynikov, A. I. (2021). Metagenomic data assembly - The way of

decoding unknown microorganisms. *Frontiers in Microbiology, 12*, 613791.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

Lee, G. H., Malietzis, G., Askari, A., Bernardo, D., Al-Hassi, H. O., & Clark, S. K. (2015). Is right-sided colon cancer different to left-sided colorectal cancer? A systematic review. *European Journal of Surgical Oncology: The Journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology, 41*(3), 300–308.

Lee, M. S., Menter, D. G., & Kopetz, S. (2017). Right versus left colon cancer biology: Integrating the consensus molecular subtypes. *Journal of the National Comprehensive Cancer Network: JNCCN, 15*(3), 411–419.

Liang, L., Zeng, J.-H., Qin, X.-G., Chen, J.-Q., Luo, D.-Z., & Chen, G. (2018). Distinguishable prognostic signatures of left- and right-sided colon cancer: A study based on sequencing data. *Cellular Physiology and Biochemistry: International Journal of Experimental Cellular Physiology, Biochemistry, and Pharmacology, 48*(2), 475–490.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. In *R News* (Vol. 2, Issue 3, pp. 18–22). https://CRAN.R-project.org/doc/Rnews/.

Li, B., Huang, Q., & Wei, G. H. (2019). The role of HOX transcription factors in cancer predisposition and progression. *Cancers, 11*(4). https://doi.org/10.3390/cancers11040528.

Liu, L. U., Holt, P. R., Krivosheyev, V., & Moss, S. F. (1999). Human right and left colon differ in epithelial cell apoptosis and in expression of Bak, a pro-apoptotic Bcl-2 homologue. *Gut, 45*(1), 45–50.

Liu, W., Zhang, X., Xu, H., Li, S., Lau, H. C.-H., Chen, Q., Zhang, B., Zhao, L., Chen, H., Sung, J. J.-Y., & Yu, J. (2021). Microbial community heterogeneity within colorectal neoplasia and its correlation with colorectal carcinogenesis. *Gastroenterology, 160*(7),

2395–2408.

Lloyd, S.P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*(2), 129-137. https://doi.org/10.1109/TIT.1982.1056489.

Lucas, C., Barnich, N., & Nguyen, H. T. T. (2017). Microbiota, inflammation and colorectal cancer. *International Journal of Molecular Sciences, 18*(6). https://doi.org/10.3390/ijms18061310.

Ludt, A., Ustjanzew, A., Binder, H., Strauch, K., & Marini, F. (2022). Interactive and reproducible workflows for exploring and modeling RNA-seq data with pcaExplorer, Ideal, and GeneTonic. *Current Protocols, 2*(4), e411.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777.

Luo, Z., Rhie, S. K., & Farnham, P. J. (2019). The enigmatic HOX genes: Can we crack their code? *Cancers, 11*(3). https://doi.org/10.3390/cancers11030323.

Marini, F., & Binder, H. (2019). pcaExplorer: an R/Bioconductor package for interacting with RNA-seq principal components. *BMC Bioinformatics, 20*(1), 331.

Martin, F. H. (1914). *Surgery, Gynecology & Obstetrics.* Franklin H. Martin Memorial Foundation.

Martin, F. J., Amode, M. R., Aneja, A., Austine-Orimoloye, O., Azov, A. G., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., Bhurji, S. K., Bignell, A., Boddu, S., Branco Lins, P. R.,

Brooks, L., Ramaraju, S. B., Charkhchi, M., Cockburn, A., Da Rin Fiorretto, L., Flicek, P. (2023). Ensembl 2023. *Nucleic Acids Research, 51*(D1), D933–D941.

Mayo Clinic. (2022). Colon Cancer. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/colon-cancer/symptoms-causes/syc-20353669.

McMurdie, P. J., & Holmes, S. (2012). Phyloseq: A bioconductor package for handling and analysis of high-throughput phylogenetic sequence data. *Pacific Symposium on Biocomputing*, 235–246.

Menzel, P., Ng, K. L., & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications, 7*(1), 11257.

Metzker, M. L. (2010). Sequencing technologies - The next generation. *Nature Reviews Genetics, 11*(1), 31–46.

Miniati, R., Iadanza, E., & Dori, F. (2016). *Clinical Engineering: From Devices to Systems.* Elsevier, Inc.

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., & Köster, J. (2021). Sustainable data analysis with Snakemake. *F1000Research, 10*, 33.

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Random forest for genomic prediction. In *Multivariate Statistical Machine Learning Methods for Genomic Prediction* (pp. 633–681). Springer International Publishing.

Mouillet-Richard, S., Cazelles, A., Sroussi, M., Gallois, C., Taieb, J., & Laurent-Puig, P. (2024). Clinical challenges of consensus molecular subtype CMS4 colon cancer in the era of

precision medicine. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research, 30*(11), 2351–2358.

Mukund, K., Syulyukina, N., Ramamoorthy, S., & Subramaniam, S. (2020). Right and left-sided colon cancers - Specificity of molecular mechanisms in tumorigenesis and progression. *BMC Cancer, 20*(1), 317.

Nagai, Y., Kiyomatsu, T., Gohda, Y., Otani, K., Deguchi, K., & Yamada, K. (2021). The primary tumor location in colorectal cancer: A focused review on its impact on surgical management. *Global Health & Medicine, 3*(6), 386–393.

Narayanan, S., Gabriel, E., Attwood, K., Boland, P., & Nurkin, S. (2018). Association of clinicopathologic and molecular markers on stage-specific survival of right versus left colon cancer. *Clinical Colorectal Cancer, 17*(4), e671–e678.

Natsume, S., Yamaguchi, T., Takao, M., Iijima, T., Wakaume, R., Takahashi, K., Matsumoto, H., Nakano, D., Horiguchi, S.-I., Koizumi, K., & Miyaki, M. (2018). Clinicopathological and molecular differences between right-sided and left-sided colorectal cancer in Japanese patients. *Japanese Journal of Clinical Oncology, 48*(7), 609–618.

NCBI. (2021). BioProject PRJNA788974. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA788974.

Okumura, S., Konishi, Y., Narukawa, M., Sugiura, Y., Yoshimoto, S., Arai, Y., Sato, S., Yoshida, Y., Tsuji, S., Uemura, K., Wakita, M., Matsudaira, T., Matsumoto, T., Kawamoto, S., Takahashi, A., Itatani, Y., Miki, H., Takamatsu, M., Obama, K., Hara, E. (2021). Gut bacteria identified in colorectal cancer patients promote tumourigenesis via butyrate secretion. *Nature Communications, 12*(1), 5674.

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv.*

http://arxiv.org/abs/1511.08458.

Owens, C. L., Epstein, J. I., & Netto, G. J. (2007). Distinguishing prostatic from colorectal adenocarcinoma on biopsy samples: The role of morphology and immunohistochemistry. *Archives of Pathology & Laboratory Medicine, 131*(4), 599–603.

Paschke, S., Jafarov, S., Staib, L., Kreuser, E.-D., Maulbecker-Armstrong, C., Roitman, M., Holm, T., Harris, C. C., Link, K.-H., & Kornmann, M. (2018). Are colon and rectal cancer two different tumor entities? A proposal to abandon the term colorectal cancer. *International Journal of Molecular Sciences, 19*(9), 2577.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., & Cournapeau, D. (2011). Scikit-learn: Machine learning in Python. *Machine Learning In Python*, 6.

Pellegrino, E., Jacques, C., Beaufils, N., Nanni, I., Carlioz, A., Metellus, P., & Ouafik, L. (2021). Machine learning random forest for predicting oncosomatic variant NGS analysis. *Scientific Reports, 11*(1), 21820.

Peters, B. A., Wilson, M., Moran, U., Pavlick, A., Izsak, A., Wechter, T., Weber, J. S., Osman, I., & Ahn, J. (2019). Relating the gut metagenome and metatranscriptome to immunotherapy responses in melanoma patients. *Genome Medicine, 11*(1), 61.

Python Language. (2019). *Python language: The Python language reference (Version 3.8.0)*. Python.org. https://docs.python.org/3/reference/.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106. https://doi.org/10.1007/BF00116251.

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Rebersek, M. (2021). Gut microbiome and its role in colorectal cancer. *BMC Cancer, 21*(1), 1325. https://doi.org/10.1186/s12885-021-09085-9.

Reimand, J., Kull, M., Peterson, H., Hansen, J., & Vilo, J. (2007). *g:Profiler - A web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Research, 35*, W193–W200. https://doi.org/10.1093/nar/gkm226.

Robinson M.D., McCarthy, D.J., Smyth, G.K. (2010). edgeR. A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics, 26*(1): 139-40. doi: 10.1093/bioinformatics/btp616.

Rohlke, F., & Stollman, N. (2012). Fecal microbiota transplantation in relapsing Clostridium difficile infection. *Therapeutic Advances in Gastroenterology, 5*(6), 403–420. https://doi.org/10.1177/1756283X12453637.

Rosendahl Huber, A., Pleguezuelos-Manzano, C., Puschhof, J., Ubels, J., Boot, C., Saftien, A., Verheul, M., Trabut, L. T., Groenen, N., van Roosmalen, M., Ouyang, K. S., Wood, H., Quirke, P., Meijer, G., Cuppen, E., Clevers, H., & van Boxtel, R. (2024). Improved detection of colibactin-induced mutations by genotoxic E. coli in organoids and colorectal cancer. *Cancer Cell, 42*(3), 487–496.e6. https://doi.org/10.1016/j.ccell.2023.06.010.

R Special Interest Group on Databases (R-SIG-DB), Wickham, H., & Müller, K. (2024). *DBI: R database interface.* https://dbi.r-dbi.org.

RStudio Team. (2020). *RStudio: Integrated development environment for R.* RStudio, PBC. http://www.rstudio.com/.

Rubinstein, M. R., Wang, X., Liu, W., Hao, Y., Cai, G., & Han, Y. W. (2013). Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/β-catenin

signaling via its FadA adhesin. *Cell Host & Microbe, 14*(2), 195–206. https://doi.org/10.1016/j.chom.2013.07.012.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics, 23*(19), 2507–2517. https://doi.org/10.1093/bioinformatics/btm344.

Sánchez-Alcoholado, L., Ramos-Molina, B., Otero, A., Laborda-Illanes, A., Ordóñez, R., Medina, J. A., Gómez-Millán, J., & Queipo-Ortuño, M. I. (2020). The role of the gut microbiome in colorectal cancer development and therapy response. *Cancers, 12*(6). https://doi.org/10.3390/cancers12061406.

Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., Thakare, R. P., Banday, S., Mishra, A. K., Das, G., & Malonia, S. K. (2023). Next-generation sequencing technology: Current trends and advancements. *Biology, 12*(7), 997. https://doi.org/10.3390/biology12070997.

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., Sherry, S. T. (2022). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research, 50*(D1), D20–D26. https://doi.org/10.1093/nar/gkab1112.

Schauberger, P., & Walker, A. (2024). *openxlsx: Read, write and edit xlsx files.* https://ycphs.github.io/openxlsx/index.html.

scikit-learn. (2022). *Cross-validation: Evaluating estimator performance.* https://scikit-learn.org/stable/modules/cross_validation.html.

Sharpton, T. J. (2014). An introduction to the analysis of shotgun metagenomic data. *Frontiers*

in Plant Science, 5, 209. https://doi.org/10.3389/fpls.2014.00209.

Sherstinsky, A. (2018). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *arXiv*. https://doi.org/10.48550/arXiv.1808.03314.

Stintzing, S., Tejpar, S., Gibbs, P., Thiebach, L., & Lenz, H.-J. (2017). Understanding the role of primary tumour localisation in colorectal cancer treatment and outcomes. *European Journal of Cancer, 84*, 69–80. https://doi.org/10.1016/j.ejca.2017.07.016.

Strimbu, K., & Tavel, J. A. (2010). What are biomarkers? *Current Opinion in HIV and AIDS, 5*(6), 463–466. https://doi.org/10.1097/COH.0b013e32833ed177.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America, 102*(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102.

Sulit, A. K., Kolisnik, T., Frizelle, F. A., Purcell, R., & Schmeier, S. (2023). MetaFunc: Taxonomic and functional analyses of high throughput sequencing for microbiomes. *Gut Microbiome, 4*(4), 1-21. https://doi.org/10.1089/gutmicro.2023.0004.

Talukder, A., Barham, C., Li, X., & Hu, H. (2021). Interpretation of deep learning in genomics and epigenomics. *Briefings in Bioinformatics, 22*(3). https://doi.org/10.1093/bib/bbaa177.

Thursby, E., & Juge, N. (2017). Introduction to the human gut microbiota. *The Biochemical Journal, 474*(11), 1823–1836. https://doi.org/10.1042/BCJ20170158.

Ushey, K., Allaire, J. J., & Tang, Y. (2024). *reticulate: Interface to Python.* https://rstudio.github.io/reticulate/.

Wang, F., Cai, K., Xiao, Q., He, L., Xie, L., & Liu, Z. (2022). Akkermansia muciniphila administration exacerbated the development of colitis-associated colorectal cancer in mice. *Journal of Cancer, 13*(1), 124–133. https://doi.org/10.7150/jca.57932.

Wang, N., & Fang, J.-Y. (2022). Fusobacterium nucleatum, a key pathogenic factor and microbial biomarker for colorectal cancer. *Trends in Microbiology.* https://doi.org/10.1016/j.tim.2022.08.010.

Wang, X., Zhou, J., Xu, M., Yan, Y., Huang, L., Kuang, Y., Liu, Y., Li, P., Zheng, W., Liu, H., & Jia, B. (2018). A 15-lncRNA signature predicts survival and functions as a ceRNA in patients with colorectal cancer. *Cancer Management and Research, 10,* 5799–5806. https://doi.org/10.2147/CMAR.S172273.

Warnes, G. R., Bolker, B., Bonebakker, L., & Gentleman, R. (2009). gplots: Various R programming tools for plotting data. R Package Version.

Waskom, M. (2021). seaborn: Statistical data visualization. *The Journal of Open Source Software, 6,* 3021. https://doi.org/10.21105/joss.03021.

WCRF International. (2022, February 15). Colorectal cancer. https://www.wcrf.org/diet-activity-and-cancer/cancer-types/colorectal-cancer/.

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer-Verlag New York.

Wickham, H. (2019). dplyr: A grammar of data manipulation. *dplyr.* https://dplyr.tidyverse.org/.

Wickham, H., Hester, J., Chang, W., & Bryan, J. (2022). devtools: Tools to make developing R packages easier.

Wu, Y., Jiao, N., Zhu, R., Zhang, Y., Wu, D., Wang, A.-J., Fang, S., Tao, L., Li, Y., Cheng, S., He, X., Lan, P., Tian, C., Liu, N.-N., & Zhu, L. (2021). Identification of microbial

markers across populations in early detection of colorectal cancer. *Nature Communications, 12*(1), 3063. https://doi.org/10.1038/s41467-021-23315-0.

Xie, B., Bai, B., Xu, Y., Liu, Y., Lv, Y., Gao, X., Wu, F., Fang, Z., Lou, Y., Pan, H., & Han, W. (2019). Tumor-suppressive function and mechanism of HOXB13 in right-sided colon cancer. *Signal Transduction and Targeted Therapy, 4,* 51. https://doi.org/10.1038/s41392-019-0072-y.

Yang, J., Du, X. L., Li, S. T., Wang, B. Y., Wu, Y. Y., Chen, Z. L., Lv, M., Shen, Y. W., Wang, X., Dong, D. F., Li, D., Wang, F., Li, E. X., Yi, M., & Yang, J. (2016). Characteristics of differently located colorectal cancers support proximal and distal classification: A population-based study of 57,847 patients. *PloS One, 11*(12), e0167540. https://doi.org/10.1371/journal.pone.0167540.

Xie, Y., Cheng, J., & Tan, X. (2022). DT: A wrapper of the JavaScript library "DataTables." https://github.com/rstudio/DT.

Yang, J., Du, X. L., Li, S. T., Wang, B. Y., Wu, Y. Y., Chen, Z. L., Lv, M., Shen, Y. W., Wang, X., Dong, D. F., Li, D., Wang, F., Li, E. X., Yi, M., & Yang, J. (2016). Characteristics of differently located colorectal cancers support proximal and distal classification: A population-based study of 57,847 patients. *PloS One, 11*(12), e0167540. https://doi.org/10.1371/journal.pone.0167540.

Yang, J., Feng, E., Ren, Y., Qiu, S., Zhao, L., & Li, X. (2021). Long non-coding (lnc)RNA profiling and the role of a key regulator lnc-PNRC2-1 in the transforming growth factor-β1-induced epithelial-mesenchymal transition of CNE1 nasopharyngeal carcinoma cells. *The Journal of International Medical Research, 49*(3), 300060521996515. https://doi.org/10.1177/0300060521996515.

Yang, J., McDowell, A., Kim, E. K., Seo, H., Lee, W. H., Moon, C.-M., Kym, S.-M., Lee, D. H., Park, Y. S., Jee, Y.-K., & Kim, Y.-K. (2019). Development of a colorectal cancer diagnostic model and dietary risk assessment through gut microbiome analysis. *Experimental & Molecular Medicine, 51*(10), 1–15. https://doi.org/10.1038/s12276-019-0252-2.

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: An R package for comparing biological themes among gene clusters. *Omics: A Journal of Integrative Biology, 16*(5), 284–287. https://doi.org/10.1089/omi.2011.0118.

Yu, H., Li, X.-X., Han, X., Chen, B.-X., Zhang, X.-H., Gao, S., Xu, D.-Q., Wang, Y., Gao, Z.-K., Yu, L., Zhu, S.-L., Yao, L.-C., Liu, G.-R., Liu, S.-L., & Mu, X.-Q. (2023). Fecal microbiota transplantation inhibits colorectal cancer progression: Reversing intestinal microbial symbiosis to enhance anti-cancer immune responses. *Frontiers in Microbiology, 14,* 1126808. https://doi.org/10.3389/fmicb.2023.1126808.

Yu, L., Lu, J., & Du, W. (2024). Tryptophan metabolism in digestive system tumors: Unraveling the pathways and implications. *Cell Communication and Signaling: CCS, 22*(1), 174. https://doi.org/10.1186/s12964-024-01168-9.

Zackular, J. P., Rogers, M. A. M., Ruffin, M. T., 4th, & Schloss, P. D. (2014). The human gut microbiome as a screening tool for colorectal cancer. *Cancer Prevention Research (Philadelphia, Pa.), 7*(11), 1112–1121. https://doi.org/10.1158/1940-6207.CAPR-14-0104.

Zhang, Z., Pan, Y., Zhao, Y., Ren, M., Li, Y., Feng, Y., Lu, G., Zhang, S., Zhao, S., & Wang, M. (2022). LncRNA CASC2 regulates colorectal cancer by modulating miR-145-5p/VEGFA axis. *Cancer Cell International, 22*(1), 1–16. https://doi.org/10.1186/s12935-022-02688-5.

Zhao, Y., Federico, A., Faits, T., Manimaran, S., Monti, S., & Evan Johnson, W. (2020). animalcules: Interactive Microbiome Analytics and Visualization in R. In bioRxiv (p. 2020.05.29.123760). https://doi.org/10.1101/2020.05.29.123760.

Zhao, Z., Wang, D.-W., Yan, N., Pan, S., & Li, Z.-W. (2020). Superior survival in right-sided versus left-sided colon signet ring cell carcinoma. *Scientific Reports, 10*(1), 17900.

Ziemann, M., Eren, Y., & El-Osta, A. (2016). Gene name errors are widespread in the scientific literature. *Genome Biology, 17*(1), 177.

Zou, S., Fang, L., & Lee, M.-H. (2018). Dysbiosis of gut microbiota in promoting the development of colorectal cancer. *Gastroenterology Report, 6*(1), 1–12.

Appendix A

Sulit, A. K., **Kolisnik, T.**, Frizelle, F. A., Purcell, R., & Schmeier, S. (2023). MetaFunc: taxonomic and functional analyses of high throughput sequencing for microbiomes. *Gut Microbiome, 4*(4), 1-21. https://doi.org/10.1017/gmb.2022.12.

This paper has been included as an Appendix with full permissions of all authors to serve as a companion to Chapter 2 of this thesis. Explicit references to the MetaFunc App are found on external page numbers 1, 4, 5, 7, 9, 10, 14, 15, 16 of this Appendix.

Gut Microbiome (2023), **4**, e4, 1–21
doi:10.1017/gmb.2022.12

GUTMICROBIOME

METHODS PAPER

MetaFunc: taxonomic and functional analyses of high throughput sequencing for microbiomes

Arielle Kae Sulit[1,2,*], Tyler Kolisnik[2], Frank Antony Frizelle[1], Rachel Purcell[1] and Sebastian Schmeier[2]

[1]Department of Surgery, University of Otago, Christchurch, New Zealand
[2]School of Natural Sciences, Massey University, Auckland, New Zealand
*Corresponding author. Email: iel_sulit@yahoo.com

(Received 26 July 2022; revised 06 November 2022; accepted 13 December 2022)

Abstract

The identification of functional processes taking place in microbiome communities augment traditional microbiome taxonomic studies, giving a more complete picture of interactions taking place within the community. While there are applications that perform functional annotation on metagenomes or metatranscriptomes, very few of these are able to link taxonomic identity to function or are limited by their input types or databases used. Here we present MetaFunc, a workflow which takes RNA sequences as input reads, and from these (1) identifies species present in the microbiome sample and (2) provides gene ontology annotations associated with the species identified. In addition, MetaFunc allows for host gene analysis, mapping the reads to a host genome, and separating these reads, prior to microbiome analyses. Differential abundance analysis for microbe taxonomies, and differential gene expression analysis and gene set enrichment analysis may then be carried out through the pipeline. A final correlation analysis between microbial species and host genes can also be performed. Finally, MetaFunc builds an R shiny application that allows users to view and interact with the microbiome results. In this paper, we showed how MetaFunc can be applied to metatranscriptomic datasets of colorectal cancer.

Keywords: Metatranscriptomics; microbiome; functional annotation; host correlation

Background

Metagenomic or metatranscriptomic studies of microbiome communities allow for characterisation of functional contributions as well as taxonomic load, by allowing the identification and quantification of genes possibly contributed by the microbial community. The ability to identify functional processes from the microbiome gives a more complete picture of microbe–microbe and/or microbe–host interactions that drive community dynamics (Langille, 2018).

There are existing bioinformatics programmes (Nayfach et al., 2015; Sharma et al., 2015; Silva et al., 2016) that perform functional annotation on metagenomes and metatranscriptomes, but most of these are unable to link taxonomies (the microbes under study) to their respective functional processes. Existing packages with this capacity include PICRUSt and PICRUSt2 (Douglas et al., 2019; Langille et al., 2013), and HUMAnN2 (Franzosa et al., 2018). PICRUSt and PICRUSt2 predict metagenome function by inferring genes present in OTUs based on their phylogenetic similarities to other OTUs with known gene content (Douglas et al., 2019; Langille et al., 2013). However, they do not directly measure the genes involved, but rather rely on 16S gene marker sequences, which, being highly conserved, are useful for the

identification of bacterial genera (Bashiardes et al., 2016; Ternes et al., 2020) and are not present in other microbes aside from Bacteria and Archaea (Ye et al., 2019). Thus 16S based taxonomic identification, and subsequent functional predictions, may be unsuitable for species-level identification, and for recognising other microbes aside from Bacteria and Archaea. HUMAnN2's taxonomic profiling, meanwhile, is reliant on MetaPhlAn2 (Segata et al., 2012; Truong et al., 2015), which uses clade-specific marker genes from reference genomes. Benchmarking efforts by Ye et al. (2019) highlight the limitations of using the MetaPhlAn2 package, and therefore HUMAnN2, which results in relatively lower precision and recall in its classification.

To augment such meta-omic studies, we present here a simple, straight-forward pipeline named MetaFunc, a snakemake workflow (Köster and Rahmann, 2012) that maps function to a microbiome (and optionally host) sample, using RNA sequences as input. MetaFunc uses Kaiju (Menzel et al., 2016) as its main taxonomic classifier. Kaiju uses protein translations of input reads to generate taxonomic profiles. By generating protein-based classifications using metatranscriptomic reads, MetaFunc identifies microbes based on their gene expression, allowing more focus on the functional contributions of microbes. MetaFunc then uses protein accession numbers from Kaiju results to obtain the set of gene ontology (GO) terms associated with the microbiome community. Furthermore, Kaiju outputs provide a direct protein – taxonomy ID relationship that makes it possible for MetaFunc to establish which organisms are contributing to the functional GO terms. MetaFunc also has options for pre-processing of reads before running Kaiju: trimming of input reads with fastp (Chen et al., 2018) can be performed in addition to pre-mapping to a host genome (eg. human) using STAR (Dobin et al., 2013). The unmapped reads following STAR processing are the input used by MetaFunc for microbe identification, while host gene expression information can be obtained from STAR-mapped reads. Thus, MetaFunc allows simultaneous investigation of host and microbe community active functional processes, as well as active host genes and microbes.

Protocol

Workflow

Figure 1 shows the workflow that takes place within MetaFunc. Paired-end and/or single-end sequencing reads are used as input in fasta or fastq format. If trimming and mapping are not enabled, reads are used as input to Kaiju and subsequent microbiome analyses (Figure 1a). If trimming is enabled, reads are

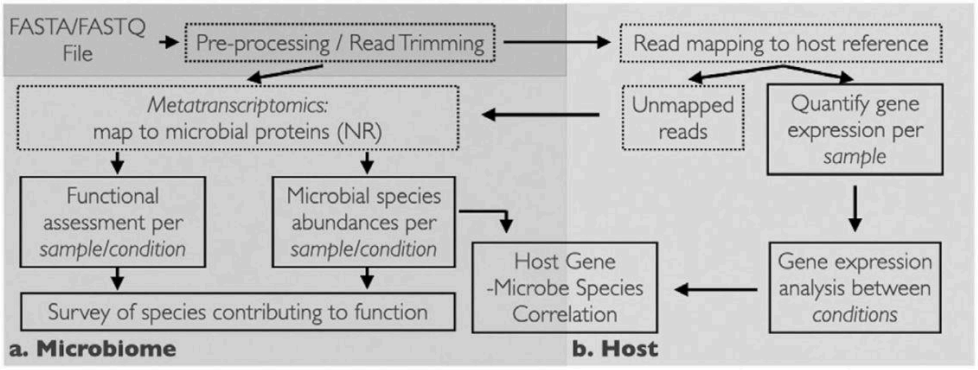


Figure 1. MetaFunc Workflow. The workflow uses FASTQ or FASTA as input and processes reads through the microbiome pipeline to give microbial abundance and function **(a)** and/or host gene analysis **(b)** which will first map reads to a host before sending unmapped reads to the microbiome pipeline. Applying host read analysis will give gene expression analysis results as well as host gene-microbial species correlation. Solid boxes indicate steps with an output while dotted boxes indicate intermediate steps in the pipeline. **NR**: NCBI Blast *nr* database.

https://doi.org/10.1017/gmb.2022.12 Published online by Cambridge University Press

trimmed for adapters and undergo quality controls using fastp. If mapping is enabled, either the trimmed reads or raw input reads are first mapped to a designated host genome using STAR. Unmapped reads after host mapping are then used as input to Kaiju. STAR results are then used to obtain host gene information (Figure 1b).

Microbiome analysis

MetaFunc parses through Kaiju results and gathers taxonomy IDs of species for taxonomic characterisation per sample and their corresponding protein accession numbers, which are subsequently annotated with GO terms (Figure 1a).

Taxonomy

Each classified read matches to a taxonomy ID in Kaiju. MetaFunc gathers the species level matches and adds up the raw reads matching to each species taxonomy ID. In cases of strain level identification, MetaFunc adds this count to its parent species. It also obtains scaled read counts in percentages by dividing the final read count of each taxonomy ID by the total reads that have mapped to species-level taxonomies (then multiplying by 100). For a dataset, the pipeline removes any taxonomy ID that is less than 0.001% in abundance in all samples of the dataset; this filter removes thousands of species that are likely to be false positives while retaining more confident classifications. Any remaining false classifications are thought not to affect downstream analyses, as the levels would be too low to impact true abundance (Ye et al., 2019), however, this value can be adjusted by the user. The taxonomy IDs that have passed the cutoff are then used in subsequent analyses. It should be noted that the pipeline still uses the original scaled percent abundances even after filtering. The pipeline would also include the lineage of the taxonomies using TaxonKit (Shen and Ren, 2021).

For a dataset, the MetaFunc pipeline outputs two tables containing species as rows and samples as columns with values being raw read counts or percent abundance for each species in the samples. If the user wishes to compare groups or conditions (eg. disease state vs. control), the pipeline calculates the average percent abundance of species among samples belonging to a group and this table is also given as an output. Differential abundance of microbes between groups is also carried out in MetaFunc using edgeR (McCarthy et al., 2012; Robinson et al., 2010). Raw read count tables are first filtered using the function *filterbyExpr* with threshold of 1, which is user-adjustable, and normalisation factors are calculated by *calcNormFactors* with default settings. *exactTest* is then applied to calculate differential abundance with *p*-values adjusted using Benjamini and Hochberg correction or false discovery rate (FDR).

Proteins

Kaiju outputs the accession number(s) of the protein match(es) with the highest BLOSUM62 alignment score of the read after translation into six open reading frames (ORF). It is possible to have more than one best protein match if two or more protein matches have equal scores in Kaiju. In order to account for this, we use proportional read counts per protein accession number where one read is divided by the number of best protein matches it has. Similar to that for taxonomy IDs, the pipeline adds up the proportional read counts per protein accession number of a species. Scaled reads as percent abundances are obtained by dividing the proportional count of each accession number by the total read counts that have mapped to a species (then multiplying by 100).

GO: database construction

MetaFunc relies on Kaiju's nr_euk database for its taxonomic identification and corresponding protein matches. The nr_euk database is built on a subset from NCBI BLAST *nr* database containing Archaea, Bacteria, Fungi, Viruses, and other Microbial Eukaryotes (see https://raw.githubusercontent.com/bioinformatics-centre/kaiju/master/util/kaiju-taxonlistEuk.tsv). Identical sequences in the *nr* database are

https://doi.org/10.1017/gmb.2022.12 Published online by Cambridge University Press

compiled into one entry and Kaiju only outputs the first protein accession number of an entry that has multiple identical sequences (Menzel et al., 2016). Thus, we needed to construct the protein-to-GO database such that all functional terms of any protein compiled in one *nr* entry are considered.

To facilitate GO annotations, we constructed an sqlite database in which GO annotations of a protein accession number from Kaiju can be looked up. We first gathered relevant NCBI *nr* database entries, converted all of the proteins of an *nr* entry into UniProt (Huang et al., 2011; The UniProt Consortium, 2017) entries, and then gathered corresponding GO annotations using the Gene Ontology Annotation (GOA) database for all those proteins (Camon et al., 2004). All GO annotations of one *nr* entry are then linked to the first protein of that entry in an sqlite database, which is used to annotate Kaiju protein accession matches with GO IDs. For more detailed information, please see the Notes section of the pipeline's documentation page (https://metafunc.readthedocs.io/en/latest/notes.html). For MetaFunc, we provide pre-made databases for download (Sulit et al., 2021a, 2021b) but users can make their own updated databases following instructions from https://gitlab.com/schmeierlab/metafunc/metafunc-nrgo.git.

GO: protein annotation

For each sample, the pipeline obtains only the proteins that are from taxonomy IDs that passed cutoffs in the section "Taxonomy" described above. Their scaled proportional read counts, as in the section "Proteins" above, are still scaled against the total number of reads that mapped to a species. In order to compare groups or conditions, the pipeline first calculates the average of the corresponding proportional reads and scaled proportional reads of a protein accession number among samples of a group. It then searches for the GO terms annotating the (*nr*) protein using the created sqlite database described in *GO: Database Construction*. Each GO term set annotating an accession number is then updated by accessing parent terms related to the GO terms by "*is_a*" or "*part_of*" using *GOATOOLS* (Klopfenstein et al., 2018). Note that this update takes the entire set of GOs annotating the accession number into consideration such that no GO terms or path/s to the top of the GO directed acyclic graph (DAG) is doubled. GOATOOLS also parses other information regarding the go term such as description, namespace, and depth through the go-basic.obo file (Ashburner et al., 2000). The proportional and scaled read counts are then added to all GO terms annotating a protein, including updated terms. Finally, the percentage of reads covering a GO term within a namespace (Biological Process, Molecular Function, and Cellular Component) is calculated by dividing the scaled read count of a GO term by the total scaled read counts covering a namespace and multiplying by 100. The final output table of the pipeline is a contingency table with GO IDs of all namespaces as rows and samples or groups as columns, with percentage within a namespace as values.

Visualisation

To facilitate the exploration of results from MetaFunc, MetaFunc automatically builds an R shiny application, such that users can view and interact with the taxonomy and GO tables. The application allows users to select GO terms and identify the species whose proteins are annotated with the searched for term. Conversely, users may search for a species and obtain all GO terms associated with the searched for species. See the pipeline's documentation page for more information (https://metafunc.readthedocs.io/en/latest/rshiny.html).

Host analyses

Many microbiome communities are often associated with a host genome (Figure 1b). Reads belonging to the host genome have the capacity to misclassify as microbiome (Ye et al., 2019) and filtering of host reads has been a part of many microbiome studies, either prior to sequencing or *in silico* (Hugerth and Andersson, 2017; Macklaim and Gloor, 2018; Xia et al., 2018). The MetaFunc pipeline offers the option

https://doi.org/10.1017/gmb.2022.12 Published online by Cambridge University Press

of mapping reads to a host genome using the programme STAR and using the unmapped reads from this step as input to Kaiju for the microbiome analysis.

MetaFunc also allows additional analyses of host reads after STAR mapping. Host genes are quantified using featureCounts (Liao et al., 2014) of the subread package. If comparisons between groups are indicated, edgeR is used to perform differential gene expression analysis (DGEA). Additionally, supplying a gene matrix transposed (*.gmt*) file from, for example, the molecular signatures database (GSEA, n.d.; Liberzon et al., 2011; Subramanian et al., 2005) allows for gene set enrichment analysis (GSEA) of host genes using the clusterProfiler package (Yu et al., 2012).

Host gene–microbe species correlation

When a comparison between groups is specified, the pipeline also performs Spearman correlation analysis between the top most significant differentially expressed genes (DEGs), expressed as transcript per million (TPM), and top most significant differentially abundant (DA) microbes, expressed as percent abundance. Results of these correlations are summarised in a matrix on which hierarchical clustering is performed and a heatmap is generated using Clustergrammer (Fernandez et al., 2017). Through this heatmap and table, a user can investigate the strength of correlation (*rho*) between a DA microbe and a DEG, and which microbes and genes have similar patterns of correlations.

Tutorial/manual

For a more detailed description of the workflow, usage instructions, and results, documentation of the MetaFunc pipeline may be found at https://metafunc.readthedocs.io/en/latest/index.html.

Illustration of tool use

Dataset PRJNA413956: matched colorectal cancer and adjacent non-tumour tissue

In order to demonstrate the utility of the MetaFunc pipeline, we obtained publicly available transcriptomics data from the study of Li et al. (2018) consisting of 10 tumours and corresponding adjacent non-tumour colorectal tissue samples. Raw sequencing data were downloaded from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE104836 and input to the pipeline and the full workflow carried out, generating data for host, microbiome, and host-microbiome correlation.

Microbiome results

Taxonomy The MetaFunc pipeline outputs a table of percent abundances of species that are identified in each sample and an average of these abundances across members of the same group if a grouping condition is applied. We ran the pipeline with the intent of comparing microbiome species and function between colon cancer samples and non-tumour matched samples.

Previous studies have already established that certain microbes associate more with colorectal cancer (CRC) samples compared to healthy controls. We searched for *Fusobacterium nucleatum*, *Parvimonas micra,* and *Porphyromonas asaccharolytica* in the averaged group results. These microbes have previously been found to be more abundant in CRC cohorts in meta-analyses of several datasets (Dai et al., 2018; Thomas et al., 2019). We also searched for *Bifidobacterium* species, *Bifidobacterium bifidum* and *Bifidobacterium longum*; Bifidobacteria are thought to confer protection from CRC (Wei et al., 2018).

The bars in Figure 2a show the average percent abundance of the species between samples from tumour and matched non-tumour tissue as identified through MetaFunc. As MetaFunc provides a per sample data, we are also able to plot individual values of CRC (red) and matched normal (blue) samples.

As seen in Figure 2a, MetaFunc identified *F. nucleatum, P. micra,* and *P. asaccharolytica* as being relatively more abundant (ie. have higher average percent abundance) in the CRC group while the *Bifidobacterium* species are relatively more abundant in the normal group.

https://doi.org/10.1017/gmb.2022.12 Published online by Cambridge University Press

MetaFunc also has a step that utilises edgeR to perform differential abundance on per sample species read counts, stratified according to CRC and non-tumour grouping. This resulted in a total of 117 species that were significantly different between the groups (FDR < 0.05). There are 59 species upregulated and 58 downregulated in colon cancer samples. Through the MetaFunc results, we identified *Tanerella forsythia* as the most prominent enriched species in the colon cancer cohort with a $\log_2$ FC = 7.40. *T. forsythia* is a known oral pathogen, thought to be part of the so-called Red complex of periodontal pathogens, along with *Porphyromonas gingivalis*, and *Treponema denticola* (Malinowski et al., 2019). Members of this Red Complex have been found to be enriched in subtype CMS1 of CRCs (Purcell et al., 2017), the subtype most associated with immune process activation in CRC (Dienstmann et al., 2017; Guinney et al., 2015; Inamura, 2018).

Function MetaFunc is intended to enable comparisons of the functional potential of the microbiome between groups. MetaFunc uses GO annotations of protein matches from Kaiju. To demonstrate, we focused on polyamine biosynthetic processes GO terms. Polyamines (PAs) are polycations found to play important biological functions in cell growth. These molecules have been found to be associated with tumour progression and growth (Gerner and Meyskens, 2004; Soda, 2011; Tofalo et al., 2019). Although cells are able to biosynthesize polyamines and even export them, a large source of cellular polyamines comes from uptake from their surroundings and, importantly, the microbiota is thought to be an essential source (Soda, 2011; Thomas et al., 2019; Tofalo et al., 2019) with spermidine and putrescine being the most common of bacterial PAs (Tofalo et al., 2019).

The bars in Figure 2b show the percent of reads among biological process GOs covering PA biosynthetic processes in the CRC and Normal conditions, superimposed with the individual values of samples from the CRC (red) and Normal (blue) groups. From Figure 2b, we saw that several of the polyamine biosynthetic processes were relatively more abundant (ie. higher percent of reads among biological process GOs) in the CRC cohort compared to the normal cohort, using protein annotations.

We used the built-in MetaFunc shiny application to facilitate an inquiry into the microbes species that may contribute to polyamine synthesis. To illustrate, we searched for "*polyamine biosynthetic process*" in the "GO to TaxIDs" tab of the application, and obtained a total of 126 TaxIDs contributing to the GO term in both CRC and normal samples. Of these TaxIDs, we identified *Escherichia coli* and *B. fragilis* to be most abundant in both cohorts. However, differences in the relative abundance of some microbial species can be identified between cancer and normal cohorts, notably several of which are oral pathogens from the genus *Prevotella.* A striking difference in abundance was seen in *T. forsythia,* which was previously found to be significantly more abundant in the CRC cohort via edgeR (Figure 2c). These data suggest that *T. forsythia* represents one of the bacterial species that most contributes to increased polyamine synthesis in CRC samples in this cohort.

Host results

The dataset we used for this study was from a total RNA transcriptomics run aiming to identify long non-coding RNAs (lncRNAs) and mRNAs in CRC samples (Li et al., 2018). Therefore, we first mapped the reads to the human genome using the STAR mapping utility of the pipeline, subsequently using only the unmapped reads for the microbiome analyses. From the reads mapped to the human genome, MetaFunc was able to obtain counts of reads covering human genes and using these, obtained DEGs between CRC and matched normal samples through edgeR. MetaFunc results showed a total of 1,476 DEGs with an FDR < 0.05 and $|\log_2 \text{fold change}| > 2$. From these, we found all the top 5 upregulated and top 5 downregulated genes as reported in the source publication (Li et al., 2018), as well as all the genes they had randomly selected for expression confirmation via qPCR. Figure 2d shows their fold change as found through MetaFunc.

MetaFunc is also able to perform host gene set enrichment analysis using the DEGs. Significant gene sets (p.adjust < 0.05) with the highest normalised positive enrichment scores (NES) included such terms as ribosome biogenesis, DNA replication, mitotic nuclear division, and condensed chromosome (see

https://doi.org/10.1017/gmb.2022.12 Published online by Cambridge University Press

https://doi.org/10.1017/gmb.2022.12 Published online by Cambridge University Press

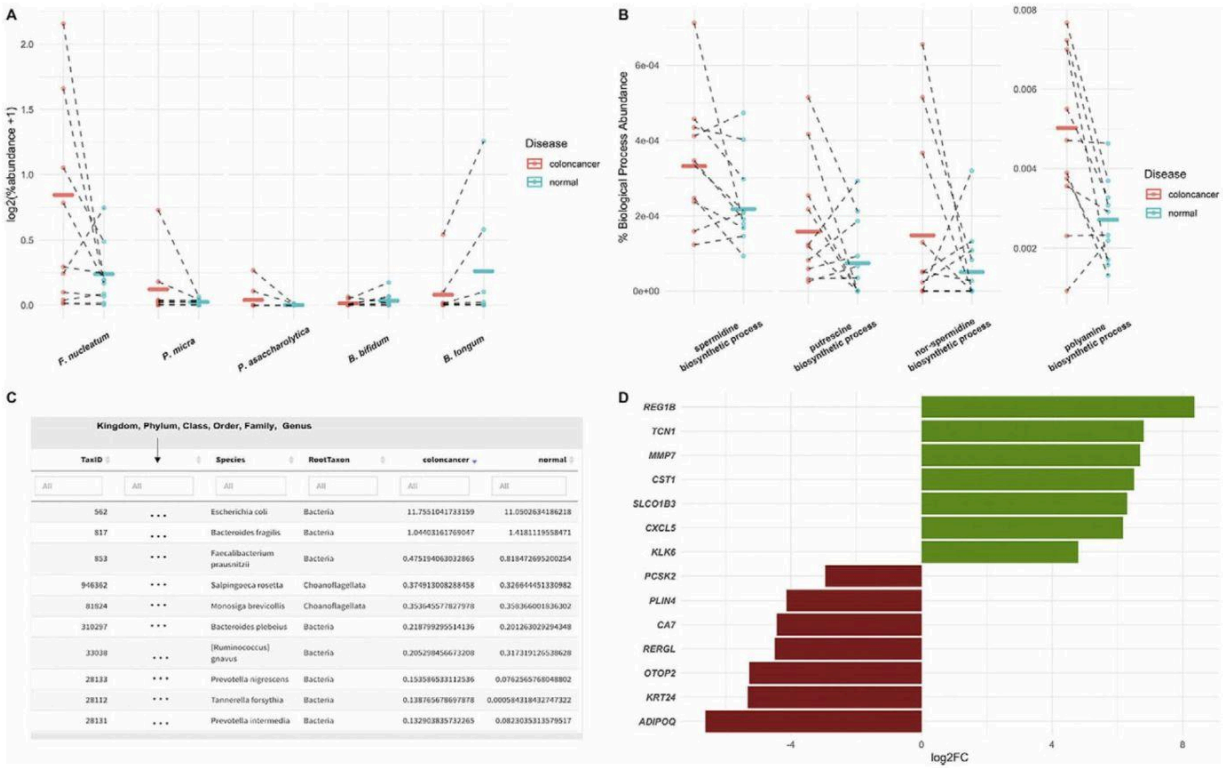


Figure 2. **MetaFunc Microbiome and Host Analyses of Dataset PRJNA413956. (a) Average percent abundance of selected bacterial species in CRC tissue compared to matched non-tumour (normal) samples.** From MetaFunc tabulated results, we plotted the percent abundances of selected bacteria in CRC and matched normal samples. Raw values were first $\log_2$ transformed, with prior addition of 1 as a pseudocount to account for 0 values. Individual points represent per sample transformed values in red (CRC) and blue (Normal). Per group means are represented by the horizontal lines. Dotted lines connect matched CRC and normal samples. **(b) Percent abundance of specific polyamine biosynthetic process GO terms among all biological process GOs in a sample/group compared between CRC (red) and normal (blue) samples.** Values were calculated as described in section "GO: protein annotation" and output in MetaFunc tables or in the R Shiny application. These values were plotted, overlaying group means (horizontal lines) and individual values (data points). **(c) Screenshot from MetaFunc R shiny application.** This view shows the first 10 species with proteins contributing to the GO *polyamine biosynthetic process.* The R Shiny application columns include a URL (not shown in screenshot), which is linked to the NCBI's Taxonomy Browser, the Species Taxonomy ID, Lineage (indicated as "…" in screenshot), Root Taxon, and percent abundances of the species in the two groups being compared: CRC and normal samples. Note that percent abundances refer to the total abundance of the species in question, not just the proteins contributing to the GO term. Results shown are sorted from highest to lowest percent abundance in the colon cancer cohort. **(d) Fold change of representative upregulated and downregulated human genes** (Li et al., 2018) between CRC and matched normal samples in this study. Fold change values were obtained from the edgeR results of the pipeline. All these genes are significant (FDR < 0.05) in both this study and the source publication.

Supplementary Table S1), many of which appear to be related to cell division or replication, consistent with the findings of the source publication (Li et al., 2018), that the upregulated lncRNAs they found were involved in mitosis, cell cycle process, and mitotic cell cycle.

Host–microbiome correlations

We set MetaFunc's default abundance cutoff for microbial identification to 0.001% to remove most probable contaminants and so as not to lose any other meaningful taxonomies. It has been shown in a prior study (Ye et al., 2019), however, that most classifiers call false positives at below 0.01% abundance. We, therefore, applied this 0.01% cutoff in looking at the host–microbiome correlations in this dataset to narrow our focus on microbes that are more likely to be involved in our test case.

In using the 0.01% cutoff, MetaFunc was able to only identify 19 DA microbes. Their correlations with the top 100 significantly abundant genes can be seen at the URL: http://amp.pharm.mssm.edu/clustergrammer/viz/5f02a49e8ec9bb33170b865c/cor.deg-tax.matrix.tsv. Table 1 highlights some notable correlations between DA microbes and differentially expressed human genes. *T. forsythia*, although significantly abundant in CRC samples, do not correlate significantly with any DEGs in CRC. Among its highest correlations, however, included the gene Colorectal Neoplasia Differentially Expressed (*CRNDE*).

Conversely, we investigated which species correlated with *CRNDE*. The highest correlations were with microbes *Candida lusitaniae, Cupriavidus necator,* and *Streptococcus pyogenes*. All correlations were determined to be significant. The same species were among the highest correlations of *TCN1*, and *WNT2*. *TCN1* was among the top DEGs in cancer identified in this study as well as in the source publication (Li et al., 2018). *WNT2* meanwhile is part of the Wnt/β-catenin pathway, which has roles in cell proliferation, cell migration, and cell differentiation. *WNT2* is responsible for the hyperactivation of β-catenin and is known to be upregulated in CRC (Jung et al., 2015).

Dataset PRJNA404030: consensus molecular subtypes of CRC samples

To illustrate MetaFunc's capacity to compare more than two sample groups, we used MetaFunc to analyse transcriptome reads from the study of Purcell and colleagues (Purcell et al., 2017) (raw reads may be accessed at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA404030), which are grouped into four CRC consensus molecular subtypes (CMS). A total of 33 samples were collected during surgical resection of tumours, and sample preparation for RNA sequencing was carried out using the Illumina

Table 1. Spearman correlation between DA microbes and DGEs in CRC.

Gene name	Gene ID	TaxID	Species	rho	*p*-value
CRNDE	ENSG00000245694.10	28112	*Tannerella forsythia*	0.29	0.22
		36911	*Clavispora lusitaniae*	0.70	0.00063
		106590	*Cupriavidus necator*	0.65	0.0019
		1314	*Streptococcus pyogenes*	0.63	0.0027
TCN1	ENSG00000134827.8	106590	*Cupriavidus necator*	0.71	0.00042
		36911	*Clavispora lusitaniae*	0.61	0.0045
		1314	*Streptococcus pyogenes*	0.60	0.0048
WNT2	ENSG00000105989.10	1314	*Streptococcus pyogenes*	0.84	4.07E-06
		106590	*Cupriavidus necator*	0.75	0.00015
		36911	*Clavispora lusitaniae*	0.75	0.00016

https://doi.org/10.1017/gmb.2022.12 Published online by Cambridge University Press

TruSeq Stranded Total RNA Library preparation kit. For these samples, fastq-mcf from ea-utils (Aronesty, 2011, 2013) and SolexaQA++ (Cox et al., 2010) were used to trim reads, which were then run through Salmon (Patro et al., 2017) to quantify transcript expression. The publicly available CRC CMS classifier (Guinney et al., 2015) was used to categorise samples into one of four CMSs. Of the 33 samples, only 27 were classified into a CMS and of these, only one sample was classified into CMS4. This sample was also removed from the dataset for lack of replicates leaving a total of 26 samples – 7 samples in CMS1, 11 in CMS2, and 8 in CMS3. Metafunc was used with default parameters, except for the following options: trimming was set to false, and featureCounts with reverse stranded option was used.

Microbiome results

Taxonomy MetaFunc performed pairwise differential abundance analysis on the three groups using edgeR. From MetaFunc's results, we considered a species to be significantly abundant in a subtype if it is significantly abundant compared to both of the other subtypes. For instance, a significantly abundant species in CMS1 must be significantly abundant in the CMS1 versus CMS2 and CMS1 versus CMS3 comparisons. Using this definition, only CMS1 had species that were significantly abundant (FDR < 0.05) compared to both CMS2 and CMS3. Figure 3a shows the false discovery rate (FDR; diamonds) and $\log_2$ fold change (bars) of the species in CMS1 compared to CMS2 (blue) and CMS3 (brown).

We take note of species in the genera *Prevotella* and *Fusobacterium,* which have previously been associated with CRC. *Fusobacterium nucleatum* in particular has strong evidence of an association with CRC (Dai et al., 2018; Gao et al., 2015; Ye et al., 2017). Most of these are also members of the oral microbiota, which have also previously been associated with cancer development particularly through inflammatory processes (Whitmore and Lamont, 2014). We found no species that were significantly abundant in CMS2 or CMS3 using the given criteria.

Function Through the microbiome functional results of MetaFunc, we then investigated if processes relating to pathogen-associated molecular patterns (PAMPs) were contributed by the microbial communities, considering that CMS1 is characterised by immune responses, which are usually triggered when the human immune system recognises such molecules. We used the MetaFunc R shiny application to search for terms "*lipopolysaccharide biosynthetic process*," "*lipid A biosynthetic process*" and "*peptidoglycan biosynthetic process*," and their relative abundances. Unsurprisingly, all PAMPs were relatively more abundant in CMS1 (Figure 3b).

Using the MetaFunc R shiny application, we also searched for which species might be contributing to the above terms. Figure 3c is a screenshot of the application showing the species contributing to any of the terms in Figure 3b. Figure 3c is arranged from highest to lowest relative abundance in CMS1 and we saw microbes that were among those identified to be significantly abundant in CMS1 such as *F. nucleatum, Hungatella hathewayi, and Prevotella* species. These microbes have previously been associated with CRC (Dai et al., 2018; Gao et al., 2015; Wirbel et al., 2019; Ye et al., 2017).

Host results

Gene set expression analysis MetaFunc calculated DEGs between subtypes in a pairwise manner (ie. CMS1 versus CMS2, CMS1 versus CMS3, CMS2 versus CMS3). From the DEGs of the results, MetaFunc was also able to calculate enriched gene sets for each comparison. Similar to identifying DA microbes, we obtained a final set of enriched gene sets for a subtype if it showed enrichment compared to both other subtypes (p.adjust < 0.05). Unsurprisingly, we saw several host GO terms involved in immune response enriched in CMS1, including regulation of innate immune response, response to interferon gamma, and positive regulation of cytokine production among others. Enriched host GOs in CMS2 are involved in the cell cycle and ribosome biogenesis, with terms such as tRNA metabolic process, ribosomal large subunit biogenesis, and DNA replication initiation, while host GOs enriched in CMS3 involve metabolic processes, for example, primary xenobiotic metabolic process, flavonoid

https://doi.org/10.1017/gmb.2022.12 Published online by Cambridge University Press

https://doi.org/10.1017/gmb.2022.12 Published online by Cambridge University Press

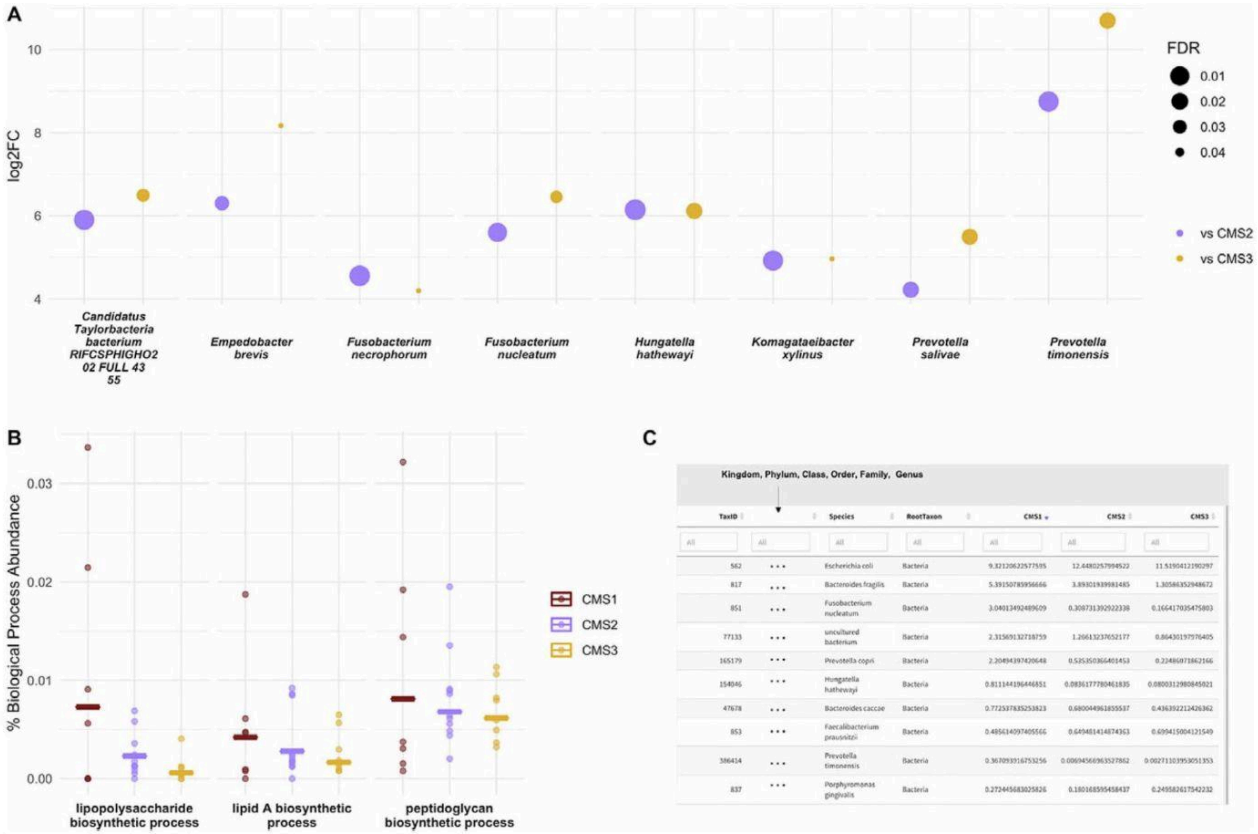


Figure 3. MetaFunc Microbiome Analysis of Dataset PRNJA4040030. (a) Microbes that are significantly more abundant (FDR < 0.05) in CMS1 compared to CMS2 (purple) and CMS3 (yellow). Microbes are considered DA in CMS1 if it is identified through edgeR as DA in both CMS1 versus CMS2 and CMS1 versus CMS3 comparisons. Log_2FC (*y*-axis) is the log_2 of the fold-change between CMS1 and the other subtypes (eg. CMS1/CMS2); FDR (point sizes) is the false discovery rate adjusted *p*-values. **(b) Percent abundance of specific PAMPs biosynthetic process GO terms among all biological process GOs in a sample/group compared between CRC subtypes, CMS1 (red), CMS2 (purple), and CMS3 (yellow).** Values were calculated as described in section "GO: protein annotation" and output in MetaFunc tables or in the R Shiny application. These values were plotted, overlaying group means (horizontal lines) and individual values (data points). **(c) Screenshot of R shiny application showing the relative abundances of species associated with PAMPs biosynthetic processes compared among CMS1, CMS2, and CMS3.** This view shows the first 10 species, with the highest abundances in CMS1, with proteins contributing to any of the PAMPs biosynthetic processes described above. The application columns show a URL (not shown in screenshot), which is linked to the NCBI's Taxonomy Browser, the Species Taxonomy ID, Lineage (shown as "…" in screenshot), Root Taxon, and percent abundances of the species in the three groups being compared: CMS1, CMS2, and CMS3. Note that percent abundances refer to the total abundance of the species in question, not just the proteins contributing to the GO term. Results shown are sorted from highest to lowest percent abundance in the CMS1 group.

metabolic process, and lipid catabolic process. These results are consistent with the description of these three CRC subtypes in the original CMS study (Guinney et al., 2015). The top enriched gene sets for each subtype can be found in Supplementary Tables S2–S7.

Table 2. Spearman correlation between DA microbes in CMS1 and DGEs in CMS1.

Gene name	Gene ID	TaxID	Species	rho	*p*-value
WARS1	ENSG00000140105.18	1802307	*Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*	0.59	0.0015
WARS1	ENSG00000140105.18	851	*Fusobacterium nucleatum*	0.55	0.0035
RNF213	ENSG00000173821.19	1802307	*Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*	0.54	0.0048
ICAM1	ENSG00000090339.9	1802307	*Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*	0.50	0.01
RNF213	ENSG00000173821.19	851	*Fusobacterium nucleatum*	0.50	0.01
PARP14	ENSG00000173193.15	851	*Fusobacterium nucleatum*	0.47	0.02
PARP14	ENSG00000173193.15	1802307	*Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*	0.47	0.02
PARP9	ENSG00000138496.16	851	*Fusobacterium nucleatum*	0.46	0.02
ICAM1	ENSG00000090339.9	851	*Fusobacterium nucleatum*	0.46	0.02
SLC15A3	ENSG00000110446.11	1802307	*Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*	0.46	0.02
STAT1	ENSG00000115415.19	386414	*Prevotella timonensis*	0.44	0.02
CD163	ENSG00000177575.12	1802307	*Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*	0.42	0.03
PARP14	ENSG00000173193.15	386414	*Prevotella timonensis*	0.42	0.03
CD163	ENSG00000177575.12	386414	*Prevotella timonensis*	0.42	0.03
ICAM1	ENSG00000090339.9	386414	*Prevotella timonensis*	0.41	0.04
PARP9	ENSG00000138496.16	1802307	*Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*	0.41	0.04
CD163	ENSG00000177575.12	851	*Fusobacterium nucleatum*	0.41	0.04
SLC15A3	ENSG00000110446.11	851	*Fusobacterium nucleatum*	0.40	0.04
STAT1	ENSG00000115415.19	851	*Fusobacterium nucleatum*	0.40	0.04
PML	ENSG00000140464.20	1802307	*Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*	0.39	0.05
GBP1	ENSG00000117228.10	28448	*Komagataeibacter xylinus*	−0.39	0.05
CEBPA	ENSG00000245848.3	154046	*Hungatella hathewayi*	−0.41	0.04
GNLY	ENSG00000115523.16	28448	*Komagataeibacter xylinus*	−0.43	0.03

https://doi.org/10.1017/gmb.2022.12 Published online by Cambridge University Press

Host–microbiome results

Next, using correlation results from MetaFunc, we investigated which of the top significantly DEGs correlated with the significantly abundant microbes in CMS1. We obtained the following statistically significant correlations between host and microbiome abundances shown in Table 2.

Some of these correlations may be found in http://maayanlab.cloud/clustergrammer/viz/610d8b3c97f268000ea37f41/cor.deg-tax.matrix.tsv. This is the hierarchical cluster obtained when correlating top DA microbes and top DGEs in CMS1 compared to CMS2. It is to be noted that there may be correlations in this clustering that are not found in CMS1 compared to CMS3 and are therefore not reported in Table 2.

The Spearman correlations (rho) between DA microbes and DEGs were quite small in value (the highest value being ~ |0.59| between *WARS1* and *Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*). Nevertheless, several of the genes appeared to have a relevant function with regards to CRC and immune responses. Table 3 shows information for genes that correlated with *Fusobacteria* and *Prevotella* species in our analyses. These two microorganisms have previously been associated with CRC.

Comparison of MetaFunc results to HUMAnN2

HUMAnN2 (Franzosa et al., 2018) is one of the packages most frequently used to assess functional pathways of the microbiome, and to determine which organisms are contributing to the functional pathways. HUMAnN2 works by pre-screening which taxonomies are present in a sample using MetaPhlAn2, afterwards aligning the reads to pangenomes of the classified taxonomies for gene hits. Unclassified reads then undergo an organism-agnostic translated search (Franzosa et al., 2018). MetaPhlAn2 has a rather limited database for the pre-screening of organisms (Ye et al., 2019), resulting in a high level of unmapped reads and a limited number of organisms identified.

We ran the same sequencing reads from the study PRJNA413956 (Li et al., 2018) through HUMAnN2, first trimming with fastp and removing human-mapped reads using the same conditions as for the MetaFunc pipeline. To be more comparable, we changed the pre-screen threshold of HUMAnN2 to 0.001% of mapped reads. Part of HUMAnN2's tiered search uses diamond (Franzosa et al., 2018), which requires higher memory and run time compared to Kaiju, used by MetaFunc (Ye et al., 2019). From taxonomy identification, using Kaiju, to the generation of GO tables, took MetaFunc 11.39 hours to complete, while a comparable analysis using HUMAnN2 took 65.9 hours to complete, almost six times slower than MetaFunc on the same machine (CentOS Linux release 7.9.2009). Notably, HUMAnN2 has an additional pathway abundance and pathway coverage analysis absent from MetaFunc. Runs for HUMAnN2 may be accessed at https://github.com/asulit08/Humann2_PRJNA413956.

Results showed that for the 20 samples analysed, 8.4–22.9% of reads were mapped after nucleotide and protein alignment steps. In contrast, using Kaiju in the MetaFunc pipeline resulted in 33.8–56.2% reads mapped to microbial species through protein matches. We also detected only 87 species across the 20 samples using HUMAnN2, compared with a total of 4,267 species using Kaiju in the MetaFunc pipeline. Further, HUMAnN2 was only able to detect Bacteria and Viruses in the samples, while MetaFunc analysis was able to detect Fungi and Archaea as well. We also investigated the concordance of the microbial GO terms that had been classified to a taxonomy from the MetaFunc run with that of HUMAnN2. We focused on only the Bacteria and Viruses – related GO terms as found in the HUMAnN2 run. We found that the majority (69–100%) of the GOs found in HUMAnN2 was also found in the MetaFunc run. There were more unique GO terms found in the MetaFunc run, which may be due to the higher number of species detected with MetaFunc (Supplementary Figure S1).

We investigated the same species and polyamine (PA) biosynthetic process GO terms in our HUMAnN2 results as we had in the MetaFunc run of dataset PRJNA413956 (Supplementary Figures S2 and S3). We see in Supplementary Figure S2 that abundances of the species in CRC and normal groups have the same trends in HUMAnN2 results as in that of MetaFunc (Kaiju) results. In HUMAnN2 runs, however, we were not able to find *B. bifidum* among the identified species in the

https://doi.org/10.1017/gmb.2022.12 Published online by Cambridge University Press

Table 3. Gene Information of DEGs correlated with DA Microbes in CMS1.

Gene name	Protein name	Relevant protein/gene function	Association with CRC and/or inflammation	Sources
ICAM1	Intercellular adhesion molecule 1	Mediates cell adhesion of cytotoxic T lymphocytes and natural killer cells	Upregulation of ICAM1 inhibits tumour growth and metastasis; a soluble form (sICAM1) is increased in CRC tissues compared to normal, and is associated with an inflammatory tumour microenvironment	Sánchez-Rovira et al., 1998; Schellerer et al., 2019; Tachimori et al., 2005
SLC15A3	Solute carrier (SLC) 15A3	Membrane transporter; highly expressed in macrophage populations	Upregulated by LPS via NF-κβ pathway; influences pro-inflammatory cytokine production triggered by TLR-4	Song et al., 2018; Wang et al., 2014
CD163	CD163 receptor	M2 Macrophage marker	M2 macrophages are anti-inflammatory macrophages and CD163+ tumour-associated macrophages are with mesenchymal transition and poor prognosis in CRC; are correlated with CCL4	Argyle and Kitamura, 2018; Bayoumi et al., 2016; De la Fuente López et al., 2018; Pinto et al., 2019
STAT1	Signal transducer and activator of transcription 1	Transcription factor for IFN signalling	Upregulated in CRCs; correlated with PD-L1 and PD1 immune checkpoint inhibitors; pro-oncogenic in MSI CRCs	Leon-Cabrera et al., 2018; Tanaka et al., 2020
PARP 9	Poly(ADP-ribose) polymerase family member 9	Involved in cell migration	Possible role in metastasis	Vyas and Chang, 2014
PARP 14	Poly(ADP-ribose) polymerase family member 14	Involved in IL-4 signalling and cell migration	Involved in anti-apoptotic effects	Vyas and Chang, 2014
RNF213	Ring finger protein 213	Involved in PI3K-AKT pathway for cell growth	Involved in endothelial angiogenesis	Ohkubo et al., 2015
WARS1	Tryptophanyl-TRNA synthetase 1	Inhibitor of angiogenesis; Involved in IFN-g signalling	Involved in immune responses; cleaved form potentially inhibits angiogenesis; increased levels indicate better CRC survival	Ghanipour et al., 2009; Jin, 2019

PRJNA413956 cohort. Meanwhile, we also see the same trends in the abundances of PA biosynthetic process GO terms in CRC samples compared to matched normal samples in HUMAnN2 runs, as in our MetaFunc run (Supplementary Figure S3), except for nor-spermidine biosynthetic process, which was not seen using HUMAnN2. Differences in abundance values were noted when comparing individual samples, however, direct comparison between HUMAnN2 and MetaFunc is difficult as raw read-counts scaled to species-classified reads are used in MetaFunc, while HUMAnN2 uses reads-per-kilobase (RPK)-based relative abundances.

Discussion

MetaFunc allowed us to investigate the relative abundances of known CRC – associated bacteria between CRC samples and matched normal tissues using the PRJNA413956 dataset. MetaFunc results show that the average abundance of microbes known to contribute to CRC progression are higher in cancer samples while those protective against CRC have higher average abundance in normal samples. Through MetaFunc, we also identified that *Tannerella forsythia*, a known oral pathogen and part of the Red Complex that causes periodontal diseases (Malinowski et al., 2019), is significantly more abundant in CRC tissues than in normal tissues. Oral pathogens have previously been seen to associate with CRC samples (Flemer et al., 2018; Koliarakis et al., 2019; Thomas et al., 2019; Whitmore and Lamont, 2014). By investigating the R shiny application from MetaFunc, we also found that *T. forsythia,* along with bacteria in the *Prevotella* genera, contributed to polyamine biosynthetic processes indicating that some oral pathogens contribute to cancer progression by producing polyamines that could be taken up by the surrounding cells.

Furthermore, we were able to find known bacteria in the MSI-Immune subset of CRCs by identifying the DA microbes in CMS1 compared to both CMS2 and CMS3 subtypes, as identified by MetaFunc's edgeR step. *Fusobacteria* have long been associated with CRC development (Dai et al., 2018; Gao et al., 2015; Thomas et al., 2019; Ye et al., 2017) while *Prevotella* includes species that inhabit the oral cavity; there have also been *Prevotella* species that were found to be abundant in CRC cohorts (Dai et al., 2018; Flemer et al., 2018; Gao et al., 2015). In line with this, PAMPs were also found to be relatively more abundant in the CMS1 cohort upon investigation through MetaFunc's R shiny application. The involvement of these bacteria in CMS1 as well as a relatively higher abundance of proteins contributing to biosynthesis of PAMPs in CMS1 indicate a role of microorganisms in the immune responses that drive the development of CRC in these tumours. This is further supported by correlation with host genes involved in inflammation and/or CRC development as found using MetaFunc's spearman correlation step. The lack of significantly abundant microorganisms in CMS2 and CMS3 may reflect that the CRC development in these subtypes are not as dependent on immune dysregulation.

We created MetaFunc with the aim of identifying microbes and their functional contribution in a microbiome environment. One of the most widely used packages for this is HUMAnN2 (Franzosa et al., 2018) but we find the taxonomic identification generated by HUMAnN2 to be limited, because of its reliance on marker genes. For our purposes, we found MetaFunc invaluable for investigating novel microbes that did not have marker gene representation, in addition to being faster for larger amounts of data, and compatible with downstream analysis programs. We showed in this paper that results from the pipeline are biologically meaningful and corroborate previous literature. It was meant to be an alternative or a complement to HUMAnN2 in this regard. Although similar trends were seen in taxa and gene ontologies of interest between CRC and matched normal samples, fewer test reads were designated as taxa using HUMAnN2 compared to MetaFunc in our comparative analysis. Unfortunately, direct comparison was not possible because HUMAnN2 and MetaFunc use different abundance outputs.

We acknowledge that, especially at the 0.001% abundance cutoff, some of these species we are seeing could be false positives, or that these could be contaminants from sequencing and processing kits used

https://doi.org/10.1017/gmb.2022.12 Published online by Cambridge University Press

(Goffau et al., 2018; Salter et al., 2014). We would caution users in interpreting data from microbes of very low abundances and would recommend following the advice of including negative control samples in sequencing (Salter et al., 2014). Indeed we could be seeing these effects upon looking at the microbes correlating with significantly abundant host genes in CRC samples from PRJNA413956. While *C. lusitaniae* is an opportunistic pathogen causing candidemia (Desnos-Ollivier et al., 2011; Krcmery et al., 1999) possibly exploiting the lowered immune responses in cancer patients (Aslani et al., 2018), and some *Streptococcus* species have previously been implicated in CRC (Kumar et al., 2017; Xia et al., 2020), with *S. pyogenes* having been known to cause invasive infections in humans (Parks et al., 2015), *C. necator* (formerly known as *Ralstonia eutropha* (Reinecke and Steinbüchel, 2009), is a soil bacterium that may be a sequencing contaminant in this dataset. *Cupriavidus* and *Ralstonia* species have been previously identified as common contaminants in meta-omics studies (Guo et al., 2019; Salter et al., 2014).

MetaFunc analyses host and microbiome reads, providing a user-friendly, interactive R-shiny application to investigate results, most useful for those with candidate microbes and function in mind, or for exploratory analyses of the characteristics of a user's dataset. It should be noted that these values are based on raw counts and percent abundances. Microbiome datasets are considered compositional (Gloor et al., 2017; Gloor and Reid, 2016), and this should be taken into consideration during further analysis. We reiterate that values shown in the shiny application (eg. average of microbial relative abundances within a group), are to be used as initial comparisons and description of the data, and care should be taken in its interpretation, especially in the light of compositional data analysis. Further downstream analysis, such as differential abundance of microbes, can also facilitate parsing of tables in the shiny application. A gold standard for differential abundance analysis in microbiome datasets is currently non-existent and different tools reach different results (Calgaro et al., 2020; Nearing et al., 2022). We offer edgeR in MetaFunc as we believe it is a good initial tool to explore DA microbes, though this is offset by being prone to false positives (Thorsen et al., 2016). MetaFunc results provide potential starting points for more in-depth analyses or hypothesis generation for experimental procedures. In this regard, we provide results in "*.tsv*" formats for use in other downstream bioinformatics applications, so users might apply their own analyses of choosing.

Correlation analysis on compositional data has the same contentious issue as differential abundance. Although there is published literature supporting the use of Spearman rank correlation coefficient in this analysis (Cremonesi et al., 2018; Dai et al., 2018; Geng et al., 2014), there are dissenting voices stating that there are spurious correlations, especially in compositional data (Aitchison, 1982; Faust et al., 2012; Friedman and Alm, 2012; Lovell et al., 2015; Pearson, 1897), and as such, conclusions from such correlations are meaningless (Lovell et al., 2015). Nevertheless, Spearman correlation serves a useful purpose, especially for an initial exploration of the data. Should users choose other analyses methods, intermediate results are provided with the pipeline.

This method was developed specifically for an RNA-seq (transcriptomic/metatranscriptomic) dataset, allowing for the common analysis applied to such studies. It is intended for an initial complete analysis of the data, with only a single configuration file and sample sheet necessary once installation of the tool has been done. Users can augment this analysis by accessing host gene and microbiome count files supplied by the pipeline and use this as input in other applications. Users can also potentially use the microbiome aspect of the pipeline on a metagenomic dataset, and can adjust this in the configuration file. As Kaiju (Menzel et al., 2016) identifies a single best protein match (or multiple matches with equal scores) of a read, we recommend its usage for short-read datasets. An exception could be made for long read sets in which the user is certain an input read will only span one protein.

We used the MetaFunc pipeline to compare genes and microbes between or among groups, but exploratory analyses of datasets from single groups can also be carried out.

While the methodology of this paper focuses on RNA sequences, metagenomic content could affect variation seen in microbial community gene expression (Franzosa et al., 2014). It should be noted that gene copy number, for instance, could affect transcript counts. Counts seen with

https://doi.org/10.1017/gmb.2022.12 Published online by Cambridge University Press

metatranscriptomic data would also reflect a species' gene expression contribution as opposed to abundance. It would be prudent to take this into consideration when interpreting biological implications of the results.

Conclusion

Here we presented MetaFunc, a single pipeline for analysing host and microbiome sequencing reads and their relationships. We found that we identified more microbes in our test datasets using MetaFunc compared to HUMAnN2, while microbes and functions of interest were comparable between the two. We have used MetaFunc to determine that microbes previously known to have associations with CRC are indeed relatively more abundant in CRC samples compared to normal samples. Furthermore, we were able to use MetaFunc to highlight that these microorganisms could contribute to CRC progression through polyamine production.

For a dataset with more than two groups, we have also used MetaFunc to identify abundant bacteria in a CRC subtype associated with immune responses, while conversely, we have not been able to identify significant microbes in the other CRC subtypes. MetaFunc's Spearman correlation step showed that the significant bacteria correlate with human DEGs that function in immune responses and CRC progression. We showed that MetaFunc was able to identify candidate microorganisms that differentiate sample groups and provide insight on the functional capacities of these candidates.

Acknowledgement. The authors would like to thank Dr. Olin Silander for valuable technical and academic advice for this manuscript.

Disclosure statement. The authors have no competing interests.

Supplementary materials. To view supplementary material for this article, please visit http://doi.org/10.1017/gmb.2022.12.

Data availability statement. MetaFunc is freely available through https://gitlab.com/schmeierlab/workflows/metafunc.git, and full documentation can be found in https://metafunc.readthedocs.io/en/latest/.

Author contributions. Conceptualisation: A.K.S, R.P., S.S.; Formal analysis: A.K.S, F.A.F., R.P., S.S.; Funding Acquisition: R.P.; Investigation: A.K.S., R.P., S.S.; Methodology: A.K.S, T.K., R.P., S.S.; Software: A.K.S, T.K., S.S.; Supervision: R.P., S.S.; Validation: A.K.S, S.S.; Visualisation: A.K.S; Writing – original draft: A.K.S; Writing – review and editing: A.K.S, T.K., F.A.F, R.P., S.S. A.K.S. and S.S. developed and co-wrote the pipeline which ultimately led to MetaFunc, and were involved with the majority of the design. T.K. developed the shiny application that is integrated in the pipeline. A.K.S. wrote the manuscript with editorial input from T.K., S.S. and R.P. R.P. further contributed to the design of the pipeline. F.A.F. provided guidance about all clinical aspects of the manuscript.

Funding. This work was supported in part by the Maurice and Phyllis Paykel Trust, Gut Cancer Foundation (NZ), with support from the Hugh Green Foundation, Colorectal Surgical Society of Australia and New Zealand (CSSANZ) and The Health Research Council of New Zealand.

Notes on Contributors. A.K.S recently finished her PhD in genetics from Massey University and is currently continuing her research on the microbiome in colorectal cancer at the University of Otago, Christchurch as a postdoctoral fellow. T.K. is a PhD student in computer science from Massey University. He is working on using computational modeling to study biomarkers in colorectal cancer. F.A.F is a professor of colorectal surgery at University of Otago, Christchurch and a colorectal surgeon at Christchurch Hospital with the Canterbury District Health Board. His research interests lie in the management and outcomes of patients with colorectal disease. RP is a senior research fellow at the University of Otago, Christchurch. Her research focuses on the microbiome of colorectal cancer. SS is a data scientist specializing in biological high-throughput data. He used to be an independent research group leader at Massey University. Currently, he is a senior scientist at Evotec, SE.

References

Aitchison J (1982) The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)* **44**(2), 139–160. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x

Argyle D and **Kitamura T** (2018) Targeting macrophage-recruiting chemokines as a novel therapeutic strategy to prevent the progression of solid tumors. *Frontiers in Immunology* **9**, 2629. https://doi.org/10.3389/fimmu.2018.02629

Aronesty E (2011) ea-utils: Command-line tools for processing biological sequencing data [C^{++}]. Available at https://github.com/ExpressionAnalysis/ea-utils (original work published 2015) (accessed 16 August 2018).

Aronesty E (2013) Comparison of sequencing utility programs. *The Open Bioinformatics Journal* **7**(1), 1–8. https://doi.org/10.2174/1875036201307010001

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM and **Sherlock G** (2000) Gene ontology: Tool for the unification of biology. *Nature Genetics* **25**(1), 25–29. https://doi.org/10.1038/75556

Aslani N, Janbabaei G, Abastabar M, Meis JF, Babaeian M, Khodavaisy S, Boekhout T and **Badali H** (2018) Identification of uncommon oral yeasts from cancer patients by MALDI-TOF mass spectrometry. *BMC Infectious Diseases* **18**(1), 24. https://doi.org/10.1186/s12879-017-2916-5

Bashiardes S, Zilberman-Schapira G and **Elinav E** (2016) Use of metatranscriptomics in microbiome research. *Bioinformatics and Biology Insights* **10**, 19–25. https://doi.org/10.4137/BBI.S34610

Bayoumi A, Sayed A, Broskova Z, Teoh J-P, Wilson J, Su H, Tang Y-L and **Kim I** (2016) Crosstalk between long noncoding RNAs and microRNAs in health and disease. *International Journal of Molecular Sciences* **17**(3), 356. https://doi.org/10.3390/ijms17030356

Calgaro M, Romualdi C, Waldron L, Risso D and **Vitulo N** (2020) Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. *Genome Biology* **21**(1), 191. https://doi.org/10.1186/s13059-020-02104-1

Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R and **Apweiler R** (2004). The gene ontology annotation (GOA) database: Sharing knowledge in Uniprot with gene ontology. *Nucleic Acids Research* **32** (database issue), D262–D266. https://doi.org/10.1093/nar/gkh021

Chen S, Zhou Y, Chen Y and **Gu J** (2018) Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Cox MP, Peterson DA and **Biggs PJ** (2010) SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* **11**(1), 485. https://doi.org/10.1186/1471-2105-11-485

Cremonesi E, Governa V, Garzon JFG, Mele V, Amicarella F, Muraro MG, Trella E, Galati-Fournier V, Oertli D, Däster SR, Droeser RA, Weixler B, Bolli M, Rosso R, Nitsche U, Khanna N, Egli A, Keck S, Slotta-Huspenina J, Terracciano LM, Zajac P, Spagnoli GC, Eppenberger-Castori S, Janssen KP, Borsig L and **Iezzi G** (2018) Gut microbiota modulate T cell trafficking into human colorectal cancer. *Gut* **67**, 1984–1994. https://doi.org/10.1136/gutjnl-2016-313498

Dai Z, Coker OO, Nakatsu G, Wu WKK, Zhao L, Chen Z, Chan FKL, Kristiansen K, Sung JJY, Wong SH and **Yu J** (2018) Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. *Microbiome* **6**(1), 70. https://doi.org/10.1186/s40168-018-0451-2

De la Fuente López M, Landskron G, Parada D, Dubois-Camacho K, Simian D, Martinez M, Romero D, Roa JC, Chahuán I, Gutiérrez R, Lopez-K F, Alvarez K, Kronberg U, López S, Sanguinetti A, Moreno N, Abedrapo M, González M-J, Quera R and **Hermoso-R MA** (2018) The relationship between chemokines CCL2, CCL3, and CCL4 with the tumor microenvironment and tumor-associated macrophage markers in colorectal cancer. *Tumor Biology* **40**(11), 1010428318810059. https://doi.org/10.1177/1010428318810059

Desnos-Ollivier M, Moquet O, Chouaki T, Guérin A-M and **Dromer F** (2011) Development of echinocandin resistance in *Clavispora lusitaniae* during saspofungin treatment. *Journal of Clinical Microbiology* **49**(6), 2304–2306. https://doi.org/10.1128/JCM.00325-11

Dienstmann R, Vermeulen L, Guinney J, Kopetz S, Tejpar S and **Tabernero J** (2017) Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nature Reviews Cancer* **17**(2), 79–92. https://doi.org/10.1038/nrc.2016.126

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M and **Gingeras TR** (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Douglas GM, Maffei VJ, Zaneveld J, Yurgel SN, Brown JR, Taylor CM, Huttenhower C and **Langille MGI** (2019). PICRUSt2: an improved and extensible approach for metagenome inference. *BioRxiv*, 672295. https://doi.org/10.1101/672295

Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J and **Huttenhower C** (2012) Microbial co-occurrence relationships in the human microbiome. *PLoS Computational Biology* **8**(7), e1002606. https://doi.org/10.1371/journal.pcbi.1002606

Fernandez NF, Gundersen GW, Rahman A, Grimes ML, Rikova K, Hornbeck P and **Ma'ayan A** (2017) Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data. *Scientific Data* **4**(1), 170151. https://doi.org/10.1038/sdata.2017.151

Flemer B, Warren RD, Barrett MP, Cisek K, Das A, Jeffery IB, Hurley E, O'Riordain M, Shanahan F and **O'Toole PW** (2018) The oral microbiota in colorectal cancer is distinctive and predictive. *Gut* **67**(8), 1454–1463. https://doi.org/10.1136/gutjnl-2017-314814

Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, Lipson KS, Knight R, Caporaso JG, Segata N and **Huttenhower C** (2018) Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods* **15**(11), 962. https://doi.org/10.1038/s41592-018-0176-y

https://doi.org/10.1017/gmb.2022.12 Published online by Cambridge University Press

Franzosa EA, **Morgan XC**, **Segata N**, **Waldron L**, **Reyes J**, **Earl AM**, **Giannoukos G**, **Boylan MR**, **Ciulla D**, **Gevers D**, **Izard J**, **Garrett WS**, **Chan AT** and **Huttenhower C** (2014) Relating the metatranscriptome and metagenome of the human gut. *Proceedings of the National Academy of Sciences* **111**(22), E2329–E2338. https://doi.org/10.1073/pnas.1319284111

Friedman J and **Alm EJ** (2012) Inferring correlation networks from genomic survey data. *PLoS Computational Biology* **8**(9), e1002687. https://doi.org/10.1371/journal.pcbi.1002687

Gao Z, **Guo B**, **Gao R**, **Zhu Q** and **Qin H** (2015) Microbiota disbiosis is associated with colorectal cancer. *Frontiers in Microbiology* **6**, 20. https://doi.org/10.3389/fmicb.2015.00020

Geng J, **Song Q**, **Tang X**, **Liang X**, **Fan H**, **Peng H**, **Guo Q** and **Zhang Z** (2014) Co-occurrence of driver and passenger bacteria in human colorectal cancer. *Gut Pathogens* **6**, 26. https://doi.org/10.1186/1757-4749-6-26

Gerner EW and **Meyskens FL** (2004) Polyamines and cancer: Old molecules, new understanding. *Nature Reviews Cancer* **4**(10), 781–792. https://doi.org/10.1038/nrc1454

Ghanipour A, **Jirström K**, **Pontén F**, **Glimelius B**, **Påhlman L** and **Birgisson H** (2009) The prognostic significance of tryptophanyl-tRNA synthetase in colorectal cancer. *Cancer Epidemiology and Prevention Biomarkers* **18**(11), 2949–2956. https://doi.org/10.1158/1055-9965.EPI-09-0456

Gloor GB, **Macklaim JM**, **Pawlowsky-Glahn V** and **Egozcue JJ** (2017) Microbiome datasets are compositional: And this is not optional. *Frontiers in Microbiology* **8**, 2224. https://doi.org/10.3389/fmicb.2017.02224

Gloor GB and **Reid G** (2016) Compositional analysis: A valid approach to analyze microbiome high-throughput sequencing data. *Canadian Journal of Microbiology* **62**(8), 692–703. https://doi.org/10.1139/cjm-2015-0821

Goffau MC, **Lager S**, **Salter SJ**, **Wagner J**, **Kronbichler A**, **Charnock-Jones DS**, **Peacock SJ**, **Smith GCS** and **Parkhill J** (2018) Recognizing the reagent microbiome. *Nature Microbiology* **3**(8), 851–853. https://doi.org/10.1038/s41564-018-0202-y

GSEA (n.d.) Available at https://www.gsea-msigdb.org/gsea/index.jsp (accessed 20 May 2020).

Guinney J, **Dienstmann R**, **Wang X**, **de Reyniès A**, **Schlicker A**, **Soneson C**, **Marisa L**, **Roepman P**, **Nyamundanda G**, **Angelino P**, **Bot BM**, **Morris JS**, **Simon IM**, **Gerster S**, **Fessler E**, **de Sousa e Melo F**, **Missiaglia E**, **Ramay H**, **Barras D**, **Homicsko K**, **Maru D**, **Manyam GC**, **Broom B**, **Boige V**, **Perez-Villamil B**, **Laderas T**, **Salazar R**, **Gray JW**, **Hanahan D**, **Tabernero J**, **Bernards R**, **Friend SH**, **Laurent-Puig P**, **Medema JP**, **Sadanandam A**, **Wessels L**, **Delorenzi M**, **Kopetz S**, **Vermeulen L** and **Tejpar S** (2015) The consensus molecular subtypes of colorectal cancer. *Nature Medicine* **21**(11), 1350–1356. https://doi.org/10.1038/nm.3967

Guo M, **Xu E** and **Ai D** (2019) Inferring bacterial infiltration in primary colorectal tumors from host whole genome sequencing data. *Frontiers in Genetics* **10**, 213. https://doi.org/10.3389/fgene.2019.00213

Huang H, **McGarvey PB**, **Suzek BE**, **Mazumder R**, **Zhang J**, **Chen Y** and **Wu CH** (2011) A comprehensive protein-centric ID mapping service for molecular data integration. *Bioinformatics* **27**(8), 1190–1191. https://doi.org/10.1093/bioinformatics/btr101

Hugerth LW and **Andersson AF** (2017) Analysing microbial community composition through amplicon sequencing: from sampling to hypothesis testing. *Frontiers in Microbiology* **8**, 1561. https://doi.org/10.3389/fmicb.2017.01561

Inamura K (2018) Colorectal cancers: an update on their molecular pathology. *Cancers* **10**(1), 26. https://doi.org/10.3390/cancers10010026

Jin M (2019) Unique roles of tryptophanyl-tRNA synthetase in immune control and its therapeutic implications. *Experimental & Molecular Medicine* **51**, 1–10. https://doi.org/10.1038/s12276-018-0196-9

Jung Y-S, **Jun S**, **Lee SH**, **Sharma A** and **Park J-I** (2015) Wnt2 complements Wnt/β-catenin signaling in colorectal cancer. *Oncotarget* **6**(35), 37257–37268.

Klopfenstein DV, **Zhang L**, **Pedersen BS**, **Ramírez F**, **Warwick Vesztrocy A**, **Naldi A**, **Mungall CJ**, **Yunes JM**, **Botvinnik O**, **Weigel M**, **Dampier W**, **Dessimoz C**, **Flick P** and **Tang H** (2018) GOATOOLS: a python library for gene ontology analyses. *Scientific Reports* **8**(1), 10872. https://doi.org/10.1038/s41598-018-28948-z

Koliarakis I, **Messaritakis I**, **Nikolouzakis TK**, **Hamilos G**, **Souglakos J** and **Tsiaoussis J** (2019) Oral bacteria and intestinal dysbiosis in colorectal cancer. *International Journal of Molecular Sciences* **20**(17), 4146. https://doi.org/10.3390/ijms20174146

Köster J and **Rahmann S** (2012) Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Krcmery V, **Mateicka F**, **Grausova S**, **Kunova A** and **Hanzen J** (1999) Invasive infections due to *Clavispora lusitaniae*. *FEMS Immunology & Medical Microbiology* **23**(1), 75–78. https://doi.org/10.1111/j.1574-695X.1999.tb01719.x

Kumar R, **Herold JL**, **Schady D**, **Davis J**, **Kopetz S**, **Martinez-Moczygemba M**, **Murray BE**, **Han F**, **Li Y**, **Callaway E**, **Chapkin RS**, **Dashwood W-M**, **Dashwood RH**, **Berry T**, **Mackenzie C** and **Xu Y** (2017) *Streptococcus gallolyticus* subsp. gallolyticus promotes colorectal tumor development. *PLOS Pathogens* **13**(7), e1006440. https://doi.org/10.1371/journal.ppat.1006440

Langille MGI (2018) Exploring linkages between taxonomic and functional profiles of the human microbiome. *MSystems* **3**(2), e00163–e00117. https://doi.org/10.1128/mSystems.00163-17

Langille MGI, **Zaneveld J**, **Caporaso JG**, **McDonald D**, **Knights D**, **Reyes JA**, **Clemente JC**, **Burkepile DE**, **Vega Thurber RL**, **Knight R**, **Beiko RG** and **Huttenhower C** (2013) Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology* **31**(9), 814–821. https://doi.org/10.1038/nbt.2676

https://doi.org/10.1017/gmb.2022.12 Published online by Cambridge University Press

Leon-Cabrera S, **Vázquez-Sandoval A**, **Molina-Guzman E**, **Delgado-Ramirez Y**, **Delgado-Buenrostro NL**, **Callejas BE**, **Chirino YI**, **Pérez-Plasencia C**, **Rodríguez-Sosa M**, **Olguín JE**, **Salinas C**, **Satoskar AR** and **Terrazas LI** (2018) Deficiency in STAT1 signaling predisposes gut inflammation and prompts colorectal cancer development. *Cancers* **10**(9), 341. https://doi.org/10.3390/cancers10090341

Li M, **Zhao L**, **Li S**, **Li J**, **Gao B**, **Wang F**, **Wang S**, **Hu X**, **Cao J** and **Wang G** (2018) Differentially expressed lncRNAs and mRNAs identified by NGS analysis in colorectal cancer patients. *Cancer Medicine* **7**(9), 4650–4664. https://doi.org/10.1002/cam4.1696

Liao Y, **Smyth GK** and **Shi W** (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Liberzon A, **Subramanian A**, **Pinchback R**, **Thorvaldsdóttir H**, **Tamayo P** and **Mesirov JP** (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**(12), 1739–1740. https://doi.org/10.1093/bioinformatics/btr260

Lovell D, **Pawlowsky-Glahn V**, **Egozcue JJ**, **Marguerat S** and **Bähler J** (2015) Proportionality: A valid alternative to correlation for relative data. *PLoS Computational Biology* **11**(3), e1004075. https://doi.org/10.1371/journal.pcbi.1004075

Macklaim JM and **Gloor GB** (2018) From RNA-seq to biological inference: Using compositional data analysis in meta-transcriptomics. In Beiko RG, Hsiao W and Parkinson J (eds), *Microbiome Analysis: Methods and Protocols*. New York, NY: Springer, pp. 193–213. https://doi.org/10.1007/978-1-4939-8728-3_13

Malinowski B, **Węsierska A**, **Zalewska K**, **Sokołowska MM**, **Bursiewicz W**, **Socha M**, **Ozorowski M**, **Pawlak-Osińska K** and **Wiciński M** (2019) The role of *Tannerella forsythia* and *Porphyromonas gingivalis* in pathogenesis of esophageal cancer. *Infectious Agents and Cancer* **14**(1), 3. https://doi.org/10.1186/s13027-019-0220-2

McCarthy DJ, **Chen Y** and **Smyth GK** (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research* **40**(10), 4288–4297. https://doi.org/10.1093/nar/gks042

Menzel P, **Ng KL** and **Krogh A** (2016) Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications* **7**, 11257. https://doi.org/10.1038/ncomms11257

Nayfach S, **Bradley PH**, **Wyman SK**, **Laurent TJ**, **Williams A**, **Eisen JA**, **Pollard KS** and **Sharpton TJ** (2015) Automated and accurate estimation of gene family abundance from shotgun metagenomes. *PLoS Computational Biology* **11**(11), e1004573. https://doi.org/10.1371/journal.pcbi.1004573

Nearing JT, **Douglas GM**, **Hayes MG**, **MacDonald J**, **Desai DK**, **Allward N**, **Jones CMA**, **Wright RJ**, **Dhanani AS**, **Comeau AM** and **Langille MGI** (2022) Microbiome differential abundance methods produce different results across 38 datasets. *Nature Communications* **13**(1), 342. https://doi.org/10.1038/s41467-022-28034-z

Ohkubo K, **Sakai Y**, **Inoue H**, **Akamine S**, **Ishizaki Y**, **Matsushita Y**, **Sanefuji M**, **Torisu H**, **Ihara K**, **Sardiello M** and **Hara T** (2015) Moyamoya disease susceptibility gene RNF213 links inflammatory and angiogenic signals in endothelial cells. *Scientific Reports* **5**(1), 13191. https://doi.org/10.1038/srep13191

Parks T, **Barrett L** and **Jones N** (2015) Invasive streptococcal disease: A review for clinicians. *British Medical Bulletin* **115**(1), 77–89. https://doi.org/10.1093/bmb/ldv027

Patro R, **Duggal G**, **Love MI**, **Irizarry RA** and **Kingsford C** (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14**(4), 417–419. https://doi.org/10.1038/nmeth.4197

Pearson K (1897) Mathematical contributions to the theory of evolution—On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London* **60**(359–367), 489–498. https://doi.org/10.1098/rspl.1896.0076

Pinto ML, **Rios E**, **Durães C**, **Ribeiro R**, **Machado JC**, **Mantovani A**, **Barbosa MA**, **Carneiro F** and **Oliveira MJ** (2019) The two faces of tumor-associated macrophages and their clinical significance in colorectal cancer. *Frontiers in Immunology* **10**, 1875. https://doi.org/10.3389/fimmu.2019.01875

Purcell RV, **Visnovska M**, **Biggs PJ**, **Schmeier S** and **Frizelle FA** (2017) Distinct gut microbiome patterns associate with consensus molecular subtypes of colorectal cancer. *Scientific Reports* **7**(1), 11590. https://doi.org/10.1038/s41598-017-11237-6

Reinecke F and **Steinbüchel A** (2009) Ralstonia eutropha strain H16 as model organism for PHA metabolism and for biotechnological production of technically interesting biopolymers. *Journal of Molecular Microbiology and Biotechnology* **16** (1–2), 91–108. https://doi.org/10.1159/000142897

Robinson MD, **McCarthy DJ** and **Smyth GK** (2010) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Salter SJ, **Cox MJ**, **Turek EM**, **Calus ST**, **Cookson WO**, **Moffatt MF**, **Turner P**, **Parkhill J**, **Loman NJ** and **Walker AW** (2014) Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology* **12**(1), 87. https://doi.org/10.1186/s12915-014-0087-z

Sánchez-Rovira P, **Jimenez E**, **Carracedo J**, **Barneto IC**, **Ramirez R** and **Aranda E** (1998) Serum levels of intercellular adhesion molecule 1 (ICAM-1) in patients with colorectal cancer: Inhibitory effect on cytotoxicity. *European Journal of Cancer* **34**(3), 394–398. https://doi.org/10.1016/S0959-8049(97)10033-8

Schellerer VS, **Langheinrich MC**, **Zver V**, **Grützmann R**, **Stürzl M**, **Gefeller O**, **Naschberger E** and **Merkel S** (2019) Soluble intercellular adhesion molecule-1 is a prognostic marker in colorectal carcinoma. *International Journal of Colorectal Disease* **34**(2), 309–317. https://doi.org/10.1007/s00384-018-3198-0

https://doi.org/10.1017/gmb.2022.12 Published online by Cambridge University Press

Segata N, **Waldron L**, **Ballarini A**, **Narasimhan V**, **Jousson O** and **Huttenhower C** (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods* **9**(8), 811–814. https://doi.org/10.1038/nmeth.2066

Sharma AK, **Gupta A**, **Kumar S**, **Dhakan DB** and **Sharma VK** (2015) Woods: A fast and accurate functional annotator and classifier of genomic and metagenomic sequences. *Genomics* **106**(1), 1–6. https://doi.org/10.1016/j.ygeno.2015.04.001

Shen W and **Ren H** (2021) TaxonKit: A practical and efficient NCBI taxonomy toolkit. *Journal of Genetics and Genomics* **48**(9), 844–850. https://doi.org/10.1016/j.jgg.2021.03.006

Silva GGZ, **Green KT**, **Dutilh BE** and **Edwards RA** (2016) SUPER-FOCUS: A tool for agile functional analysis of shotgun metagenomic data. *Bioinformatics (Oxford, England)* **32**(3), 354–361. https://doi.org/10.1093/bioinformatics/btv584

Soda K (2011) The mechanisms by which polyamines accelerate tumor spread. *Journal of Experimental & Clinical Cancer Research* **30**(1), 95. https://doi.org/10.1186/1756-9966-30-95

Song F, **Yi Y**, **Li C**, **Hu Y**, **Wang J**, **Smith DE** and **Jiang H** (2018) Regulation and biological role of the peptide/histidine transporter SLC15A3 in toll-like receptor-mediated inflammatory responses in macrophage. *Cell Death & Disease* **9**(7), 1–15. https://doi.org/10.1038/s41419-018-0809-1

Subramanian A, **Tamayo P**, **Mootha VK**, **Mukherjee S**, **Ebert BL**, **Gillette MA**, **Paulovich A**, **Pomeroy SL**, **Golub TR**, **Lander ES** and **Mesirov JP** (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102

Sulit AK, **Kolisnik T**, **Frizelle FA**, **Purcell R** and **Schmeier S** (2021a) *MetaFunc Databases: Kaiju Database* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5602178

Sulit AK, **Kolisnik T**, **Frizelle FA**, **Purcell R** and **Schmeier S** (2021b) *MetaFunc Databases: Nr-go Database* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5602157

Tachimori A, **Yamada N**, **Sakate Y**, **Yashiro M**, **Maeda K**, **Ohira M**, **Nishino H** and **Hirakawa K** (2005) Up regulation of ICAM-1 gene expression inhibits tumour growth and liver metastasis in colorectal carcinoma. *European Journal of Cancer* **41**(12), 1802–1810. https://doi.org/10.1016/j.ejca.2005.04.036

Tanaka A, **Zhou Y**, **Ogawa M**, **Shia J**, **Klimstra DS**, **Wang JY** and **Roehrl MH** (2020) STAT1 as a potential prognosis marker for poor outcomes of early stage colorectal cancer with microsatellite instability. *PLoS One* **15**(4), e0229252. https://doi.org/10.1371/journal.pone.0229252

Ternes D, **Karta J**, **Tsenkova M**, **Wilmes P**, **Haan S** and **Letellier E** (2020) Microbiome in colorectal cancer: how to get from meta-omics to mechanism? *Trends in Microbiology* **28**, 401–423. https://doi.org/10.1016/j.tim.2020.01.001

The UniProt Consortium (2017) UniProt: The universal protein knowledgebase. *Nucleic Acids Research* **45**(D1), D158–D169. https://doi.org/10.1093/nar/gkw1099

Thomas AM, **Manghi P**, **Asnicar F**, **Pasolli E**, **Armanini F**, **Zolfo M**, **Beghini F**, **Manara S**, **Karcher N**, **Pozzi C**, **Gandini S**, **Serrano D**, **Tarallo S**, **Francavilla A**, **Gallo G**, **Trompetto M**, **Ferrero G**, **Mizutani S**, **Shiroma H**, **Shiba S**, **Shibata T**, **Yachida S**, **Yamada T**, **Wirbel J**, **Schrotz-King P**, **Ulrich CM**, **Brenner H**, **Arumugam M**, **Bork P**, **Zeller G**, **Cordero F**, **Dias-Neto E**, **Setubal JC**, **Tett A**, **Pardini B**, **Rescigno M**, **Waldron L**, **Naccarati A** and **Segata N** (2019) Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nature Medicine* **25**(4), 667–678. https://doi.org/10.1038/s41591-019-0405-7

Thorsen J, **Brejnrod A**, **Mortensen M**, **Rasmussen MA**, **Stokholm J**, **Al-Soud WA**, **Sørensen S**, **Bisgaard H** and **Waage J** (2016) Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. *Microbiome* **4**(1), 62. https://doi.org/10.1186/s40168-016-0208-8

Tofalo R, **Cocchi S** and **Suzzi G** (2019) Polyamines and gut microbiota. *Frontiers in Nutrition* **6**, 16. https://doi.org/10.3389/fnut.2019.00016

Truong DT, **Franzosa EA**, **Tickle TL**, **Scholz M**, **Weingart G**, **Pasolli E**, **Tett A**, **Huttenhower C** and **Segata N** (2015) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods* **12**(10), 902–903. https://doi.org/10.1038/nmeth.3589

Vyas S and **Chang P** (2014) New PARP targets for cancer therapy. *Nature Reviews. Cancer* **14**(7), 502–509. https://doi.org/10.1038/nrc3748

Wang Y, **Sun D**, **Song F**, **Hu Y**, **Smith DE** and **Jiang H** (2014) Expression and regulation of the proton-coupled oligopeptide transporter PhT2 by LPS in macrophages and mouse spleen. *Molecular Pharmaceutics* **11**(6), 1880–1888. https://doi.org/10.1021/mp500014r

Wei H, **Chen L**, **Lian G**, **Yang J**, **Li F**, **Zou Y**, **Lu F** and **Yin Y** (2018) Antitumor mechanisms of bifidobacteria. *Oncology Letters* **16**(1), 3–8. https://doi.org/10.3892/ol.2018.8692

Whitmore SE and **Lamont RJ** (2014) Oral bacteria and cancer. *PLoS Pathogens* **10**(3), e1003933. https://doi.org/10.1371/journal.ppat.1003933

Wirbel J, **Pyl PT**, **Kartal E**, **Zych K**, **Kashani A**, **Milanese A**, **Fleck JS**, **Voigt AY**, **Palleja A**, **Ponnudurai R**, **Sunagawa S**, **Coelho LP**, **Schrotz-King P**, **Vogtmann E**, **Habermann N**, **Niméus E**, **Thomas AM**, **Manghi P**, **Gandini S**, **Serrano D**, **Mizutani S**, **Shiroma H**, **Shiba S**, **Shibata T**, **Yachida S**, **Yamada T**, **Waldron L**, **Naccarati A**, **Segata N**, **Sinha R**, **Ulrich CM**, **Brenner H**, **Arumugam M**, **Bork P** and **Zeller G** (2019) Meta-analysis of fecal metagenomes reveals global microbial

https://doi.org/10.1017/gmb.2022.12 Published online by Cambridge University Press

signatures that are specific for colorectal cancer. *Nature Medicine* **25**(4), 679–689. https://doi.org/10.1038/s41591-019-0406-6

Xia Y, **Sun J** and **Chen D-G** (2018) Bioinformatic analysis of microbiome data. In Xia Y, Sun J and Chen D-G (eds), *Statistical Analysis of Microbiome Data with R*. Singapore: Springer, pp. 1–27. https://doi.org/10.1007/978-981-13-1534-3_1.

Xia X, **Wu WKK**, **Wong SH**, **Liu D**, **Kwong TNY**, **Nakatsu G**, **Yan PS**, **Chuang Y-M**, **Chan MW-Y**, **Coker OO**, **Chen Z**, **Yeoh YK**, **Zhao L**, **Wang X**, **Cheng WY**, **Chan MTV**, **Chan PKS**, **Sung JJY**, **Wang MH** and **Yu J** (2020) Bacteria pathogens drive host colonic epithelial cell promoter hypermethylation of tumor suppressor genes in colorectal cancer. *Microbiome* **8**(1), 108. https://doi.org/10.1186/s40168-020-00847-4

Ye SH, **Siddle KJ**, **Park DJ** and **Sabeti PC** (2019) Benchmarking metagenomics tools for taxonomic classification. *Cell* **178**(4), 779–794. https://doi.org/10.1016/j.cell.2019.07.010

Ye X, **Wang R**, **Bhattacharya R**, **Boulbes DR**, **Fan F**, **Xia L**, **Adoni H**, **Ajami NJ**, **Wong MC**, **Smith DP**, **Petrosino JF**, **Venable S**, **Qiao W**, **Baladandayuthapani V**, **Maru D** and **Ellis LM** (2017) *Fusobacterium Nucleatum* subspecies *Animalis* influences Proinflammatory cytokine expression and monocyte activation in human colorectal tumors. *Cancer Prevention Research* **10**(7), 398–409. https://doi.org/10.1158/1940-6207.CAPR-16-0178

Yu G, **Wang L-G**, **Han Y** and **He Q-Y** (2012) clusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology* **16**(5), 284–287. https://doi.org/10.1089/omi.2011.0118

Cite this article: Sulit A.K., Kolisnik T., Frizelle F.A., Purcell R., and Schmeier S. 2023. MetaFunc: taxonomic and functional analyses of high throughput sequencing for microbiomes. *Gut Microbiome,* **4**, e4, 1–21. https://doi.org/10.1017/gmb.2022.12