Identification of prognostic indicators of healthy and unhealthy conditions with a machine learning-based systems biology approach using gut microbiome data

by

Matthew Jack Madgwick

A thesis submitted for the degree of Doctor of Philosophy

Earlham Institute
Quadram Institute
University of East Anglia
BenevolentAI

Supervisors:
Dr Tamas Korcsmaros
Dr Samer Abujudeh

United Kingdom

November 2022

Abstract

Inflammatory bowel disease (IBD) is associated with alterations in the intestinal microbiome. However, the precise nature of these microbial changes remains unclear. With billions of microbes within the gut, novel and powerful computational techniques are required to identify the relevant shifts in the microbiota that contribute to healthy and unhealthy conditions.

Machine learning (ML) allows a data-driven approach to identify these discrete dynamic changes. However, the interpretation and biological validation of the findings from ML algorithms remain a challenge. By combining ML and Systems Biology (SB) approaches, this thesis aims to characterise key microbial factors in IBD pathogenesis by extracting prognostic indicators from the human gut microbiome.

The causal relationship between the changes in the gut microbiome and IBD is difficult to establish. Data from cross-sectional studies are plagued by confounding factors and inconsistencies between cohorts. Rich longitudinal datasets and integrated metagenomic, multi-omic, and electronic healthcare records can be used to overcome these limitations. In this PhD thesis, I have developed an integrated ML-based microbiome analysis pipeline to identify prognostic indicators for IBD from longitudinal microbiome data. Furthermore, using a variety of SB approaches, the interplay between the host and the microbiome has been explored to provide insights into the mechanisms during healthy and unhealthy conditions.

Acknowledgements

For a PhD to be successful, one must have the support of their research group, family and friends. I want to express my gratitude and thanks to those who made this thesis possible.

First and foremost. Tamas Korcsmaros - thank you. You saw something in me that many others dismissed, enabling me to actualise my ideas and dreams. Without your guidance, friendly advice, support, trust and belief in me, I would not be the scientist I am today. I could not have had a better supervisor, mentor or role model than you. You have set the bar very high for whoever comes after you. This gratitude extends to the entirety of the Korcsmaros group, past and present. Agatha, Marton, Lejla, David, Mate, Dezso, Martina, Isabelle, John and Polina, for helping me learn and expand my biological, network biology and computer science knowledge. To Lucie and Elena, thank you for your friendship and belief in me throughout my PhD. In particular, I would like to highlight Lejla Gul for her guidance, mentorship and friendship during my PhD. I would also like to express my sincere gratitude to Junaid Butt for his invaluable mathematical assistance. Finally, I would like to thank my industrial partner, BenevolentAI, specifically Sam Abujudeh, for being an exceptional mentor by devoting additional time and effort to supervising my PhD project.

I am eternally grateful to my parents; you are both my inspiration. From a young age, I have always wanted to be a scientist like you both, and you nurtured that passion. Your unwavering support, love and encouragement have contributed more to this thesis than you know or I can describe in a paragraph. These words do not adequately express how deeply I appreciate everything you have done for me. Thank you to my brother, James, who was always there to bring me back up, celebrate the good times and make me laugh during the worst of times (that said, you must only address me as Dr. now).

Finally - Sam Witham, Peter Osborne and Anita Scoones. This journey would not have been the same without you. For all the late nights, the motivation to keep going and the memories we made together will always be one of the most defining parts of my PhD. You are all outstanding, talented, and incredible individuals I am proud and privileged to call my friends. You will all go far in the future, and I cannot wait to see what you all achieve.

Table of contents

List of abbreviations

16S: 16S ribosomal RNA

AD: Alzheimer's Disease

AI: Artificial Intelligence

ANCOM: Analysis of compositions of microbiome

ANOVA: Analysis of variance

AUC: Area under receiver operating curve

BAI: Bile-acid-induced

CCF: Cross correlation function

CD: Crohn's disease

CLR: Centred log-ratio

CoDa: Compositional data analysis

CTF: Compositional Tensor Factorization

DA: Dalton/unified atomic mass unit

DBN: dynamic Bayesian network

dbRDA: Distance-based redundancy analysis

DDI: Domain-domain interaction

DEG: Differentially expressed genes

DL: Deep Learning

DMI: Domain-motif interaction

EDA: Exploratory data analysis

ELM: Eukaryotic Linear Motif

EMBL: European Molecular Biology Laboratory

FA: Factor Analysis

FCBT: Log Fold Change baseline transformation

FDR: False discovery rate

FMT: faecal microbiota transplantation

GI: Gastrointestinal

GPCR: G protein-coupled receptor

HBI: Harvey-Bradshaw Index

HC: Healthy control

HMI: Host-microbe interactions

HMP: Human Microbiome Project

HPC: High Performance Computing

i.i.d: Independent and identically distributed random variables

IBD: Inflammatory bowel disease

IC: Independent Component

ICA: Independent Component Analysis

IL10: Interleukin 10

ILR: Isometric log-ratio

IVA: Independent Vector Analysis

LC-MS: Liquid chromatography-mass spectrometry

LDA: Linear Discriminant Analysis

LOGOCV: Leave-one group out cross validation

LOOCV: Leave-one out cross validation

LR: Linear regression

MAPK: Mitogen-activated protein kinases

MCMC: Markov chain Monte Carlo

ML: Machine Learning

mLDM: Environmental factor-microbe association

MMinte: Microbial metabolic interactions

MMPC: Max-Min parents and Child

MSE: Mean squared error

NGS: Next generation sequencing

NGS: Next-generation sequencing

NMF: Non-negative Matrix Factorisation

OPLS-DA: Orthogonal partial least squares discriminant analysis

OTU: Operational taxonomic unit

P-gp: P-Glycoprotein

PBA: Primary bile acid

PC: Principal Component

PCA: Principal Component Analysis

PCoA: Principal Coordinates Analysis

PCR: Polymerase chain reaction

PERMANOVA: Permutational analysis of variance

PFAM: Proteins Families annotation

PI3K: Phosphatidylinositol-3 kinase

PLS-DA: Partial least squares-discriminant analysis

PPI: Protein-proteins interaction

PQN: Probabilistic quotient normalisation

PSM: Phenol-soluble modulins

QC: Quality control

RDA: Redundancy analysis

RF: Random Forest

ROC: Receiver Operator Curve

RprY: DNA-binding response regulator

RTF: Rotation Forest

SB: Systems biology

SBA: Secondary bile acid

SBT: Subtracted baseline transformation

SCCAI: Simple Clinical Colitis Activity Index

SCFA: Short-chain fatty acid

SLiM: Short linear motif

SPM: Species Precision Model

SVD: Singular value decomposition

TCA: Tensor Component Analysis

TDR: True discovery rate

TF: Transcription Factor

TNF: Tumour necrosis factor

TSS: Total Sum Scaling

UC: Ulcerative colitis

WGS: Whole genome sequencing

XGBoost: eXtreme Gradient Boosting

List of figures

Figure 5.6. Inferred multi-layer host-microbe protein-protein interaction (PPI) network from the source *Bacteroides vulgatus* microbial proteins (red triangles) to the host differentially expressed genes through downstream signalling proteins in ulcerative colitis.

Figure 5.7. Enriched functions among the DEGs in TieDie compared to the top 150 upregulated genes in UC.

Figure 5.8. Subset network modelling the host-microbe interactions and regulatory interactions between bacterial proteins and GPCR/MAPK pathway in ulcerative colitis.

List of tables

List of supplementary materials

Supplementary Figure 4.5. Metabolomic CD vs HC Scree plot for assessing the number of components needed for matrix factorisation algorithms based on Horn's parallel analysis.

Supplementary Figure 4.6. Metabolomic UC vs HC Scree plot for assessing the number of components needed for matrix factorisation algorithms based on Horn's parallel analysis.

Supplementary Figure 4.7. Metagenomic UC vs CD Scree plot for assessing the number of components needed for matrix factorisation algorithms based on Horn's parallel analysis.

Supplementary Table 4.1. Table of results for Metagenomic classifiers using different normalisation methods.

Supplementary Figure 4.8. Critical difference between models built for predicting phenotype based on metabolomic profiles.

Supplementary Table 4.2. Top ICA loadings for UC vs nonIBD with metabolites as sources.

Supplementary Figure 4.9. Inverse Kurtosis from ICA with metabolites as sources after FCBT between UC and nonIBD patients.

Supplementary Figure 4.10. Inverse Kurtosis from ICA with metabolites as sources after FCBT between CD and nonIBD patients.

Supplementary Figure 4.11. Hierarchical clustering of the ICs extracted between UC and healthy controls when using Metabolites as the sources.

Supplementary Table 5.1. 65 bacterial proteins that had domain-domain interactions.

Supplementary Figure 5.1. Inferred multi-layer host-microbe protein-protein interaction (PPI) network from the source *Bacteroides vulgatus* microbial proteins to the host target proteins in ulcerative colitis.

List of peer-review publications

Peer-reviewed journal articles published during my PhD from October 2018 - November 2022.

Publications represented in the thesis:

Chapter 1:

- Tabib, N. S. S., **Madgwick, M.**, Sudhakar, P., Verstockt, B., Korcsmaros, T., & Vermeire, S. (2020). Big data in IBD: big progress for clinical practice. Gut, 69(8), 1520–1532.

Publications not represented in the thesis:

1. Bohar, B., Fazekas, D., **Madgwick, M**., Csabai, L., Olbei, M., Korcsmáros, T., & Szalay-Beko, M. (2022). Sherlock: an open-source data platform to store, analyze and integrate Big Data for computational biologists. F1000Research, 10, 409.

2. Brooks-Warburton, J., Modos, D., Sudhakar, P., **Madgwick, M.**, Thomas, J. P., Bohar, B., Fazekas, D., Zoufir, A., Kapuy, O., Szalay-Beko, M., & others. (2022). A systems genomics approach to uncover patient-specific pathogenic pathways and proteins in ulcerative colitis. Nature Communications, 13(1), 1–12.

3. Csabai, L., Fazekas, D., Kadlecsik, T., Szalay-Bekő, M., Bohár, B., **Madgwick, M.**, Módos, D., Ölbei, M., Gul, L., Sudhakar, P., & others. (2022). SignaLink3: a multi-layered resource to uncover tissue-specific signaling networks. Nucleic Acids Research, 50(D1), D701–D709.

4. Gul, L., Modos, D., Fonseca, S., **Madgwick, M.**, Thomas, J. P., Sudhakar, P., Booth, C., Stentz, R., Carding, S. R., & Korcsmaros, T. (2022). Extracellular vesicles produced by the human commensal gut bacterium *Bacteroides thetaiotaomicron* affect host immune pathways in a cell-type specific manner that are altered in inflammatory bowel disease. Journal of Extracellular Vesicles, 11(1), e12189.

5. McKee, A. M., Kirkup, B. M., **Madgwick, M.**, Fowler, W. J., Price, C. A., Dreger, S. A., Ansorge, R., Makin, K. A., Caim, S., Le Gall, G., & others. (2021). Antibiotic-induced

disturbances of the gut microbiota result in accelerated breast tumor growth. Iscience, 24(9), 103012.

6. Olbei, M., Bohar, B., Fazekas, D., **Madgwick, M.**, Sudhakar, P., Hautefort, I., Métris, A., Baranyi, J., Kingsley, R. A., & Korcsmaros, T. (2022). Multilayered Networks of SalmoNet2 Enable Strain Comparisons of the *Salmonella* Genus on a Molecular Level. Msystems, 7(4), e01493-21.

7. Olbei, M., Thomas, J. P., Hautefort, I., Treveil, A., Bohar, B., **Madgwick, M.**, Gul, L., Csabai, L., Modos, D., & Korcsmaros, T. (2021). CytokineLink: A Cytokine Communication Map to Analyse Immune Responses—Case Studies in Inflammatory Bowel Disease and COVID-19. Cells, 10(9), 2242.

8. Pavlidis, P., Tsakmaki, A., Pantazi, E., Li, K., Cozzetto, D., Digby-Bell, J., Yang, F., Lo, J. W., Alberts, E., Sa, A. C. C., **Madgwick, M.**, & others. (2022). Interleukin-22 regulates neutrophil recruitment in ulcerative colitis and is associated with resistance to ustekinumab therapy. Nature Communications, 13(1), 1–17.

9. Poletti, M., Treveil, A., Csabai, L., Gul, L., Modos, D., **Madgwick, M.**, Olbei, M., Bohar, B., Valdeolivas, A., Turei, D., & others. (2022). Mapping the epithelial–immune cell interactome upon infection in the gut and the upper airways. NPJ Systems Biology and Applications, 8(1), 1–19.

10. Treveil, A., Bohar, B., Sudhakar, P., Gul, L., Csabai, L., Olbei, M., Poletti, M., **Madgwick, M.**, Andrighetti, T., Hautefort, I., & others. (2021). ViralLink: An integrated workflow to investigate the effect of SARS-CoV-2 on intracellular signalling and regulatory pathways. PLoS Computational Biology, 17(2), e1008685.

11. Lo, J., Cozzetto, D., Liu, Z., Ibraheim, H., Sieh, J., Olbei, M., Alexander, J., Blanco, J. M., **Madgwick, M.**, Kudo, H., & others. (2022). Immune checkpoint inhibitor-induced colitis is mediated by CXCR6+ polyfunctional lymphocytes and is dependent on the IL23/IFNγ axis. (*In Press*)

List of software developed

A list of both internal and open-source software packages, tools and pipelines which I designed, implemented, optimised, or contributed during the course of my PhD from October 2018 - November 2022.

Software packages/tools/pipelines represented in the thesis

- **LongitOmix:** A wrangler for the pipeline for the identification of prognostic indicators of between conditions with a machine learning-based systems biology approach from longitudinal gut microbiome data using independent component analysis (*Chapter 3 and 4*)
- **MetabolomiX:** a lightweight code base for handling, parsing, performing ID conversion and simple preprocessing steps of metabolomics data (*Chapter 4*)
- **Microbiolink2:** An Integrated Computational Pipeline to Infer Functional Effects of Microbiome–Host Interactions (*Chapter 5*)

Software packages/tools/pipelines/web resources developed that are not represented in the thesis

- **Integrated Single Nucleotide Polymorphism (iSNP) pipeline:** A novel precision medicine workflow designed to determine the mechanisms by which SNPs affect cellular regulatory networks, and how SNP co-occurrences contribute to disease pathogenesis in ulcerative colitis (UC). https://github.com/korcsmarosgroup/iSNP
- **ScOmix:** An internal single-cell and low-input preprocessing and downstream analysis code base developed to make single-cell analysis more efficient and interpretable for bioinformaticians (Internal Tool).

- **TranscriptOmix:** Bulk-RNA preprocessing, downstream and functional analysis pipeline for large-scale and efficient batch processing of bulk RNA-seq datasets. Designed to improve collaboration and preprocessing of publicly available datasets (Internal Tool).

- **CHAT:** Reimplemented a faster, more efficient and updated version of Conext Hub Analysis Tool (CHAT) for use on internal projects where a graphical user interface would not be usable. This is a python port of the Contextual Hub Analysis Tool for the application onto multiple patient-specific networks (Internal Tool).

- **PyDyNet:** A python port of DyNet, a tool for the analysis of protein-protein interaction networks to identify rewiring in response to different stimuli and in disease. https://github.com/korcsmarosgroup/pyDyNet

- **BioHandler:** Fast and efficient parsing and serialisation of biological data to different formats. This tool was used as the backend for web resources like Signalink, Salmonet and Autophagy Regulation Network. An example of this can be seen: http://signalink.org/download

- **ViralLink:** A systems biology workflow which reconstructs and analyses networks representing the effect of viral infection on specific human cell types. https://github.com/korcsmarosgroup/ViralLink

- **CytokineLink:** A map of cytokine communication for inflammatory and infectious diseases. https://github.com/korcsmarosgroup/CytokineLink

- **Signalink3**: An integrated resource to analyse signalling pathway cross-talks, transcription factors, miRNAs and regulatory enzymes. http://signalink.org/

- **SalmoNet**: an integrated network resource containing regulatory, metabolic and protein-protein interactions of *Salmonella*. http://salmonet.org/

- **AutophagyNet**: Autophagy Regulatory Network 2 (ARN2) is the updated version of the previous autophagy-focused network resource. The aim of the tool is to aid omics analysis and experiment planning. https://www.autophagynet.org/

- **Sherlock:** an open source data platform, developed in the Korcsmaros Group to store, analyse and integrate bioinformatics data. https://earlham-sherlock.github.io/

List of conference papers, posters and presentations

List of conference papers, presentations and posters I either presented or contributed to during the course of my PhD. Conferences where the posters and presentations are not published in a journal are not included in this list.

1. **Madgwick, M**; Sudhakar, P; Tabib, NS; Norvaisas, P; Creed, P; Verstockt, B; Vermeire, S; Korcsmáros, T; P070 Machine learning approaches to identify IBD biomarkers from longitudinal microbiome data, Journal of Crohn's and Colitis,14,Supplement_1, S170-S171,2020, Oxford University Press US.

2. Verstockt, B; Sudahakar, P; Creyns, B; Verstockt, S; Cremer, J; Wollants, WJ; Organe, S; Korcsmaros, T; **Madgwick, M**; Van Assche, G; , DOP70 An integrated multi-omics biomarker predicting endoscopic response in ustekinumab treated patients with Crohn's disease, Journal of Crohn's and Colitis,13, Supplement_1,S072-S073,2019, Oxford University Press US.

3. Verstockt, Bram; Sudhakar, Padhmanand; Creyns, Brecht; Verstockt, Sare; Cremer, Jonathan; Wollants, Willem-Jan; Organe, Sophie; Korcsmaros, Tamas; **Madgwick, Matthew**; Van Assche, Gert A; Predicting endoscopic response in ustekinumab treated patients with Crohn's disease through an integrated multi-omics biomarker,, Gastroenterology, 156,6,S656-S656,2019, Elsevier.

4. Lo, Jonathan; Cozzetto, Domenico; Sieh, Jillian Yong Xin; **Madgwick, Matthew**; Kudo, Hiromi; Alexander, James; Miguens-Blanco, Jesus; Korcsmaros, Tamas; Goldin, Robert; Marchesi, Julian; , PMO-4 Immune checkpoint inhibitor-induced colitis is mediated by IL23 responsive CD90+ cytotoxic lymphocytes, 2021, BMJ Publishing Group

5. Kornilova, P; Potari-Gul, L; Modos, D; **Madgwick, M**; Haerty, W; Korcsmaros, T; , P004 Critical paralog proteins has a cell-type specific rewiring role in Ulcerative Colitis associated signalling processes, Journal of Crohn's and Colitis,15,Supplement_1,S126-S127, 2021, Oxford University Press US

6. Potari-gul, L; Modos, D; Turei, D; Valdeolivas, A; **Madgwick, M**; Saez-Rodriguez, J; Korcsmaros, T; ,P020 Mapping the changing intercellular communication and its downstream effect in Ulcerative Colitis, Journal of Crohn's and Colitis,15, Supplement_1,S138-S139,2021, Oxford University Press US

7. Modos, D; Brooks-Warburton, J; Sudhakar, P; **Madgwick, M**; Fazekas, D; Szalay-Beko, M; Thomas, JP; Verstockt, B; Watson, A; Tremelling, M; ,DOP07 Ulcerative Colitis associated single nucleotide polymorphisms found in transcription factor binding sites effect key pathogenesis pathways and facilitate patient stratification,Journal of Crohn's and Colitis, 15, Supplement_1,S045-S046,2021, Oxford University Press US

8. Modos, D; Gul, L; Lo, J; **Madgwick, M**; Cozzetto, D; Lord, G; Korcsmaros, T; Powell, N; , P011 New insights into the pathogenic potential and signalling network of NKG2D+ CD4+ T-cells in Crohn's Disease,Journal of Crohn's and Colitis,16, Supplement_1,i141-i141,2022, Oxford University Press US

9. **Madgwick, M**; Sudhakar, P; Korcsmáros, T; ,P027 Machine learning approaches to identify prognosis indicators from microbiome data,Journal of Crohn's and Colitis,13, Supplement_1,S99-S100,2019, Oxford University Press US

10. Pavlidis, Polychronis; Tsakmaki, Anastasia; Pantazi, Eirini; Li, Katherine; Cozzetto, Domenico; Yang, Feifei; Lo, Jonathan; Alberts, Ms Elena; Sa, Ana Caroline Costa; Niazi, Umar; **Matthew Madgwick** .... ,O32 The interleukin 22//neutrophil axis is associated with treatment resistance in ulcerative colitis,,,,,2022,BMJ Publishing Group

11. Pavlidis, Polychronis; Tsakmaki, Anastasia; Pantazi, Eirini; Li, Katherine; Cozzetto, Domenico; Yang, Feifei; Lo, Jonathan W; Alberts, Ellie; Sa, Ana Caroline Costa; Niazi, Umar; **Matthew Madgwick** ...,795: AN INTERLEUKIN 22/CHEMOKINE AXIS DRIVES COLONIC NEUTROPHIL RECRUITMENT AND TREATMENT RESISTANCE IN ULCERATIVE COLITIS,Gastroenterology,162,7,S-192,2022, WB Saunders

12. Kottoor, Sherine Hermangild; Cozzetto, Domenico; **Madgwick, Matthew**; Olbei, Marton; Lo, Jonathan W; Digby-Bell, Jonathan; Ibraheim, Hajir; Constable, Laura E; Pavlidis, Polychronis; Korcsmaros, Tamas; ,792: MICROBIOTA-RESPONSIVE CYTOTOXIC, POLYFUNCTIONAL CD4+ T CELLS ARE ENRICHED IN ULCERATIVE

COLITIS AND ARE ASSOCIATED WITH RESISTANCE TO ANTI-CYTOKINE THERAPIES,Gastroenterology,162,7,S-191,2022,WB Saunders

13. Lo, Jonathan W; Cozzetto, Domenico; **Matthew Madgwick**; Sieh, Jillian Y; Olbei, Marton; Alexander, James L; Blanco, Jesús Miguéns; Kudo, Hiromi; Ibraheim, Hajir; Liu, Zhigang; ,374: IMMUNE CHECKPOINT INHIBITOR-INDUCED COLITIS IS MEDIATED BY CYTOTOXIC LYMPHOCYTES AND IS RELIANT ON THE IL23/IFNγ AXIS,Gastroenterology,162,7,S-78-S-79,2022,WB Saunders

14. Scoones, Anita; Wojtowicz, Edyta; **Matthew Madgwick**; Rushworth, Stuart; Haerty, Wilfried; Macaulay, Iain; 3185–FROM STEM CELL TO MEGAKARYOCYTE: DELINEATING LINEAGE COMMITMENT IN MURINE MEGAKARYOPOIESIS USING SINGLE-CELL RNA-SEQ,Experimental Hematology, 111,,S137, 2022, Elsevier

15. Lo, Jonathan; Cozzetto, Domenico; Liu, Zhigang; Ibraheim, Hajir; Sieh, Jillian; Olbei, Marton; Alexander, James; Blanco, Jesus Miguens; **Matthew Madgwick**; Kudo, Hiromi; ,P38 Immune checkpoint inhibitor-induced colitis is mediated by CXCR6+ polyfunctional lymphocytes and dependent on IL23/IFNg axis,,,,, 2022, BMJ Publishing Group

16. Saifuddin, A; **Matthew Madgwick**; Cozzetto, D; Pavlidis, P; Verstockt, B; Hart, A; Korcsmaros, T; Powell, N; ,DOP71 TREM1, OSM and a co-expressed transcriptional module are core components of the molecular resistome to anti-cytokine therapy in Ulcerative Colitis, Journal of Crohn's and Colitis, 17, Supplement_1,i146-i148,2023,Oxford University Press US

Chapter 1: General Introduction

1.1 Preface

The human gut microbiome plays a vital role in human health. An example where disruptions of the microbiome can lead to increased inflammation and disease pathogenesis is a disorder called inflammatory bowel disease. Due to the nature of the disease, it is difficult to collect biopsies from patients, and therefore, faecal samples provide a non-invasive way to study the gut microbiome as well as the progression of the disease. Currently, methods to investigate and extract biomarkers and prognostic indicators from these datasets remain an active field of research. The majority of current approaches rely on correlation or compositional approaches but these lack mechanistic or functional insight. Furthermore, when applying the same approach to a different dataset, the results can be dramatically different, which points to these approaches' inability to generalise well to new datasets.

This chapter introduces and summarises the literature on fundamental biological concepts of the gut microbiome and meta-omics data. Then, it will outline the current state-of-art methods used in the application of machine learning, bioinformatics and systems biology approaches to human gut microbiome data. Concluding with an overview of inflammatory bowel disease (IBD) as a case study for the application and investigation of the microbiome's contribution to human health. In Chapter 2, I conduct a "classical" analysis of a publicly available dataset and perform exploratory data analysis to outline the current limitations in analytical approaches used to investigate metagenomics. This is twinned with Chapter 3, which shows the development of a new approach to investigating the dynamics of the microbiome with respect to its temporal component. Chapter 4 combines the findings from Chapter 3 to apply dimensionality reduction methodologies to meta-omics data and to predict disease activity increase in inflammatory bowel disease patients. Chapter 5 explains the findings of Chapter 4 using systems biology approaches to gain insights into how host-microbe interaction affects the host. Finally, chapter 6 will summarise the overall conclusions of the thesis, which are discussed, with an evaluation of the methods developed and an exploration of potential future directions and developments to explore in more detail the applications of predicating gut microbes in health and disease.

This BBSRC iCASE PhD scholarship was supported by BenevolentAI. Together, we aimed to develop methods to analyse, predict and interpret the human gut microbiome during healthy and unhealthy conditions. As part of this iCASE project, I worked on placement within the Precision Medicine Product Team at BenevolentAI to build further on the methods developed and described in this thesis. In addition to furthering my professional, research, and personal skills, I have worked closely with BenevolentAI throughout this project to extend my knowledge and understanding of the application of machine learning and data science skills in both research and production settings. BenevolentAI's contribution and support have resulted in the methodologies and analysis described in exploring the temporal dynamics of the gut microbiome outlined in Chapter 3 and used to predict disease activity in inflammatory bowel disease in Chapter 4.

1.2 The human gut microbiome

The human microbiome can be defined as the entirety of the microorganisms that colonise individual sites in the human body; these include the skin, oral mucosa, lung and gastrointestinal tract. As a result of the adoption of DNA-sequencing technologies to investigate, characterise and identify microbes within the human body at the turn of the century, hundreds of previously unknown microbial communities have been discovered. A microbiome is not solely composed of bacterial microbes but also contains a vast number of archaebacterial, protozoan, fungi and viruses (Hill et al., 2014).

The human gut microbiome consists of a vast number and a high diversity of microbes operating within a complex and dynamic ecosystem. The human gut is colonised by commensal and pathogenic bacteria along the entire gastrointestinal tract. Furthermore, it is the largest reservoir of microbes in the human body. The gut microbiota composition continuously evolves either during its development in the early stages of life or through perturbations, such as diet, lifestyle and medication, which can lead to dynamic changes in the abundance levels of specific microbes (Hildebrand et al., 2019; Nayfach et al., 2019).

Although there remain many similarities in bacterial species across individuals, for example, bacterial phyla like Bacteroidetes, Firmicutes and Actinobacteria, the abundance levels of the subpopulations of these bacteria can represent differentially. The role that the diversity

of the microbial communities within the microbiome plays in regulating the host's health is well established and can provide preliminary insights into disease progression and regulation. Consequently, the dysbiotic states of the microbiome and key subpopulations have been suggested to be a critical prognostic indicator for diseases and disorders, such as inflammatory bowel disease, irritable bowel syndrome, type 2 diabetes and atopy (Bull and Plummer, 2014).

Host-microbiota interactions play a key role in maintaining host homeostasis. It is generally accepted that the regulatory effects of health-promoting interactions contribute to a symbiotic microbiota or, conversely, a perturbed system that drives a dysbiotic microbiota. Interestingly, complex and coevolved interdependencies between microbial communities are commonly observed between individuals within the same ecological niches (Alkasir et al., 2017; Filyk and Osborne, 2016). This implies that individuals with the same environmental factors can have contrasting microbiota composition, suggesting that the host's genetics and environmental factors are interacting with the host's gut microbiome and, therefore, contributing to the shift from symbiotic microbiota and healthy host to a dysbiotic microbiota and unhealthy host (Figure 1.1).

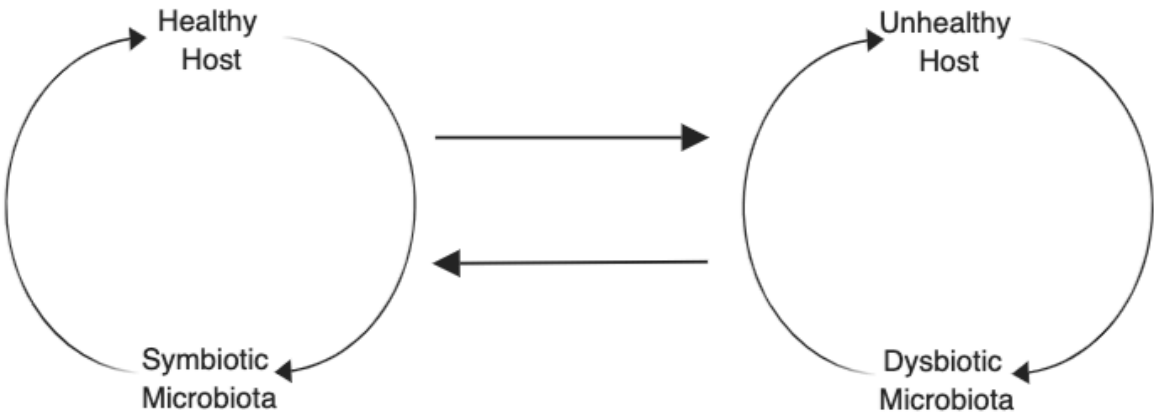


Figure 1.1. Schematic of the regulatory effect of the microbiome on human health.

A healthy gut, therefore, is a balancing act between the gut microbiota composition, host immune response and the physical barrier of the epithelial layer, separating microbes and the host (Figure 1.2.). The intestinal epithelium prevents microbes from leaving the gut and

regulates inflammatory states by warning immune cells of injury or pathogen exposure. Importantly, this means the mucosal surface and its interactions with microbes also contribute to regulating a symbiotic microbiota (Dovrolis et al., 2019; Eckburg et al., 2005).

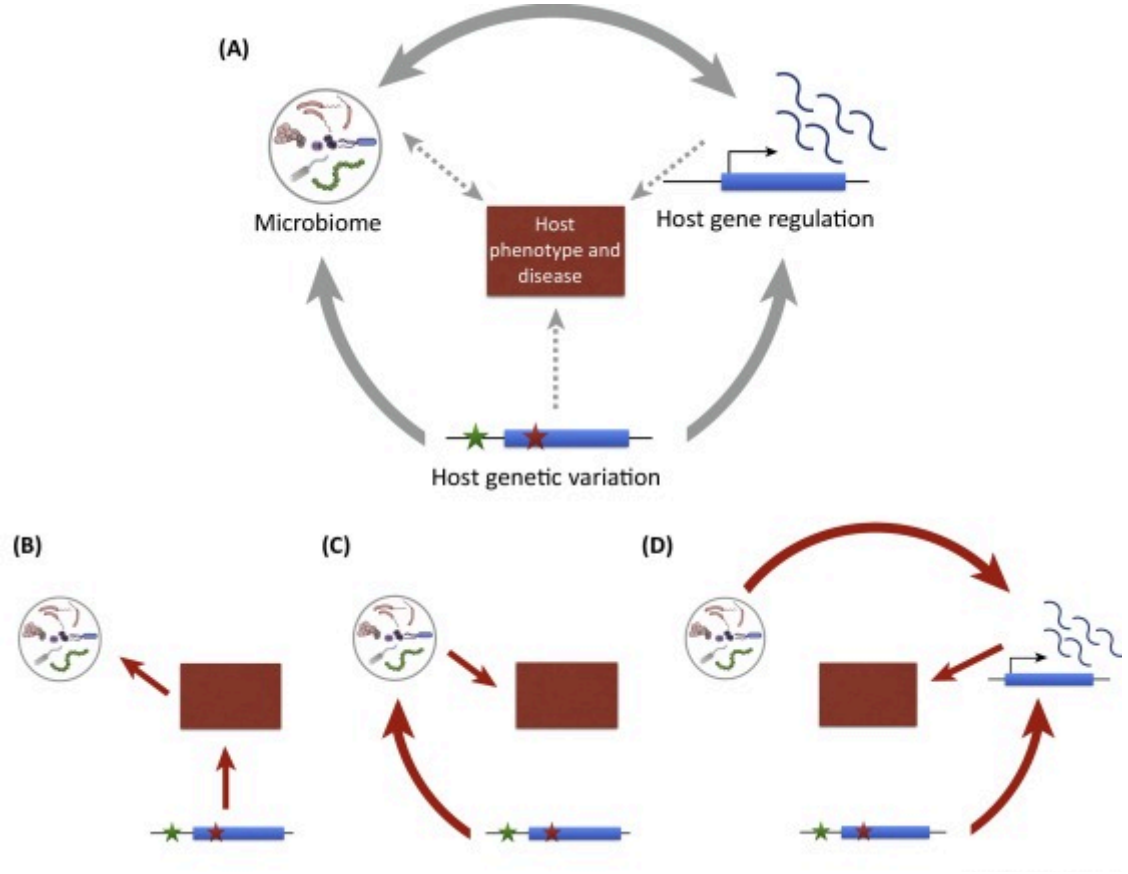


Figure 1.2. Functional effect of host-microbiome interactions in humans. (A) A schematic of how the microbiome influences the host's phenotype through causal/regulatory interactions between itself and the host's genetic/transcriptomic processes. (B,C,D) demonstrates the potential feedback loop between the systems at play. (B) This shows how host genetics directly controls the phenotype, and this in turn can lead to alterations in the microbiome. (C) The host genetics can also control the microbiome first and affect the host's phenotype indirectly through the microbiome. (D) And finally, the host genetic variation leads to different/dysfunctional gene regulation resulting in the microbiome and the host affecting the host phenotype. *Figure adapted from* (*Luca et al.*, 2018)

However, although there is increasing evidence of the microbiome's role in both healthy, acute and chronic disease states, no microbiome-based test has been clinically validated for

either disease diagnosis or treatment (Chiu and Miller, 2019). This is likely due to the microbiome's complexity in such disease pathogenesis. Accordingly, longitudinal studies are required to study the disease to identify more robust prognostic indicators.

1.3 Meta-Omics

1.3.1 Metagenomics

Metagenomics has enabled the characterisation of the microbial communities within the human microbiome and the determination of the relationship between the resident microbiome and invasive pathogens. The data produced by metagenomic studies have contributed to understanding the dynamic nature of microbial communities and the impact these changes have on human health (Malla et al., 2018; Eckburg et al., 2005). There are numerous protocols and tools that can be used to analyse metagenomic data. In this section, the advantages and disadvantages of metagenomics protocols will be outlined, and then bioinformatic pipelines that can be used to conduct downstream analysis of the datasets produced will be highlighted.

The earliest methods to investigate the microbiome used culture-dependent approaches to investigate host-microbe interactions. In culture-dependent methods, samples from patients (humans or animals) are cultured to isolate microbes present within a sample, and then each cultured microbe interaction with co-cultured microbial taxa is studied (Parker and Snyder, 1961; Gibbons et al., 1964). However, this approach not only produced a limited set of microbial taxa and, thus, microbial interactions but also failed to consider spurious interactions that occur within the microbiome (Malla et al., 2018). Accordingly, with the emergence of next-generation sequencing (NGS), culture-independent methods are now the most widely used approach to determine the abundance level of microbes within a community (Strobl et al., 2008; Bent et al., 2007). There are two main culture-independent approaches: (1) 16S ribosomal RNA (rRNA) targeted sequencing and (2) shotgun metagenomic sequencing. In both cases, these approaches cared for small reads, approximately 25-500 base pairs in length, allowing for microbes to be detected, either if they are unknown or in low abundance.

The targeted sequencing of the 16S ribosomal RNA (rRNA) subunit gene is the most commonly used protocol for the identification and classification of microbial taxa within a community (Weinstock, 2012). The 16S rRNA gene has a high degree of conservation (Alves et al., 2018; Tessler et al., 2017), assumed as the result of the importance of the 16S rRNA as a critical component of the ribosome. Thus, the area between the conserved regions of the 16S rRNA varies among bacterial species and is known to be species-specific. However, the 16S rRNA sequencing standard operating procedure dictates a library to be built from the amplification of the variable regions of the 16S gene using multiplex polymerase chain reaction (PCR) primers. This further step adds more uncertainty to this approach, resulting in lower-resolution sequencing results. Nevertheless, 16S sequencing is faster, accessible and inexpensive, therefore better suited for large control and patient-based studies (Dovrolis et al., 2019).

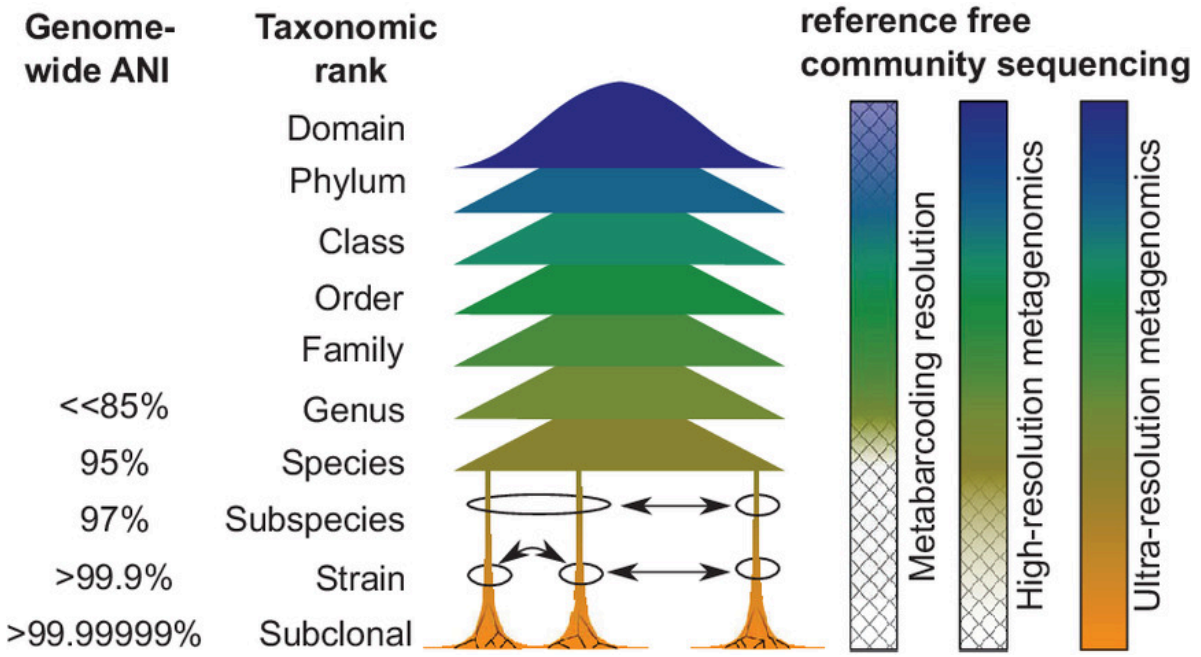


Figure 1.3. Schematic of taxonomic rank vs sequencing depth in metagenomics. Technologies such as 16S rRNS are placed under metabarcoding resolution, while metagenomic approaches and whole genome sequencing are in the high-resolution metagenomics category. *Figure from (Hildebrand, 2021).*

The other culture-independent metagenomic approach is whole genome sequencing (WGS). This approach is considered the best method for identifying and characterising microbial communities as it results in high-resolution metagenomics (Figure 1.3.). This is due to its

ability to provide a much greater level of diversity compared to the targeted approach of 16S sequencing. Shotgun sequencing takes a whole-genome approach by sequencing random string fragments of the DNA sequences and using either common sequences or clade-specific markers to match these fragments to an annotated database of known DNA sequences (Tessler et al., 2017; Alves et al., 2018; Malla et al., 2018). Therefore, shotgun metagenomics is more commonly used when cataloguing genes or making a functional inference (Tessler et al., 2017). In addition to being more expensive, WGS also has the added complexity of the results, including all the microorganisms within the sample, including the host, and thus requires a copious amount of processing power, memory and storage. Metagenomics sequencing remains a very active research space, and there is a need to increase resolution in metagenomic sequencing approaches (Hildebrand, 2021).

1.3.1.2 Bioinformatic pipelines for metagenomic data

The ability to analyse the human microbiome in its entirety, introduced from culture-independent such as WGS, enabled the characterisation of all DNA or RNA present within a sample, resulting in the generation of an enormous quantity of metagenomic data. This, in turn, has transitioned a microbiology and bioinformatics problem into a big data challenge. With WGS producing datasets in the magnitude of Gigabytes (10^9 bytes) per patient, a patient cohort can now easily exceed Terabytes (10^{12} bytes) of data. The standard output of a metagenomic protocol is a taxonomic unit (OTU), which holds information related to clusters of similar sequences (Figure 1.4).

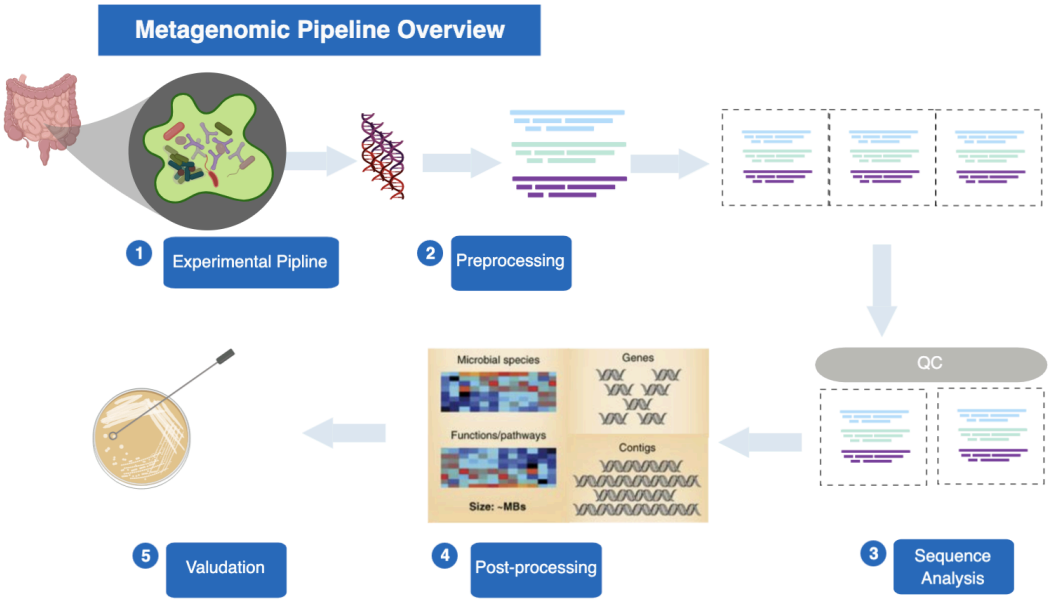


Figure 1.4. A high-level overview of metagenomic methods. The metagenomic pipeline can usually be defined in 5 steps; (1) Experimental pipeline, (2) Pre-processing, (3) Sequence analysis, (4) Post-processing and (5) Validation. During the post-processing stage, further downstream analysis can be conducted. This can include multivariate statistical methods, machine learning (ML) methods and network analysis to interpret the data.

There are many metagenomic pipelines to investigate the microbial composition within an individual sample. The main objective of workflow and tools is to bin each isolated genome into a bin such that functional downstream analysis can be conducted. This can be achieved through two methods; assembly-based or assembly-free (read-based) profiling (Chiu and Miller, 2019). For case-control design, the idea is to determine the encoded functions from the identified species and match them back to the case or control condition, thus suggesting a characterisation of the condition within the sample set.

The bioinformatic challenge of the metagenomics pipeline remains a difficult process as it requires fine-tuning for each dataset; however, with many new models being developed, a number of different advanced algorithms can be used to better determine and explore the parameter space. Moreover, depending on the research questions and the method used to sequence the samples, there are two main methods to analyse the output of the metagenomic protocols; homology- and prediction-based methods (Dovrolis et al., 2019).

These methods are both hybrid implementations combining two different approaches to determine the microbiome composition from small read fasta or FASTAQ files and mapping files (which contain all the metadata required to conduct the analysis). For 16S rRNA data *de novo* and closed-reference OTU picking is used while for shotgun sequencing homology-independent or -dependent binning methods are used (Dovrolis et al., 2019).

An example of tools used to achieve this approach on shotgun raw sequences is HUMAnN2 (Franzosa et al., 2018), which provides species-resolved functional profiles of both host-associated and environmental communities, and MetaPhlAn2 (Truong et al., 2015), which provides methods for metagenomic phylogenetic analysis. These two pipelines are commonly used in combination to investigate the effects within the microbiome in case-control studies. Other tools such as metagenomeSeq, QIIME, Phyloseq and PICRUSt (Paulson et al., 2013; Caporaso et al., 2010; McMurdie and Holmes, 2013; Langille et al., 2013) also allow similar analysis, providing α-diversity, β-diversity, and microbe-microbe associations, which enable the characterisation of the overall properties of a microbiome. Specific algorithms such as Bayesian models to infer environmental factor-microbe association (mLDM) and a large-scale assessment of microbial metabolic interactions (MMinte) (Mendes-Soares et al., 2016; Yang, Chen and Chen, 2017) allow for a more semantic analysis of the microbiome.

From the introduction to metagenomics protocols and analysis pipelines, it is evident that copious amounts of data are being produced. This is particularly the case when studying the disease state and healthy state in a longitudinal study to determine biomarkers for the disease (Vázquez-Baeza et al., 2018). This is framing metagenomic biomarker discovery as a big-data challenge that requires novel analysis methods (Vázquez-Baeza et al., 2018; Luna, Mansbach and Shaw, 2020; Kodikara, Ellul and Lê Cao, 2022).

1.3.3 Metaproteomics

Proteomics is the study of all proteins present expressed in a sample and their functions. Metaproteomics is the extension of proteomics to identify the protein content with microbial communities, for example, in the gut microbiome from a faecal sample. The main advantage of metaproteomics over metagenomics for example is the functional information it provides. In turn, it complements the genetic potential described by metagenomics, enabling the discovery of potential genotype-phenotype linkages (Van Den Bossche et al.,

2021; Issa Isaac et al., 2019). A typical analytical approach to metaproteomics would be 1) extract and purify proteins from the samples, 2) use enzymes to digest the proteins into peptides, 3) perform mass spectrometric analysis on the separated proteins, and 4) identify and annotate proteins using large sequence databases (Kolmeder and de Vos, 2014; Petriz and Franco, 2017; Lee et al., 2017; Issa Isaac et al., 2019).

Metaproteomics leverages the power of mass spectra to identify these microbial communities however, this has some limitations. The size of data produced is often vast as each species contains millions of proteins, which leads to an order of magnitude more peptides to process (Zhang et al., 2018b). This can then result in a large false discovery rate (FDR) during the protein identification stage of the analysis (Zhang et al., 2018a; Van Den Bossche et al., 2021). However, multiple bioinformatic approaches, search algorithms, datasets and ensemble machine learning approaches have been developed to combat this issue (Issa Isaac et al., 2019).

1.3.4 Metabolomics

Like metaproteomics, metabolomics also provides insights into the functional potential of the gut microbiome. The metabolome is widely said to be the closest representation of the phenotype and, therefore, is essential in understanding how cellular processes respond in both healthy and unhealthy conditions (Bauermeister et al., 2022; Vernocchi, Del Chierico and Putignani, 2016; Nguyen et al., 2021; Johnson, Ivanisevic and Siuzdak, 2016). Metabolites are defined as low molecular weight molecules (<1500 Da). These small molecules show both host and microbe activity. In the case of the host, these molecules appear as byproducts of host-microbe co-metabolism involved in the regulation of host metabolic homeostasis (Nicholson et al., 2012; Heinken and Thiele, 2015). Alternatively, molecules act as nutrients for bacterial species within the gut microbiome which can directly affect the overall composition (Oliphant and Allen-Vercoe, 2019).

Once again, mass spectrometry is often used to study metabolomics as it has the ability to process complex biological samples and still quantify a large range of molecules (Bauermeister et al., 2022). This results in large and complex datasets, particularly in the case of untargeted metabolomic studies, which require computational methods to handle and interpret the results. The general approach to processing the result is to 1) feature/speak detection from data, 2) align and normalise the data, and 3) annotate the

results from comprehensive metabolite databases. However, not all metabolites can be annotated from these resources and this remains a major challenge in the field of metabolomics (Johnson, Ivanisevic and Siuzdak, 2016). Another computational after identifying the metabolites within the sample is to infer the biological meaning and their mechanism within the host (Johnson and Gonzalez, 2012; Johnson, Ivanisevic and Siuzdak, 2016).

1.4 Machine learning

Machine Learning (ML) provides the ability to discover hidden structures within datasets. Going beyond the power of traditional statistics, it can achieve this without explicitly being programmed to achieve this task. An example of this could be to predict an outcome from historical data or to determine the cluster of multiple data points in a dataset. Going further still, Deep Learning (DL) provides architectures which operate in a fashion similar to that of the brain through the use of Artificial Neural Networks.

There are three main categories for ML algorithms. Supervised, unsupervised and reinforcement learning. This thesis will focus on supervised and unsupervised learning (Figure 1.5). In supervised learning, the input vector along with the target vector is used to train the model, such that a function can calculate a value for the error. This then alters the function in an attempt to learn the mapping of the data (Bishop, 2006). In unsupervised learning, the training vector only consists of the input vector with no target vector provided. The goal here is to cluster the data into groups, project data from a high-dimension space into a low-dimensional space or determine the distribution of data in an input space (Bishop, 2006).

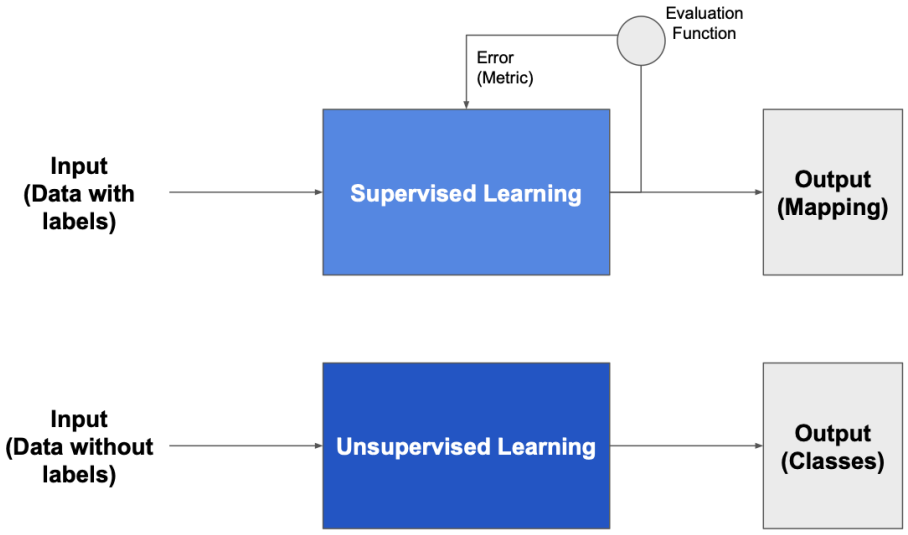


Figure 1.5. Overview of machine learning categories. The two most common classical machine learning strategies. The key difference is the feedback and training loop found in the supervised learning strategy and the input of data with labels defining a description of the data.

1.4.1 Bayesian Models

Bayesian models are based on Bayes' theorem, which describes the probability (p) of an event based on prior knowledge of the conditions that might be related to that event. Bayes' theorem takes the form:

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)} = \frac{p(\theta)p(D|\theta)}{\int p(\theta)p(D|\theta)d\theta} \propto p(\theta)p(D|\theta)$$

In the form above Θ represents a parameter of an unknown quantity. The prior $p(\theta)$ is an estimation of the uncertainty of the parameter Θ is usually guided by domain knowledge, for example, research questions, literature reviews and historical data. D is a vector $\{x_1, x_2, x_3, \ldots, x_n\}$ which represents the collected data in an attempt to gain more information about the unknown parameter Θ. The joint probability of observed data D as a function of Θ

is known as likelihood, $P(D|\Theta)$. The posterior distribution, $P(\Theta|D)$, is a conditional probability that describes the uncertainty about the inference (Bishop, 2006; Casella and Berger, 2001; van de Schoot et al., 2021). The posterior can then be used to make predictions or assumptions based on the research question. Bayes' theorem can be simplified to form below:

$$\text{posterior} \propto \text{likelihood x prior}$$

The life cycle of creating a Bayesian model as described above is repeated and updated based on new domain knowledge by updating the prior or from the collection of new data (Figure 1.6).

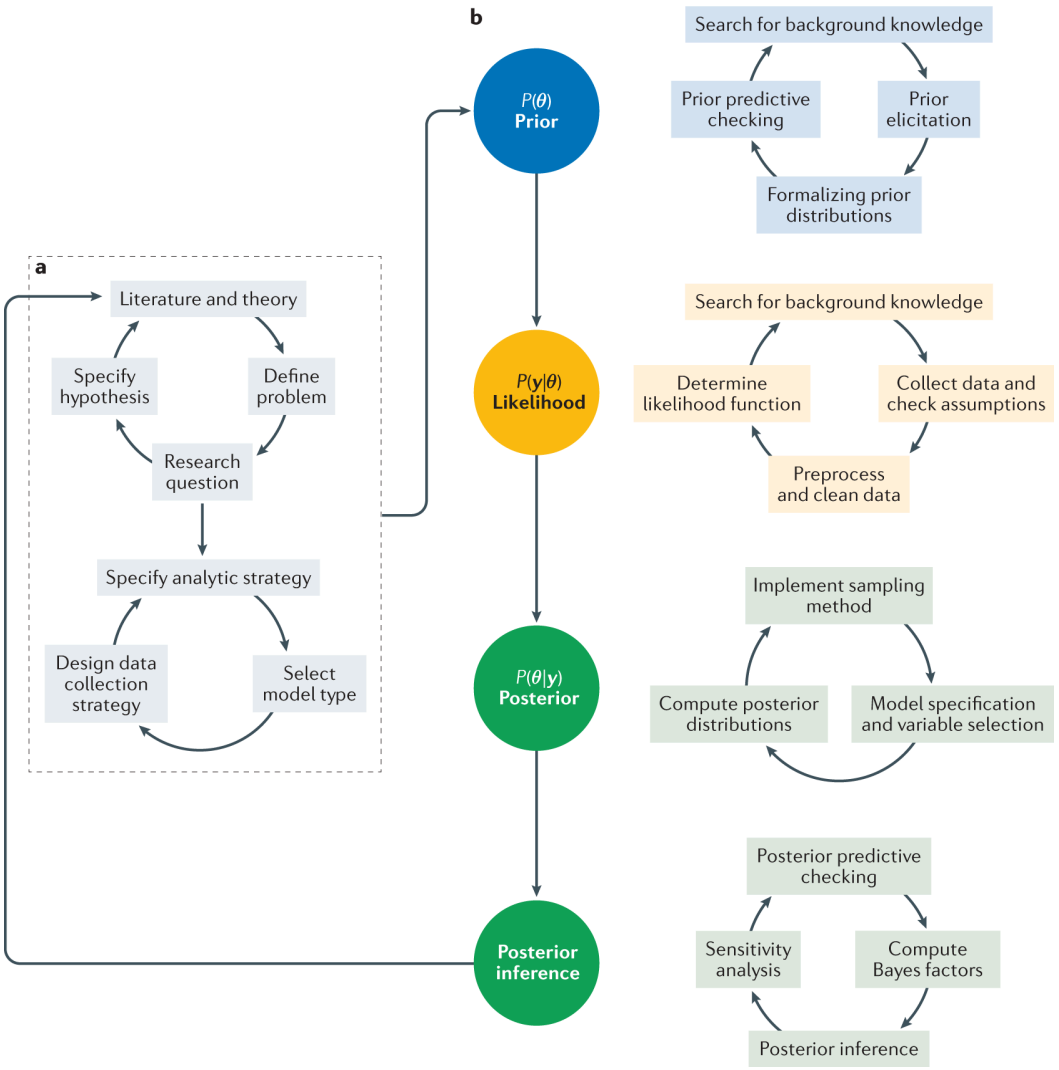


Figure 1.6. Schematic of a bayesian model life system. (A) Demonstrates the importance of background research taken before developing a Bayesian model and how to incorporate this information into the prior. (B) The figure shows the feedback loop involved with the development of Bayesian models. The likelihood: $p(D|\Theta)$. Data: D. Prior: $p(\Theta)$. Posterior distribution: $p(\Theta|D)$. Figure from (van de Schoot et al., 2021).

In both Bayesian and frequentist approaches, Bayes' theorem, and more specifically, likelihood function, plays a vital role in model fitting. In a frequentist approach, the most used approach is the maximum likelihood, where the aim is to set the value Θ to maximise the likelihood function $p(D|\Theta)$. On the other hand, a Bayesian approach estimates the entire posterior distribution Θ. Therefore, the posterior distribution is usually summarised by the mean of the posterior and the credible interval (Bishop, 2006; van de Schoot et al., 2021).

Denote N as the number of instances of evidence we possess. As we gather an infinite amount of evidence, say as $N \rightarrow \infty$, our Bayesian results (often) align with frequentist results. Hence for large N, statistical inference is more or less objective. On the other hand, for small N, the inference is much more unstable; frequentist estimates have more variance and larger confidence intervals. This is where Bayesian analysis excels. By introducing a prior and returning probabilities (instead of a scalar estimate), we preserve the uncertainty that reflects the instability of statistical inference of a small-N dataset.

As the Bayesian model wants to estimate the entire posterior distribution, the direct inference is usually not tractable, particularly for large, highly-dimensional datasets (van de Schoot et al., 2021). This was one of the reasons frequentist statistics became more popular than Bayesian statistics. However, multiple methods have been developed for sampling the posterior distribution and, therefore fitting the models more efficiently. Markov chain Monte Carlo (MCMC) can be used to fit models by indirectly obtaining inference on the posterior distribution. The algorithm samples the posterior distribution where the next sample is dependent on the current sample and thus guides the algorithm to find the values being estimated. This is known as the Markov Chain. This enables the approximation of the posterior distribution without having to sample every variable (Titterington, 1997; van de Schoot et al., 2021). Loosely, MCMC uses the following process to solve Bayesian models; 1) it starts with an initial guess of the parameters, 2) based on the current parameters, generates a new set of parameters from a distribution, 3) then according to the posterior distribution accepts or rejects the new set of parameters and 4) continues to iteratively repeat these steps. The idea is that after many iterations, the Markov Chains will converge to the target posterior and this can be used to approximate the posterior.

1.4.2 Dimensionality Reduction

Dimensionality reduction can be used for multiple different feature engineering, machine learning and statistical analyses. This is most commonly referred to when you have more features than samples in your data. When data has such high dimensionality, it is not only difficult to visualise but also due to the amount of noise and redundancy in the data, it can be challenged to extract statistically meaningful results. The core principle of dimensionality reduction is to transform the data from a high dimensional state to a low dimensionality state while preserving the information present in the raw data (Velliangiri, Alagumuthukrishnan and Thankumar Joseph, 2019). Moreover, the run time complexity of

analysing a large number of features means that the downstream analysis is often not tractable. There are numerous different techniques for dimensionality reduction but some of the most widely used methods are; principal component analysis (PCA), independent component analysis (ICA), non-negative matrix factorisation (NMF), factor analysis (FA) and manifold learning.

Dimensionality reduction is a cornerstone of omic data analysis. This is in part because a typical omic study will have an order of magnitude more biological features (genes, proteins, metabolites, microbes, etc.) than samples but also because of the complexity of visualising biological data (Ma and Dai, 2011). Specific tools have been developed for performing dimensionality reduction on omics data, such as MOFA (Argelaguet et al., 2019) which uses FA to extract biomarkers and other methods for visualisation and analysis like Poincare maps for visualisation of single-cell data (Klimovskaia et al., 2020). Dimensionality reduction methods will be explored further in Chapter 4.

1.4.3 Application of Machine Learning in microbiome studies

To be able to embrace the heterogeneity of the microbiota and thus utilise the robust random processes employed by ML, and to more of an extent DL algorithms, a large number of samples need to be collected. This is even more prominent within biological systems. This can be put down to several different intrinsic factors associated with omic's data. One of which is the phenomenon of the curse of dimensionality (Bellman, 1966). This phenomenon states that when the dimensionality increases, the volume of the space increases so fast that the available data becomes increasingly sparse. Consequently, it complicates ML applications to problems, as the essential task is to learn from a finite number of data samples in a high-dimensional feature space. ML learning algorithms are, therefore, incredibly well suited to finding prognostic indicators across this wealth of data as they leverage the ability to learn the subtle underlying structure within both molecular and clinical datasets.

ML and DL have been used extensively within computational biology, and have many applications within healthcare (Table 1.1.). An example of the application of DL on metagenomic data can be seen in the work of (Arango-Argoty et al., 2018) where the authors developed a DL approach for predicting antibiotic resistance genes from metagenomics data (Arango-Argoty et al. 2018). Here, the authors developed a multi-layered neural network that

utilises a dissimilarity matrix generated from all known antibiotic-resistant genes. The results outperformed other classifiers or search algorithms that produced many false positives (Arango-Argoty et al., 2018).

There is increasing evidence that longitudinal multi-omic studies provide more actionable biomarkers (Schüssler-Fiorenza Rose et al., 2019). Schüssler-Fiorenza Rose *et al.* showed this through the deep-omic profiling of 109 type 2 diabetes mellitus patients over an 8-year period. The authors created predictive models for insulin resistance using Max-Min parents and Child (MMPC) to identify the features within the Bayesian Networks constructed from the integrated dataset. After this feature selection stage, these most informative features were used to create a ridge-regression model, validated through leave-one-out cross-validation. Most significantly, using just the clinical data, the model achieved a cross-validated R^2 of 0.59 (MSE=0.55) and with all integrated data R^2 of 0.87 (MSE=0.16), with the transcriptome, metabolome and microbiome models achieving the highest accuracy of the individual models (Schüssler-Fiorenza Rose et al., 2019).

Table 1.1. Current studies using ML methods can result in a clinically translatable result in IBD. Most of the work around IBD has been on investigating the disease pathogenesis or disease courses. However, this is closely followed by diagnostics and investigating disease severity. Interesting disease subtyping, treatment responses and disease risk lag behind in being actively clinically translatable. This table was modified from work done by Stafford et al. where they investigated how ML methods have been used in investigating IBD in a clinical setting (Stafford et al., 2022).

Task	No. Studies	Chosen ML Models	Data Types Used
Disease Course	22	Bayes Network, Boosting, Decision Tree, Hierarchical Clustering, Neural Network, Partial Least Squares Discriminant Analysis, Random Forest, Regression, Support Vector Machine	Clinical, Gene Expression, Genetic, Imaging, Metabolomic, Metatranscriptomic, Microbiome
Diagnosis	18	Boosting, Hierarchical Clustering, Neural Network, Random Forest, Regression, Support Vector Machine	Gene Expression, Genetic, Imaging, Metabolomic, Microbiome
Disease Severity	16	Bayes Network, Boosting,	Clinical, Gene

		Decision Tree, Hierarchical Clustering, Intelligent Monitoring, Neural Network, Regression, Support Vector Machine	Expression, Genetic, Imaging, Protein Biomarkers
Disease Subtype	8	Boosting, Hierarchical Clustering, Random Forest, Similarity Network Fusion Clustering, Support Vector Machine	Clinical, Gene Expression, Metabolomic, Microbiome
Treatment Response	7	Neural Network, Random Forest	Clinical, Gene Expression, Microbiome
Risk of Disease	6	Ensemble Model, Random Forest, Regression	Clinical, Gene Expression, Genetic
Patient Clustering	4	Gaussian Mixture Model, Hierarchical Clustering, Latent Dirichlet Allocation, Neural Network	Immunoassay, Metagenomic, Online Posts, Questionnaire
Medication Adherence	4	Support Vector Machine	Clinical
Metabolite Abundance	1	Sparse Neural Encoder-Decoder Network	Metabolomic, Microbiome
Identification of Patients	1	Natural Language Processing	Clinical

Furthermore, Haran *et al.* conducted a study which employed all of the currently outlined approaches to investigate the effects of the microbiome on Alzheimer's disease (AD). They looked at the effect of dysregulation of the anti-inflammatory P-Glycoprotein (P-gp) pathway. Following a patient cohort of 108 patients for up to 5 months, taking stool samples, and performing metagenomic sequencing in addition to the metadata of G-gp expression gained from *in vitro* T84 intestinal epithelial cell functional assays. Then combining machine learning approaches using clinical and metagenomics data to identify specific predictors of the bacterial species that lead to the dysregulation of the G-gp pathway. They also differentiated the microbiome of patients with AD and to those of patients without AD. Overall, they observed that patients with AD had a higher proportion of microbes responsible for the synthesis of butyrate and taxa that are linked to proinflammatory

conditions. This study demonstrated the link between intestinal homeostasis by regulating inflammatory pathways and microbial metabolism. However, they didn't look at the effect across multiple clinical layers and, more importantly, the transition from a healthy to a diseased state (Haran et al., 2019).

There is an extensive amount of research in the application of ML methods to the microbiome of IBD patients. This has been in an unsupervised approach with the aim to explore the structure of sub-communities of the microbes or a supervised approach to extract biomarkers. Some of the most commonly explored supervised models include; gradient boosting, random forests, support vector machines and neural networks. This research is not just limited to the methods themselves but also the preprocessing, feature selection, feature engineering and model evaluation stages of the machine learning life cycle. For example, studies have shown that taxonomic data outperforms pathways (Kubinski et al., 2022). Moreover, the same study also highlighted the performance of different normalisation and transformation methods applied to microbiome data, further highlighting the importance of using the correct normalisation method for the model you have selected (Kubinski et al., 2022). Bakir-Gungor et al benchmarked different feature selection methods for biomarker selection from microbiome data. Of the approaches the authors tested, XGBoost (Chen and Guestrin, 2016), Information Gain (Kent, 1983) and Select K Best (Alex et al., n.d.) obtained the highest overall performance. The combination of Select K Best and Random Forest classifier outperforms other methods to predict between healthy controls and IBD patients (0.85 F1-score, 0.93 AUC, and accuracy 88%) (Bakir-Gungor et al., 2022). However, it should be noted that feature selection can result in a reduction in the model's ability to generalise to different datasets, particularly between different cohorts or sequencing technologies.

1.5 Network Biology and Systems Biology

As described in Barabasi *et al.*, a disease rarely results from an abnormality in a single gene or factor. Therefore, in multifactorial diseases, a systems-level approach is required to elucidate the complex perturbations of the intracellular and intercellular mechanisms that link between organs and systems within the body (Barabási, Gulbahce and Loscalzo, 2011; Gosak et al., 2018). Systems biology is a multidisciplinary field, which through a holistic

approach models complex interactions within biological systems (Chuang, Hofree and Ideker, 2010; Tavassoly, Goldfarb and Iyengar, 2018; Gosak et al., 2018).

This can be conducted through computational and mathematical analysis of biological data. One such method to analyse these systems is to model the biological system in a graph data structure, known as a network. In a biological network, nodes represent components of the biological system (e.g. a protein) and the edges represent the relationship between these components (e.g. an interaction). The same is true for metagenomic data, where a node can show taxa and the edge can show the interaction/relationship between other microbes or in fact the host. These interaction networks enable us to determine functional connectivity patterns in multicellular systems. Hence by employing network metrics, mutually exclusive microbes, co-occurring or associations with metadata can be identified. Computational tools provided by network biology enable the systematic transverse of multiple molecular layers of a particular disease, but also the molecular associations among seemingly distinct phenotypes. Besides phenotype classification these methods also allow the identification of disease modules and pathways of these phenotypes (Barabási, Gulbahce and Loscalzo, 2011).

Networks can represent a microbial community structure by integrating multiple types of information and providing the causal relationships between layers allowing for the generalisation of the knowledge. More importantly, microbiome networks have been used in longitudinal studies to determine prognostic indicators. Layeghifard *et al.* used microbiome networks and change-point detection statistical methods to determine the point of change in the distribution of stochastic processes to identify dynamic microbial communities which lead to cystic fibrosis pulmonary exacerbations (Layeghifard et al., 2019). Meanwhile, combining systems biology and machine learning approaches, Lugo-Martinez *et al.* developed a pipeline that enables the integration of longitudinal data across samples to investigate dynamic interactions from networks. This was achieved through a dynamic Bayesian network (DBN), which represents the causal relationships between the clinical and the taxa (Lugo-Martinez et al., 2019). To test their model applied their DBN model on the infant's gut, finding 14 microbial taxa, and 4 clinical and one demographic variables node (Lugo-Martinez et al., 2019).

1.5.1 Graph Theory and Network Science

The key to understanding complex systems is knowing how the system's components interact. One approach is to represent the system as a network consisting of pairwise connections between the components (nodes) and the interactions (edges) between them.

Although graph theory and network science are often used interchangeably, there are subtle differences between the two terminologies. A network refers to a real system, while a graph is a mathematical network representation. For example, we can model the sum of all chemical reactions between a metabolite and a host as a network. However, the mathematical representation we can apply would be a metabolic graph. That is to say, the foundation of a network is underpinned by graph theory. Therefore, there are some overlaps in the terminology between network science and graph theory (Table 1.2.) which can be used interchangeably when talking about networks and graphs.

Table 1.2. Terminology between network science and graph theory.

Network Science	**Graph Theory**
Network	Graph
Node	Vertex
Link	Edge

A network can be directed or undirected. Directed networks have signed interactions and describe a connection between a source node and a target node. In contrast, an undirected network does not have the same signed interactions (e.g. protein-protein interactions). The edges within a network can have attributes applied known as weights. A network is weighted if the edges have weights and unweighted if the edges are not weighted. Finally, a node can also encode additional information. Either by applying a weight, statistic or other attributes to the node.

Once a network has been created, certain metrics can be used to describe the properties of the network. This is often known as the topology of the network. These metrics can be used to compare networks to one another in a global approach or to look into the local patterns within the network. This thesis's main network metrics are *degree*, *hub* and *shortest path*.

The *degree* represents the sum of the total number of links one node has to other nodes. The *average degree* is defined as the total number of edges over the total number of nodes. A node is defined as a *hub* when it has a higher level of connectivity than the *average degree* of that network.

A *hub* is an intrinsic property of a scale-free network and is not observed in a random network generated using Erdős–Rényi model. A network is said to be a scale-free network where its degree distribution follows the power law. However, interestingly, not all biological networks show evidence for being scale-free. Nevertheless, by the nature of a *hub*, it is highly connected within a network, and therefore removing these nodes results in disconnected graphs, i.e. there exist two nodes within the network that are not connected.

The final metric is *path length*, which can be considered a network's "*distance*" metric. A *path* is a journey one would take between linking nodes of a network, and the number of links within that journey is presented as the *length*. The *shortest path* is the fewest number of links between nodes *i* and *j*.

Beyond holistic data analysis, visualising the network can be extremely beneficial. Often allowing for a visual and interpretable representation of a complex system. Further information can be encoded through the representation of nodes (size, shape, colour, label, layout, etc.), edges (thickness, colour, arrow, etc.) and network layout (hierarchical, force directed, etc.).

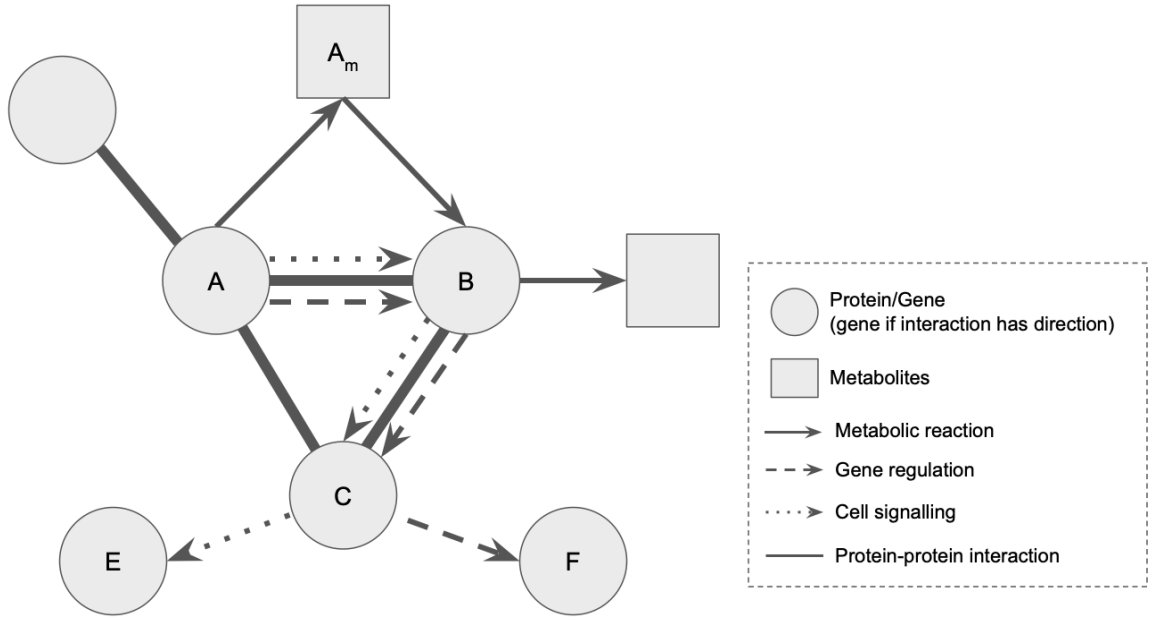


Figure 1.7. Biological interactions are represented as a network. This figure shows how interactions between biological molecules can be represented as a network. The circles and squares represent molecules of interest (nodes), and the connections between them are the interactions (edges). Those that are signed show the direction of the interaction and those without a sign show a potential interaction can occur.

1.6 Biological databases and tools

To utilise the analytical approaches provided by network and systems biology, typically prior or reference data is required. Molecular databases provide essential biological, contextual and domain-specific information to enable not just the identification of biological molecules but also to aid in determining the biological function as well. These databases have been rapidly increasing in numbers, and as of writing, there are over 1700 publicly available biological databases (Imker, 2018).

1.6.1 Sequence databases

The largest and central database for protein sequences and annotations is the UniProt resource (UniProt Consortium, 2021). The aim of UniProt is to provide a knowledge base of all protein sequences with high-quality functional metadata. As of writing, there are

approximately 190 million unique sequences held in UniProt's sequence database UniProtKB, which has almost doubled in past years despite the author's best efforts to reduce the amount of redundancy in the database (UniProt Consortium, 2021).

In addition to the protein sequence and functional annotations, UniProt also holds taxonomy, interactions, subcellular locations, post-translational modifications, expression and other biological database information. For example, UniProt holds extensive gene ontology and alternative identifiers from other databases like PFAM (which is a large-scale, complete and accurate classification of protein families and domains (Mistry et al., 2021)). This makes Uniprot well-suited as a central repository to access any protein information. Table 1.3 outlines UniProt's proteomes summary statistics as of November 2022.

Table 1.3. Uniprot proteomes summary statistics as of November 2022. Up-to-date statistics can be found at https://www.uniprot.org/proteomes **(UniProt Consortium, 2021)**

Proteome Type/Superkingdom	Number
Reference proteomes	22,114
Other proteomes	137,331
Redundant proteomes	282,657
Excluded proteomes	27,603
Bacteria proteomes	349,114
Viruses proteomes	115,399
Eukaryotic proteomes	4,342
Archaea proteomes	3,844

1.6.2 Protein structure databases

Metagenomics, the identification of the composition of the microbiome, frames the potential of the microbiome between conditions. However, within the gut microbiome resides microorganisms which are commensal, symbiotic and pathogenic and under most circumstances, these bacteria and the host are in symbiosis. That is to say, the functional effect of these same bacteria can change during times of dysbiosis. Two key ways these

microbes communicate with the host are; (1) through protein-protein interactions and (2) through the production of metabolites. Therefore, they are key to understanding how the bacterial proteins and metabolites within the gut interact with the host.

For protein-protein interactions, there are two predominant types of interactions. Domain-domain interactions, where a domain of one protein is physically interacting with a domain of the other leading to one protein exerting its effect on the other (Itzhaki et al., 2006). Alternatively, domain-motif interactions occur when a protein domain interacts with a protein-containing motif (Akiva et al., 2012).

In a domain-motif interaction, the protein with the domain exerts its effect on the protein containing the motif. In particular, these interactions are regulated by short linear motifs (SLiMs), which are short amino acid sequences of approximately 3-10 base pairs in length (Brito and Pinney, 2017; Idrees, Pérez-Bercoff and Edwards, 2018). The current standard of the database holding SLiM is the ELM database developed by the European Molecular Biology Laboratory (EMBL) (Kumar et al., 2022).

Domain-domain interactions can be identified experimentally by inferring their three-dimensional structures (Raghavachari et al., 2008). However, it is becoming increasingly common to use a computational approach instead through methods such as sequence co-evolution, phylogenetic profiling, probabilistic frameworks and machine learning approaches (Yellaboina et al., 2011). The largest collection of domain-domain interactions can be found in the Pfam Database (Mistry et al., 2021).

More recent approaches using Deep Learning architectures have yielded more accurate results than competing methods. An example of this is Google's AlphaFold (Jumper et al., 2021) or Evolutionary Scale Modeling (ESMFold) (Rives et al., 2021). Briefly, AlphaFold employs a network-based approach and works by incorporating novel neural network architectures and training procedures based on the evolutionary, physical and geometric constraints of protein structures. While ESMFold utilised a transformer, a large-scale language model, which leverages the improved performance in structural learning and Natural Language Processing (NLP) evaluation methods like perplexity (Rives et al., 2021).

1.6.3 Protein-protein interaction databases

Protein-protein interaction (PPI) databases are a collection of both experimental and *in silico* interactions which have been integrated together to provide fast and efficient access to this data. The most complete resources are STRING (Szklarczyk et al., 2021), IntACT (Del Toro et al., 2022), UniHI (Kalathur et al., 2014) and BioGrid (Oughtred et al., 2021). The key advantage of PPI databases is they give a confidence score to the interaction reflecting the evidence supporting the interaction. The highest confidence interactions come from experimentally obtained interactions and the lowest confidence comes from those that are solely based on predicted interactions.

One of the limitations of PPI databases is that each curation effort takes a different approach, leading to PPI databases holding differing attributes. An example of this would be the introduction of new protein identifiers (ID) as the primary key or in some cases a unique database-specific protein ID. This gives added complexity when performing PPI network analysis downstream as you need to ensure the quality of any ID.

A database that aims to solve this issue is Omnipath (Türei, Korcsmáros and Saez-Rodriguez, 2016; Türei et al., 2021). The Omnipath database (https://omnipathdb.org/) is a large collection of more than 100 resources that have collected and standardised the data. The standardised data is then held in five different knowledge bases (sub-databases); network, enzyme-PTM, Complexes, Annotations and Intercell (Türei, Korcsmáros and Saez-Rodriguez, 2016; Türei et al., 2021). The database has an Application Programming Interface (API) to request data but is also available as a python, R and Cytoscape package.

1.6.4 Metabolic pathway resources databases

There are multiple large databases used for metabolite identification including; HMBD, METLIN, GMD and MassBank (Wishart et al., 2007; Smith et al., 2005; Vinaixa et al., 2016; Horai et al., 2010). However, as of present, they lack high-quality interaction databases for metabolomics. Typically, metabolomic pathways have been used to fill this gap. A database such as BioGRID (Oughtred et al., 2021), KEGG pathways (Kanehisa et al., 2023), and MetaCyc (Caspi et al., 2014), provide manually drawn pathways to aid in mapping metabolomic signatures to functional and regulatory mechanisms. More recently, a new database was released called gutMGene (Cheng et al., 2022), which provides a manually curated database

of microbial gene and microbial metabolites interaction through potential intermediate targets.

1.7 Inflammatory Bowel Disease

Inflammatory bowel disease (IBD) is a chronic multi-systemic inflammatory disorder of the gut. There are two distinct disorders which encapsulate IBD; ulcerative colitis (UC) and Crohn's disease (CD) (Roda et al., 2020; Ungaro et al., 2017; Kobayashi et al., 2020). Although often grouped together, the two diseases differ in pathophysiology, symptoms, complications, therapeutic management and disease course. More specifically, CD presents with patchy lesions throughout the gastrointestinal tract. In contrast, UC presents mucosal inflammation, starting at the rectum and continually propagating throughout the colon (Kobayashi et al., 2020). A key difference between the two diseases' pathophysiology is that the inflammation is typically restricted to only the mucosal layer in UC. In contrast, in CD, the inflammation can affect all layers of the bowel, which results in added complications, such as fibrosis, fistulas and strictures.

The exact pathogenesis of UC and CD is still unknown, however, multiple factors have been implicated in the disease development (de Souza and Fiocchi, 2016). These factors include a dysregulated immune system, genetic factors, alterations in the gut microbiota (microbes, fungi and viruses abundances) and external factors (environment, diet, therapy etc.) (de Souza and Fiocchi, 2016). Each of these factors contributes in part to disease pathogenesis in IBD (Figure 1.8.). However, the complex interaction between these factors results in IBD is not completely understood (Roda et al., 2020; Kobayashi et al., 2020).

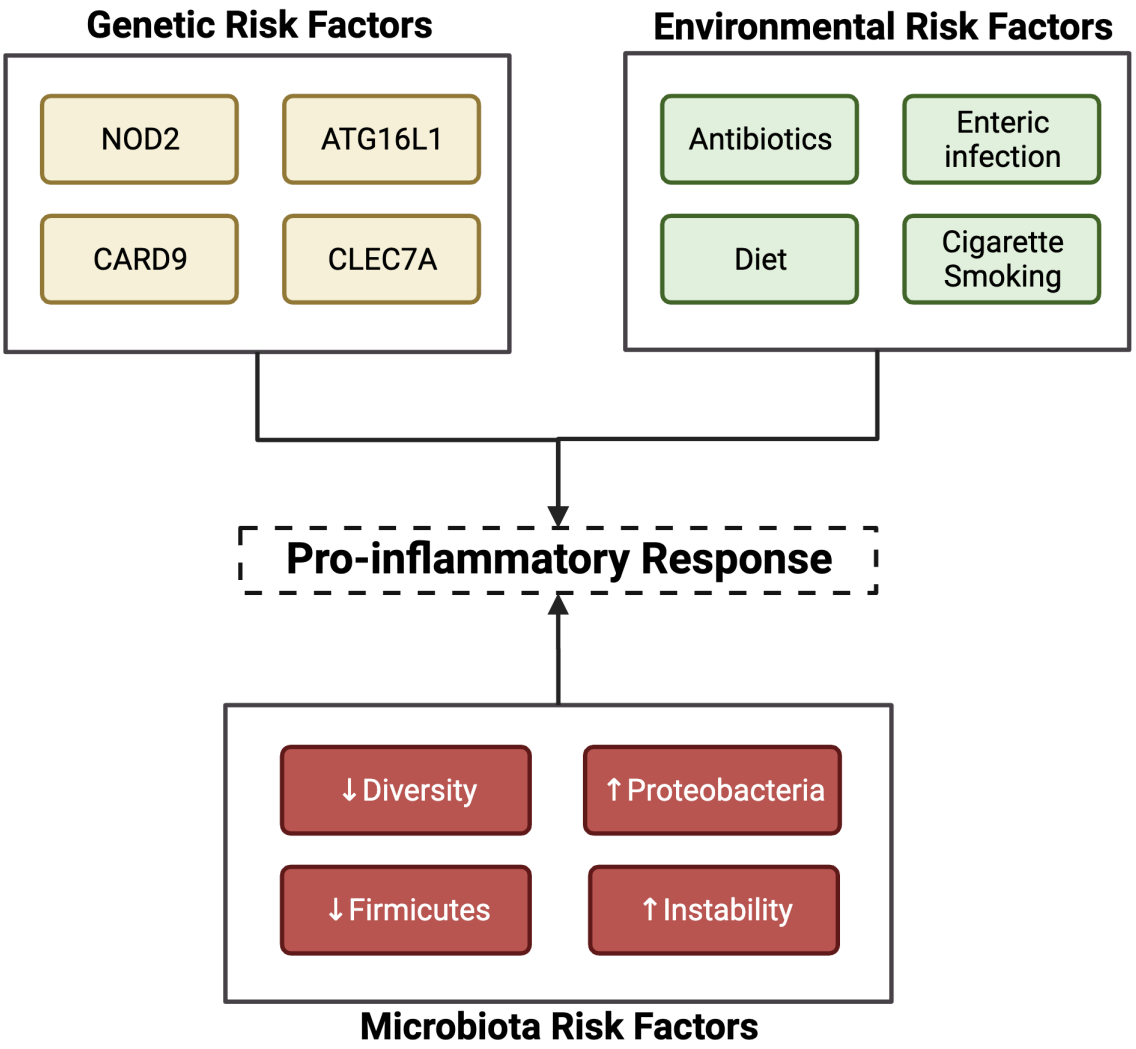


Figure 1.8. A high-level overview of the multifactorial nature of IBD. IBD is considered to have 3 main risk groups. Genetic risk factors, environmental risk factors and microbiome-related risk factors. These different risk categories together lead to the pro-inflammatory response.

The incidence and prevalence of both UC and CD are rapidly increasing worldwide. With both diseases being defined as progressive diseases (i.e. an individual's disease will spread or get worse), IBD is putting an ever-increasing strain on healthcare systems worldwide. As of present, there is no known cure for IBD.

1.7.1 Gut bacterial composition in IBD

Since the implication of the microbiome in IBD disease development, gut dysbiosis (i.e. the alterations in the gut microbial composition) has been studied extensively over the past decade to try and determine if there are a defined microbiota composition or marker microbes that are specific to CD and UC (Glassner, Abraham and Quigley, 2020). Studies have shown how the gut microbiome differs between IBD patients and healthy controls. These studies demonstrate the reduction in microbiome diversity, lower levels of abundance of anti-inflammatory taxa and an increase in invasive bacterial species (e.g. *Escherichia coli*) (Glassner, Abraham and Quigley, 2020; Lee and Chang, 2021).

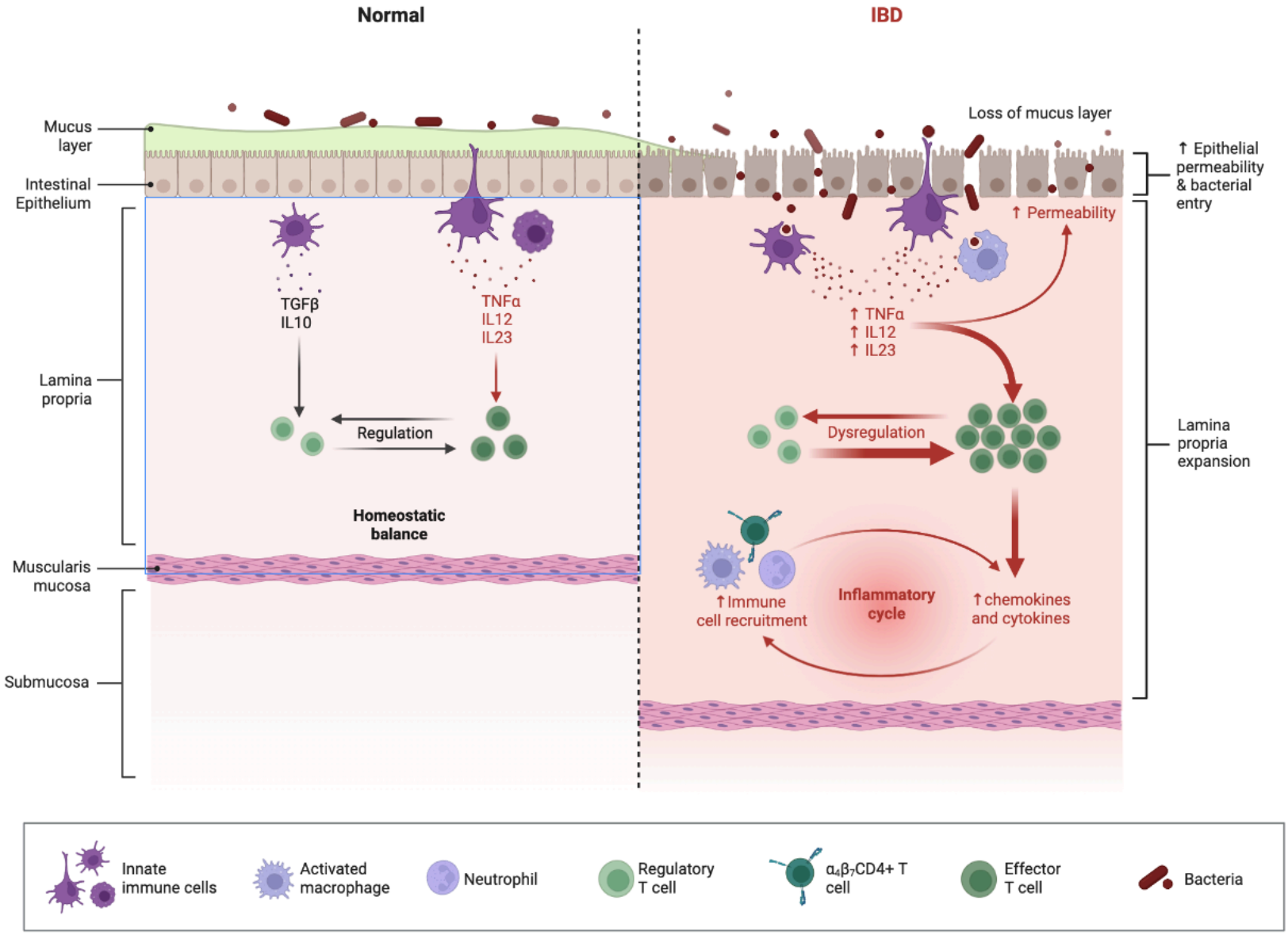

Normal
IBD
Mucus layer
Intestinal Epithelium
Lamina propria
Muscularis mucosa
Submucosa
Loss of mucus layer
↑ Epithelial permeability & bacterial entry
↑ Permeability
TGFβ
IL10
TNFα
IL12
IL23
↑ TNFα
↑ IL12
↑ IL23
Regulation
Dysregulation
Lamina propria expansion
Homeostatic balance
↑ Immune cell recruitment
Inflammatory cycle
↑ chemokines and cytokines
Innate immune cells
Activated macrophage
Neutrophil
Regulatory T cell
α4β7CD4+ T cell
Effector T cell
Bacteria

Figure 1.9. Compared to a healthy gut, a schematic and overview of the pathophysiology of IBD. In the healthy condition (left), a thick and intact mucus layer acts as a barrier between the gut and the intestinal epithelium. However, in patients with IBD (right), this layer of protection is missing, leading to bacterial invasion of the intestinal epithelium. In combination with a dysregulation of the host's immune system, which leads to a pro-inflammatory response (Figure adapted from BioRender).

When compared to healthy controls, specific changes in the gut microbiome composition have been identified. Within CD patients, a reduction in the abundance of Firmicutes and Bacteroidetes and an overrepresentation of Enterobacteria has been characterised in the microbiota. Furthermore, CD patients have seen an increase in pro-inflammatory bacteria such as *Escherichia coli* and in a reduction in anti-inflammatory bacterial species such as *Faecalibacterium prausnitzii* (*Quévrain et al.*, 2016). UC studies have linked *Akkermansia muciniphila*, and also the genus of bacteria *Desulfovibrio* and *Clostridium* (*Manichanh et al.*, 2012; *Bajer et al.*, 2017).

Table 1.4. Bacterial species extracted from the literature whose change in abundance levels has been implicated in IBD (CD and UC) compared to healthy control.

Increased abundance in IBD	**Decreased abundance in IBD**
Fusobacterium species	*Bacteroides* species
Pasteurellaceae	*Bifidobacterium* species
Proteobacteria	*Clostridium* XIVa, IV
Ruminococcus gnavus	*Roseburia* species
Veillonellaceae	*Sutterella* species

The microbiota has also been implicated in the disease progression as well as the disease development. For example, when looking into the disease activity of IBD patients, studies have linked two locations of the gastrointestinal tract where the bacterial population is the highest (i.e. the colon) and where the faecal matter remains at equilibrium (i.e. the terminal ileum and rectum). *Cloony et al* investigated the microbiome variance in patients during inactive and active states defined by the clinical marker faecal calprotectin (inactive $\leq$ 250

μg/g; active > 250 μg/g). The authors used a random forest and a ratio of two-time points to implicate *Bifidobacterium* and *Streptococcus* bacteria as the top determinants between active and inactive UC (Clooney et al., 2021). The same analysis was performed in CD, which suggested that *Hydrogenoanaerobacterium saccharovorans* and *Clostridiales* were the top contributors to stratifying by disease activity (Clooney et al., 2021). Nonetheless, it still remains unclear whether these shifts in composition in a dysbiotic state are causative or a response to the increase in intestinal inflammation.

The current therapeutic practice focuses on regulating the host's immune system through the use of include mesalazine, corticosteroids, thiopurines, methotrexate, ciclosporin, anti-TNF, vedolizumab, ustekinumab, tofacitinib and antibiotics (Lamb et al., 2019). These approaches largely ignore the role of the microbiome in disease pathogenesis. The potential for the microbiota to act as a therapeutic intervention has shown great promise since the introduction of faecal microbiota transplantation (FMT) in IBD patients (Costello et al., 2017; Sokol et al., 2020; Shen et al., 2018). FMT aims to reset the entire microbiome in an IBD patient from a healthy individual's faecal sample. Another treatment which is being increasingly used in the clinic is the use of probiotics. Probiotics aim to help restore symbiosis in the gut by inhibiting pathogenic bacteria, aiding the restoration of the disturbed mucosal barrier and enhancing the intestinal barrier function (Sartor, 2006; Shen et al., 2018).

1.7.2 Metaproteomics studies in IBD

Studies have suggested that only limited variance can be explained by the microbiome composition alone in IBD patients. Although metagenomic outlines the genetic potential of the microbiome in IBD patients, exploring what happens functionally during IBD could reveal associations between different microbial taxa as well as the host. Therefore, there has been an increasing focus on exploring the metaproteome present in the gut of IBD patients (Lehmann et al., 2019).

Previous studies have investigated the functional potential by investigating the pathways associated with the annotated metaproteomes. For example, a twin study extracted metaproteomics data from six pairs of twins that were either healthy or had CD observed an increase in carbohydrate transport and metabolism, an increase in host-bacterial interactions and an increase in host-secreted enzymes (Erickson et al., 2012). Comparing

IBD patients' metaproteomes to control studies have found associations between the reduction of RprY protein from *Bacillus fragilis* in both UC and CD (Lehmann et al., 2019). Moreover, *Mills et al.* demonstrated how *Bacteroides vulgatus* proteases are overabundant in UC patients. To validate this they used an monocolonised IL10-deficient mice model was with *Bacteroides vulgatus* and found that mice given broad spectrum-protease prevented colitis further suggesting the role of overabundant *Bacteroides vulgatus* proteins play in UC (Mills et al., 2022).

1.7.3 Metabolomics studies in IBD

The metabolome has been extensively researched in IBD. There have been six main areas of biosample research; urine, blood (plasma or serum), tissue, breath and stool. However, in this section, the focus will be on metabolites extracted from stool samples (Gallagher et al., 2021). Metabolomics has the potential to link and reveal the underlying mechanisms between the microbiota and the intestinal mucosa (Thomas et al., 2022). There are currently three main candidates for IBD-related metabolites; Bile acids, Short Chain Fatty Acids (SCFA) and Tryptophan (Zheng, Wen and Duan, 2022).

Bile acids have been shown to be perturbed in IBD patients compared to the health control, with IBD patients having a reduction in both primary and secondary metabolites (Weng et al., 2019; Franzosa et al., 2019). Conversely, other studies have suggested bile acids are increased within IBD patients when comparing dysbiotic and non-dysbiotic microbiomes (Lloyd-Price et al., 2019; Gallagher et al., 2021). This contradiction in results can be explained when looking at integrating these samples with paired metagenomic samples, as bacterial species associated with an increase in bile acid production were also increased in abundance in these samples (Gallagher et al., 2021). The resulting shifts in the microbiome composition and bile acid production have also been seen in blood-based metabolomics. Work done by Roda et al, where CD patients with impaired primary and secondary bile acid production saw an increase in production post-treatment of anti-TNF patients (Aden et al., 2019; Roda et al., 2019).

Another class of metabolites which have seen marked changes in IBD are SCFA. SCFA are a byproduct of microbial fermentation in the gut. Compared to healthy controls; Acetate, propionate and butyrate have been found at lower concentrations in IBD patients (Machiels et al., 2014; De Preter et al., 2015; Bjerrum et al., 2015). For example, SCFA like butyrate is

reduced in active IBD and associated with the reduction of bacterial species *Roseburia inulinivoransa*, which is known to be an SCFA-producing bacteria (Bjerrum et al., 2021; Aden et al., 2019). Conversely, when anti-TNF is given to the patient butyrate levels increase and a reduction in inflammation is observed (Aden et al., 2019). As well as the anti-inflammatory effects, SCFA can act as an energy source for the host cell. Once again, a good example of this is Butyrate which is also a primary source of energy for colonocytes (epithelial cells in the colon) (Litvak, Byndloss and Bäumler, 2018; Parada Venegas et al., 2019).

The final class of metabolites we will discuss here are amino acids. From stool samples, patients with IBD have increased levels of both amino acids and branched-chain amino acids when compared to healthy controls. It is considered that due to increased inflammation and therefore intestinal instability in IBD patients there is a reduction in the gut's ability to effectively digest food (malabsorption) (Marchesi et al., 2007). During increased disease activity, tryptophan metabolism also increases leadings (Nikolaus et al., 2017).

Table 1.5. Stool metabolites associated with IBD. A summary of 11 stool-based metabolomics studies in IBD and the aggregated results of metabolite class changes in IBD data compared to controls **(Gallagher et al., 2021)**

Metabolite Class	**Increase/Decrease in IBD**
Lipid classes	Increased
Amino Acids (Alanine, Glycine, Lysine, Phenylalanine, Taurine, Tyrosine)	Increased
Primary and secondary bile acids	Decreased
Branched-chain amino acids	Decreased
SCFA	Increased

1.8 Mathematical Notation

Uppercase letters such as X are matrices and X_{ij} is the i-th row and j-th column of matrix X.

A matrix $X_{i,}$, states the i-th row of that matrix as a vector of length D.

Lowercase letters such as x are vectors and x_i is the i-th element of vector x.

$\sum_{i=a}^{b} x_i$ is just a for-loop that iterates x from a to b, summing all the x_i.

Notation $f(x)$ refers to a function called f with an argument of x.

The dot product w · x is the summation of the element-wise multiplication of the elements, such that $\sum_{i}^{n} (w_i x_i) = sum(\mathbf{w} \otimes \mathbf{x})$.

{ } is a set of elements and is a [ ] vector of elements.

$\{X_k\}_{k=1}^{K}$ represents a set of matrices of length K where the k-th element of the vector is a matrix.

$\mathbb{R}^{NxD}$ are real numbers of size N rows and D columns.

$\mathbb{N}^{NxD}$ are natural numbers of size N rows and D columns.

H_0: gives the null hypothesis, while H_a gives the alternative hypothesis.

$NORMAL$ represents a normal distribution.

$BETABINOMIAL$ represents a beta-binomial distribution.

$a \sim$ represents simulation of a vector given some distribution and any interactions terms.

1.9 Aims

My PhD project aims to develop an integrated machine learning-based systems biology workflow to analyse gut microbiome data and identify prognostic indicators of healthy and unhealthy conditions, using IBD as a case study. The approach is based on gut microbiome data (e.g. metagenomics, metatranscriptomics, metaproteomics and metabolomic data), which capture the composition and functional potential of the microbiome in modulating host processes. Machine learning (ML) can efficiently model microbiome interactions as ML can (1) learn novel features (by the automatic discovery of "regularities" without relying on a priori knowledge); (2) capture multiple features (strains, proteins, pathways, etc.) and model these for prediction; (3) quickly learn complex patterns across large datasets.

Combining ML-based features with host-microbiome interactions and systems biology (SB) will improve our understanding of how microbiota contribute to health using generated microbiome datasets. The project outcomes are expected to overcome critical challenges, leverage existing data, and contribute towards BenevolentAI's efforts in creating ML-based solutions for inflammation-related disease treatments.

The research hypothesis of this project is that the machine learning-based analytical pipeline utilising sequence and systems biology information will identify microbiome-related features implicated in the transition between healthy and unhealthy conditions in inflammatory bowel disease (IBD).

1.10 Objectives

Objective 1: Predict dynamic changes in critical features during the transition between healthy and unhealthy conditions. Benchmarking and testing existing tools and developing new methods to fill the gaps identified.

Objective 2: Identifying condition-related features in microbiome data. Available microbiome data will be collected and analysed using various ML approaches to identify critical communities/microbial products upon the switch between healthy and unhealthy conditions.

Objective 3: Combining systems biology with the developed ML approaches to identify microbiome-host mechanisms. Analyse the ML-based communities and proteins and predict changes using bioinformatic workflows developed previously at the Korcsmaros group to infer microbe-host interactions (Sudhakar et al., 2019).

Objective 4: Create an automated ML-SB pipeline for reproducibility and scalability when running on a complex condition-related dataset. Create proper documentation and integrate codes into a unified pipeline for repeated use with similar datasets or projects initiated by BenevolentAI.

Chapter 2: Exploratory data analysis on longitudinal metagenomics samples using traditional microbiome analysis methods

2.1 Introduction

It has been well studied how the microbiome of Inflammatory Bowel Disease (IBD) patients differs from healthy controls or non-IBD patients (Chapter 1, Table 1.2). The difference between a healthy and a diseased (dysbiotic) microbiome can be measured in many different ways. The data extracted from these high-throughput DNA sequencing studies can be represented as counts, proportions or as ratios. One such approach is compositional data analysis (CoDa) (Gloor and Reid, 2016; Mandal et al., 2015; Greenacre, Martínez-Álvaro and Blasco, 2021). CoDA differs from more standard approaches as it describes the dataset as an arbitrary sum (Aitchison, 1982).

A lot of microbiome studies rely on relative abundance (or proportions). Although this is suitable for some analyses, if one would like to apply an approach that uses Euclidean distance, the resulting representation of the data could induce biases, which would lead to incorrect conclusions being drawn (Ricotta, 2021). Therefore, it is generally accepted that for a CoDa approach, the counts or proportions need to be a transformed ratio between all parts (Gloor and Reid, 2016). An example implementation of this is the centred log-ratio (CLR) transformation (Aitchison, 1982) :

$$clr(X) = (log(x_1/g_x),\ log(x_2/g_x) \ldots,\ log(x_D/g_x))$$

(Equation .2.1)

where $g(x)$ is the geometric mean of all values in the vector X. CLR is a transformation method that can be used to remove the constraint that is present in compositional data. This enables the data to be used by statistical methods and other downstream approaches and is a fundamental tool used by researchers to explore the complexities of compositional data (Faith, 2015). This approach would be robust if microbiome data were not sparse. The

sparsity of the data is problematic for these transformation algorithms as they cannot compute the geometric mean if the vector they are being applied to is 0 (Gloor and Reid, 2016; Mandal et al., 2015). Methods have been developed to address these issues, for example, robust centre log ratio (RCLR), which accounts for the sparsity of microbiome data sets (Martino et al., 2019). However, this transformation requires changes to the ordination algorithm used, as it treats 0 as missing. Another approach would be the use of a Bayesian-based approach. Here, the parameters and transformations can be made in the model to account for the over-dispersed and zero-inflated count matrix (Sankaran and Holmes, 2019; Zhou et al., 2022) (this will be explored further in Chapter 3).

After the correct normalisation and transformation of the data, the next approach could be to find differences between samples or groups of phenotypes. A good first measure is to assess the diversity of the microbiome. This can be done using alpha or beta diversity. Alpha diversity observes the number of taxa present in a sample. The simplest example of this metric is richness, which is defined as the total number of different species within the sample. Meanwhile, beta diversity measures the variability of the communities of taxa by calculating the dissimilarity or distance between samples. The resulting measures can then be used in ordination or by clustering methods to try and group similar samples together (Walters and Martiny, 2020). These measures are important and fundamental to human gut microbiome analysis, as associations between healthy and unhealthy conditions, because they provide insights into the differences and similarities in microbial composition. By understanding how microbial communities vary from person to person or in different conditions, patterns and factors that might influence health and disease can be identified (Manor et al., 2020; Hou et al., 2022). Beta diversity thus plays a crucial role in elucidating the complex interactions within the gut microbiome and how these interactions might be linked to various health outcomes, dietary habits, environmental exposures, or disease states. This level of analysis is essential for advancing personalised medicine and developing targeted interventions to modulate the gut microbiome for better health outcomes (Petrosino, 2018; Cammarota et al., 2020).

2.1.1 Aims

In this chapter, I develop a preprocessing pipeline to enable the fast, efficient and structured preprocessing of metagenomic datasets and then apply this pipeline to publicly

available IBD and healthy control datasets. I will then perform exploratory data analysis on this dataset using both typical data science and microbiome analysis approaches. This chapter aims to:

- Develop a flexible and scalable metagenomic preprocessing pipeline
- Perform exploratory data analysis on longitudinal metagenomics samples from UC, CD patients and healthy controls to gain a greater understanding of longitudinal microbiome data in IBD
- Describe and identify the potential limitations of using traditional microbiome analysing approaches on a longitudinal dataset

2.2 Methods

To assess the variability of the microbiome over time in IBD patients, we used the largest publicly available longitudinal metagenomic study available, created by *Lloyd-Price et al.* This patient cohort consists of 132 individuals who were recruited as part of the Human Microbiome Project (HMP) (Human Microbiome Project Consortium, 2012; Lloyd-Price et al., 2019). The patients were from four US hospitals and were made up of three paediatric and two adult cohorts. In total, the authors collected 1,785 stool samples along with various meta-data, including disease activity metrics, diet, therapy, disease age and more (Figure 3).

Table 2.1. Patient Cohort breakdown per sub-disease. A number of patient data were extracted, including the total cohort, and then patients were removed if they did not have enough metadata, sequencing depth, or enough time points (t > 4).

Disease	N patients before QC	N patients after QC
Crohn's Disease	66	50
Ulcerative Colitis	38	30
Healthy controls (non-IBD)	27	27

2.2.1 Data preprocessing

Raw reads were downloaded from SRA BioPorject PRJNA398089. Quality control was performed using KneadData (https://github.com/biobakery/kneaddata), and the reads that mapped to the human genome were first filtered out (although the number of human reads mapped was kept as a quality control metric). To assign taxonomic profiles to the shotgun metagenomes MetaPhlAn3 was used (Human Microbiome Project Consortium, 2012; Truong et al., 2015; Beghini et al., 2021). MetaPhlAn3 is a shotgun sequencing data-specific tool that uses a library of clade-specific markers to provide pan-microbial (e.g. bacterial, archaeal, viral, and eukaryotic) profiling from a database of ~17.000 reference genomes (Truong et al., 2015; Beghini et al., 2021). MetaPhlAn3 is a fast and efficient way to accurately estimate the number of microbial DNA captured in a sample and map that DNA sequence to a taxa. This is one of the key advantages of MetaPhlAn3 compared to k-mer-based tools, as it achieves very similar accuracy with significantly reduced memory (Yang et al., 2021a). Finally,

MetaPhlAn3 has a large and active community and support network, making it ideal for the pipeline to be used in a production setting due to its reliability and robustness, which come from such open-source projects.

For functional profiling, I used the companion tool to MetaPhlAn3, called HUMAnN3 (http://huttenhower.sph.harvard.edu/humann3). HUMAnN3 constructs a sample-specific reference database from the pangenomes of the subset of species detected in the sample by MetaPhlAn3 (pangenomes are precomputed representations of the open reading frames of a given species) (Beghini et al., 2021). Sample reads are mapped against this database to quantify gene presence and abundance on a per-species basis. A translated search is then performed against a UniRef-based protein sequence catalogue (Suzek et al., 2015) for all reads that fail to map at the nucleotide level. The result is abundance profiles of gene families (UniRef90s), for both metagenomics and metatranscriptomics, stratified by each species contributing those genes, which can then be summarised to higher-level gene groupings such as ECs or KOs.

To ensure a reasonable read depth in each sample, only samples (metagenomes and metatranscriptomes) with at least 1 million reads (after human filtering) and at least one non-zero microbial abundance detected by MetaPhlAn3 were used in downstream analyses. In total, this resulted in 1,595 metagenome profiles across all patient cohorts. The final preprocessing step was to remove individuals with inconsistent metadata or time points. Individuals were removed if they did not meet the following criterion: 1) had fewer than 4 times points and 2) did not have a disease activity index present in more than 4 times points. This left 107 individuals for downstream analysis (Table 2.1.)

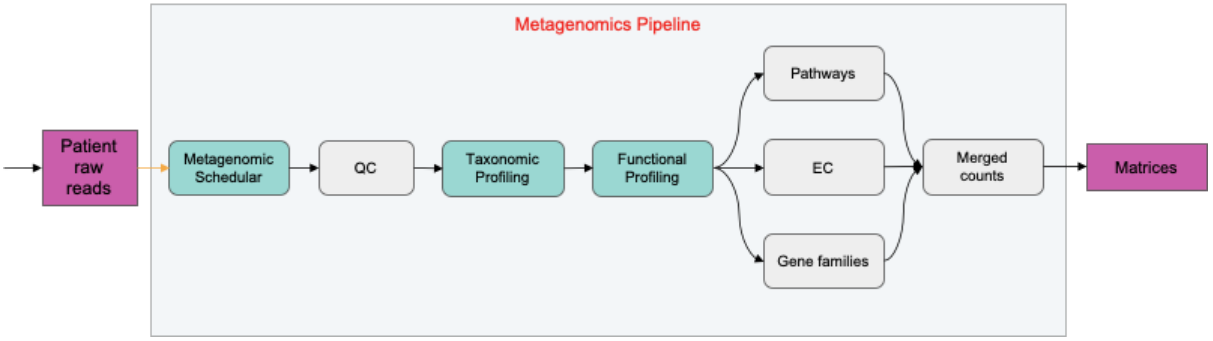


Figure 2.1. Metagenomics workflow with a custom scheduler to take raw reads as input and output annotated count matrices for downstream analysis. This schematic

represents a single run of a parallelised implementation. The metagenomic workflow is used to process the raw reads, perform quality control, conduct taxonomic and functional profiling, and export the data. The output of this workflow is the count's matrices for pathways, enzymes and gene families.

2.2.2 Exploratory data analysis

Before building the new model, I first performed exploratory data analysis (EDA) on the resulting data from the pipeline outlined in Methods 2.2.1 (Figure 2.1). EDA is primarily used to see what data can reveal before or even after modelling or hypothesis testing. Therefore, EDA provides insights into the relationships between features within the data and helps determine if the desired modelling or statistical analysis techniques are appropriate for the dataset.

Due to the data from *Lloyd-Price et al.* being collected across multiple locations, it required a lot of metadata wrangling, cleaning and processing. There was a wealth of metadata collected for this study, but the majority of the clinical metadata was either undersampled (e.g., for patients with fistulas of the 3292 samples collected, only 6 patients presented with a new fistula over the course of the study), or too general (e.g. therapy for antibiotics being boolean variable and not giving further information about what extract treatment the patient received).

2.2.2.1 Compositional Abundance

To assess the compositional abundance between each condition, the top 9 most abundant microbes and microbial genes in the study were extracted on a patient-specific level. For each sample, the relative abundance was calculated by performing total sum scaling, i.e. dividing the feature counts by the total count in the sample. The mean of the relative abundance was then taken over the entire time course of that patient. The resulting means were then ranked in descending order.

2.2.2.2 Diversity and Ordination

Diversity is a measure used in ecology to show how many bacteria, usually at the species level, are within a community. Alpha diversity is the measure of the total number of species

found within the community. Beta diversity, however, describes the difference between species composition between individual samples within the community.

To assess the alpha diversity of the processed samples, Gini-Simpson alpha diversity was calculated for each sample and compared using the Kruskal-Wallis test between the resulting diversity scores. Gini-Simpson is defined mathematically as:

$$D_{Gini}(p_1,\ldots,p_n) = 1 - \sum_{i=1}^{n} p_i^2$$

(Equation.2.2)

where $D_{Gini}(p_1,\ldots,p_n)$ denotes the diversity of the community, p_i is the relative abundance of the i-th species and n is the species within the community, $\{s_1,\ldots,s_n\}$.

To compute the beta diversity, the distance between the samples, Bray-Curtis distances are calculated between all pairs of samples and quantify the dissimilarity between them (Bray and Curtis, 1957). Bray-Curtis dissimilarity takes the form:

$$D_{BC}(x_1, x_2) = \frac{\sum_{j=1}^{p} |y_{1j} - y_{2j}|}{\sum_{j=1}^{p} (y_{1j} + y_{2j})} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

(Equation.2.3)

Where $D_{BC}(x_1, x_2)$ is the distance between two samples, p is the number of taxa, y is each species. C_{ij} is the sum of the lesser values between common species between samples, while S_i and S_j is the total number of species counted at each site (Bray and Curtis, 1957).

Principle coordinate analysis (PCoA) was used to perform ordination as an exploratory data analysis step. PCoA is similar to Principal Components Analysis (PCA) but differs as it can be applied to any form of distance matrix and it is not just limited to Euclidean distance and can better handle the sparsity of microbiome data, which would skew the covariance in PCA. In microbiome analysis, PCoA is used to visualise the distances between these samples in a lower dimensional space while preserving their distance relationships.

The distance matrices and PCoA were performed using *Scikit-bio* (http://scikit-bio.org/, version 0.5.7) functions *skbio.diversity.beta_diversity* and *skbio.stats.ordination.pcoa*.

2.2.3 Defining disease activity

In Crohn's disease, the Harvey–Bradshaw Index (HBI) is used to determine disease activity (Harvey and Bradshaw, 1980). The HBI was created in 1980 to aid the systematic collection of clinical data and consists of 5 parameters which are scored. Remission is defined by an HBI score < 5, mild activity of 5-7, moderate activity of 8-16 and severe activity >16.

Alternatively, in ulcerative colitis, the Simple Clinical Colitis Activity Index (SCCAI) was developed in 1998 in an attempt to simplify the current scoring index to help evaluate levels of exacerbation of colitis (Walmsley et al., 1998). It consists of 6 parameters whose sum is defined as the score. The exact cut-offs for activity are less well defined as in HBI. When defining remission, an SCCAI score < 2.5 is seen as being in remission or mild activity. While a score >= 2.5 is seen as active.

The last disease activity indicator is applicable to both UC and CD. Faecal Calprotectin is an inflammatory marker, which like the scoring indexes above, is a non-invasive method of assessing disease activity in IBD. For example, a patient with asymptomatic IBD with a high calprotectin level has an 80% chance of clinical relapse in the next 6 months. On the other hand, a patient with a low calprotectin level has a ~20% chance of experiencing a clinical relapse (Pavlidis et al., 2016; Smith and Gaya, 2012). The exact cutoff value for the distinction between high and low calprotectin levels in this context is debated. Most studies suggest a cutoff between remission and active disease being 250 µg/mg, though studies have suggested somewhat lower cutoff points (Pavlidis et al., 2016; Smith and Gaya, 2012). For this study, I have used cutoffs for SCCAI, HBI and Fecal Calprotectin as defined in Table 2.

Table 2.2. Breakdown of the cutoffs for disease activity. Remission was defined within UC (SCCAI) and CD (HBI), respectively. Then, the assessment of intestinal inflammation for both UC and CD via Fecal Calprotectin was performed.

Activity Index/Marker	Remission Score	Active Score
SCCAI	< 2.5	>=2.5
HBI	< 5	>=5
Faecal Calprotectin	< 250 μg/mg	> 250 μg/mg

2.2.4 Differential abundance analysis

2.2.4.1 Analysis of compositions of microbiome

Analysis of compositions of microbiome (ANCOM) method performs differential abundance from microbiome data (Mandal et al., 2015; Li, Shen and Li, 2021). This is done by calculating pairwise log ratios between all features and performing a significance test to determine if there is a significant difference in feature ratios with respect to the variable of interest (Mandal et al., 2015).

ANCOM relies on two assumptions:

1. The mean log absolute abundance of 2 taxa is not different within the ecosystem (dataset).
2. The mean log absolute abundance of all taxa in the ecosystem (dataset) does not differ by the same amount between the two study groups.

In an experiment with only two treatments, this tests the following hypothesis for feature i,

$$H_{0ri}: E\left[log\left(\frac{\mu_i^{(1)}}{\mu_{i'}^{(1)}}\right)\right] = E\left[log\left(\frac{\mu_i^{(2)}}{\mu_{i'}^{(2)}}\right)\right],$$
$$against\ H_{ari}: E\left[log\left(\frac{\mu_i^{(1)}}{\mu_{i'}^{(1)}}\right)\right] \neq E\left[log\left(\frac{\mu_i^{(2)}}{\mu_{i'}^{(2)}}\right)\right].$$

(Equation.2.4)

where $\mu_i^{(1)}$ is the mean abundance for i-th feature in the first group, $\mu_i^{(2)}$ is the mean abundance for feature i in the second group and i' is the abundance of every feature that is not i, e.g. $i \neq i'$ (Mandal et al., 2015; Li, Shen and Li, 2021).

Using the hypotheses described in Equation 2.4, the test can then be formulated using a standard ANOVA model:

$$log\left(\frac{r_{ij}^{(g)}}{r_{i'j}^{(g)}}\right) = \alpha_{ii'} + \beta_{ii'}^{(g)} + \sum_k x_{jk}\beta_{ii'k} + \epsilon_{ii'j}^{(g)} \ .$$

(Equation.2.5.)

Where r is the relative abundances of the i-th taxon and j-th samples, i' is the reference taxon $i' \neq 1, 2, ..., m$. And $g = 1, 2, ..., G.$ is the number of study groups. $\alpha_{ii'}$ is the overall common mean and $\beta_{ii'k}$ gives the effect of the g-th group. Finally, $\epsilon_{ii'j}^{(g)}$ is an i.i.d normal distribution used within standard ANOVA, $\epsilon_{ii'j}^{(g)} \sim NORMAL\left(0, \sigma_{ii`}^2\right)$, where σ is the variance.

Due to the log-ratio approach taken by this method, it cannot handle zero counts as input, as the logarithm of zero cannot be computed. In this case, zero counts are handled by the imputation of a pseudocount calculated via multiplicative replacement. In multiplicative replacement, zero counts are replaced with a small positive δ ($\delta = \frac{1}{N^2}$ where N is the number of components in the sample) whilst still preserving that the total compositions sum to 1 (Mandal et al., 2015).

Taxa were identified as differentially abundant if they had a p-value < 0.05 after the Benjamini-Hochberg multiple correction procedure. The topmost differentially abundant taxa were then plotted against each other as boxplots, with their abundances being CLR (Eq.2.1) transformed with a pseudocount imputed by multiplicative replacement as defined above. This was performed using *Scikit-bio* (http://scikit-bio.org/, version 0.5.7) *skbio.stats.composition.ancom*.

2.2.4.2 Distance-based redundancy analysis

Distance-based redundancy analysis (dbRDA) is another ordination method similar to PCoA. dbRDA is an extension of redundancy analysis, a constrained analysis method which aims to explore the feature space and determine the feature(s) which are the most separate class

labels. The main difference between the two methods is dbRDA's ability to utilise non-Euclidean dissimilarity indices, such as Bray–Curtis distance. Importantly, dbRDA makes the assumption that the dependent variables respond in a linear nature, and thus non-linear relationships cannot be found using this method.

Briefly, given the response variable, Y a multiple linear regression is run on all variables in the observation matrix X. Each variable within the set Y is regressed against all variables in the set X, leading to the computed fitted values. This can be defined as a matrix equation

$$\widehat{Y} = X[X'X]^{-1}X'Y$$

(Equation.2.6.)

where $\widehat{Y}$ is the fitted values from the multiple regressions, X is the matrix of observations, Y is the response variable and X' is the transformed matrix of X. After this, a PCoA is performed on these fitted values, resulting in the extraction of eigenvalues and eigenvectors. This process yields two distinct ordinations: 1) denoted as YU, is derived from response variables Y and 2) $\widehat{YU}$ is derived from the explanatory variables X. Additionally, a separate PCoA ordination can be conducted on the matrix of residuals, again providing eigenvalues and eigenvectors (Numerical Ecology, 2012).

To determine which species differ the most between UC, CD and non-IBD patients, dbRDA was run with Bray-Curtis dissimilarities on the relative abundances of the microbiome samples. A permutational test (PERMANOVA) is then applied to the results of dbRDA. This results in coefficients, which can then be used to determine how much each species differs between samples. The coefficients are then visualised using a bar plot, which represents the weighting of how much each species differs between samples.

Counts data for each species was loaded into a `SummarizedExperiment` (version 1.28.0) object and then passed to `mia` (version 1.6.0) to transform the data into relative abundances. `Vegan` (version 2.6.4) was used to perform dbRDA through the `dbrda()` function. Finally, the resulting ordination from dbRDA was used to perform a permutation test from the `Vegan` (version 2.6.4) package function `anova.cca()` with the number of permutations set to 9999.

2.2.4.4 Mixed effects model

Finally, the same mixed effects model as the original authors (Lloyd-Price et al., 2019) of the dataset was implemented. The model was implemented in R using nlme (Pinheiro and Bates, 2000; Pinheiro et al., 2023) and the code was extracted from the author's original code repository and run in isolation of the author's workflow. The implementation and original code for performing differential abundance analysis can be found in this repository: https://bitbucket.org/biobakery/hmp2_analysis/src/master/differential_abundance/src/core_DA_functions.r.

A linear mixed model can be represented as:

$$y_{ij} = f\left(\phi_{ij}, \nu_{ij}\right) + \epsilon_{ij},$$
$$i = 1, \ldots, M, \; j = 1, \ldots, n_i$$

(Equation.2.7.)

where M is the number of groups, n_i is the number of observations for the group, and f is function of parameter vector ϕ_{ij} and ν_{ij}. ϕ_{ij} is a linear mixed-effects model $\phi_{ij} = A_{ij}\beta + B_{ij}b_i$ where β is a vector of mixed-effects and b_i is the random effects associated with group i. The final term ϵ_{ij} is the random variable describing additive noise (Pinheiro and Bates, 2000).

For completeness, prior to fitting the model all data was arcsine square-root transformation and features with no variance or with >90% zeros were removed before fitting linear models. These steps were taken to reduce the effects of zero inflation caused by the sparsity of microbiome data. The formula and design of the mixed effects model can be seen below:

$$feature \sim (intercept) + diagnosis + antibiotic + age + (1|site) + (1|subject)$$

(Equation.2.8)

2.3 Results

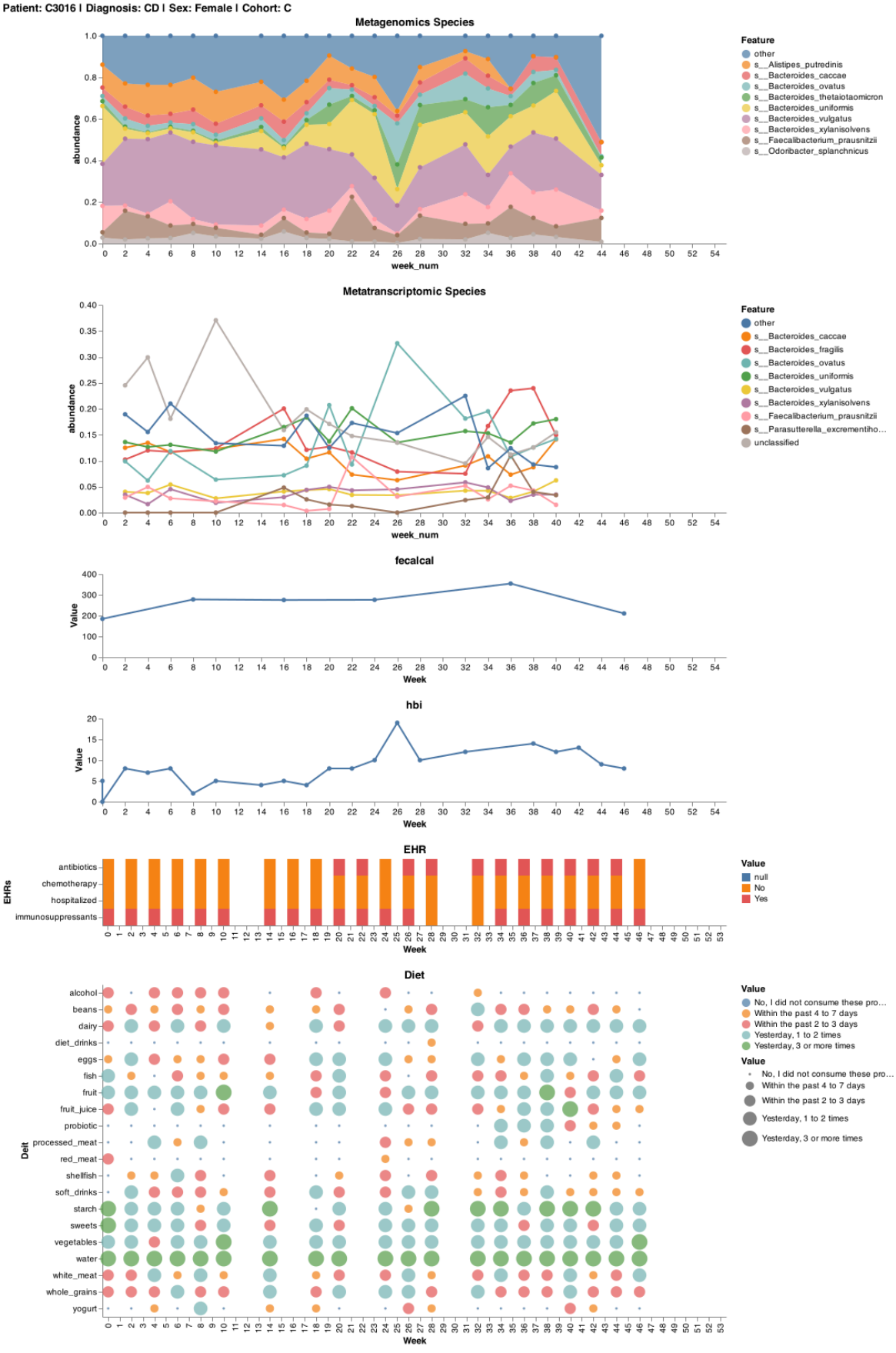
2.3.1 Exploratory Data Analysis

2.3.1.1 Temporal Component

After processing the data using the pipeline described in section 2.2.1, the data was then mapped to its metadata. In order to assess the quality of the metadata on a patient-level, a patient report was generated for every patient. Figure 2.2 shows an example of this report. The reports are aligned by the time along the axis. The report contains (a) the most abundant species, (b) the most abundant metagenes expressed, (c) the faecal calprotectin score, (d) the disease activity score (either SCCAI for UC or HBI for CD), (e) electronic health care records as a boolean value, and (f) results of the diet survey for this patient over time. All patient reports can be found in the supplementary materials.

The reports provide a visual tool to determine how consistent the sampling was during the course of the study conducted by *Lloyd-Price et al.* Prior to any further analysis, visually, the time component of the data across all patients is not stable or irregularly sampled. This results in difficulty in applying any sort of time-series analysis to the data. Time-series data requires regularly sampled data and often has a trend, seasonality and other time-dependent structures. As this dataset had too many missing and irregularly spaced time points, time-series analysis was not suitable for this data.

Figure 2.2 (Next page) An example of the summarised patient reports after preprocessing and metadata extraction. The report is fixed on the x-axis with respect to time, and the y-axis presents the extracted data. From top to bottom, this report shows (a) the most abundant species, (b) the most abundant metagenes expressed, (c) the faecal calprotectin score, (d) the disease activity score (either SCCAI for UC or HBI for CD), (e) electronic health care records as a boolean value, and (f) results of the diet survey for this patient over time. This report was produced for each patient individually.


Patient: C3016 | Diagnosis: CD | Sex: Female | Cohort: C
Metagenomics Species
abundance
week_num
Feature
other
s__Alistipes_putredinis
s__Bacteroides_caccae
s__Bacteroides_ovatus
s__Bacteroides_thetaiotaomicron
s__Bacteroides_uniformis
s__Bacteroides_vulgatus
s__Bacteroides_xylanisolvens
s__Faecalibacterium_prausnitzii
s__Odoribacter_splanchnicus
Metatranscriptomic Species
abundance
week_num
Feature
other
s__Bacteroides_caccae
s__Bacteroides_fragilis
s__Bacteroides_ovatus
s__Bacteroides_uniformis
s__Bacteroides_vulgatus
s__Bacteroides_xylanisolvens
s__Faecalibacterium_prausnitzii
s__Parasutterella_excrementiho…
unclassified
fecalcal
Value
Week
hbi
Value
Week
EHR
EHRs
antibiotics
chemotherapy
hospitalized
immunosuppressants
Week
Value
null
No
Yes
Diet
Diet
alcohol
beans
dairy
diet_drinks
eggs
fish
fruit
fruit_juice
probiotic
processed_meat
red_meat
shellfish
soft_drinks
starch
sweets
vegetables
water
white_meat
whole_grains
yogurt
Week
Value
No, I did not consume these pro…
Within the past 4 to 7 days
Within the past 2 to 3 days
Yesterday, 1 to 2 times
Yesterday, 3 or more times
Value
No, I did not consume these pro…
Within the past 4 to 7 days
Within the past 2 to 3 days
Yesterday, 1 to 2 times
Yesterday, 3 or more times

2.3.1.2 Clinical metadata

The unstructured metadata was cleaned manually to ensure overlapping fields were concatenated together. The manual curation focused on metadata that would be informative about the disease activity of the patient. These fields included therapies (oral corticosteroids, chemotherapy, antibiotics and immunomodulators), general information (sex, age, cohort and location) and electronic healthcare records information (diagnosis and hospitalisation). These fields were chosen in particular as they were the most regularly sampled metadata during the study. This was then mapped to the quality control metrics extracted from the output files of each sample (pipeline shown in Methods section 2.2.1).

A correlation matrix was calculated for each sample's metadata for all IBD samples that passed quality control to assess the correlation between the metadata features. Spearman's correlation between the clinical metadata showed a high correlation coefficient between the clinical data and the quality control metrics. More specifically, the most correlated features were between the disease activity metrics (SCCAI, HBI and faecal calprotectin) and the number of human or bacterial reads extracted from the sample (Figure 2.3).

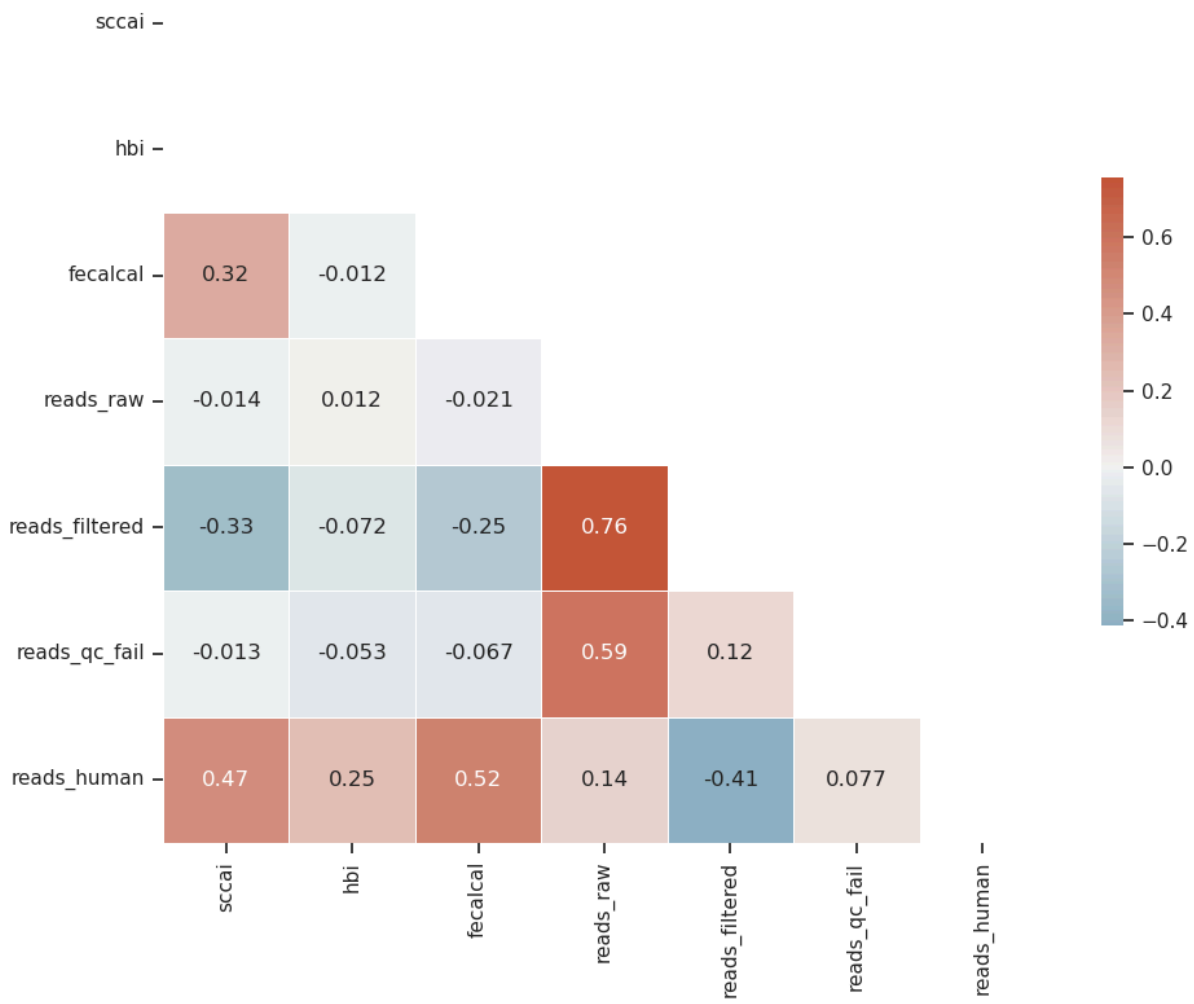


Figure 2.3. Correlation analysis between metadata across all IBD (both UC and CD) samples. Pearson correlation between the disease activity metrics and quality control metrics. HBI and SCCAI have no values as they are disease-specific. Only samples with full rank (i.e. no missing values) were used to ensure a fair comparison between variables.

To further investigate the relationship between disease activity metrics and read counts, linear regression and Pearson correlation coefficient were performed to observe the trend between data points. A histogram and univariate KDE curves were also plotted to assess the distribution of the data. Figure 2.4 shows the correlation between the number of human reads detected in the sample and the disease activity for both UC and CD. This demonstrates a positive correlation between higher levels of disease activity. Interestingly, a higher correlation between UC and SCCAI (r=0.467) and the number of human reads is observed when compared to CD and HBI (r=0.248). Faecal calprotectin is

similarly positively correlated for both UC (r=0.589) and CD (r=0.447). The data is skewed in its distribution, however, with the majority of the human reads nearing zero. It should also be noted that the sampling rate of faecal calprotectin was lower than that of disease activity, hence the difference in the number of points between the plots. The opposite case was true when applying the same analysis to the number of bacterial reads extracted from the sample, which showed a more negative correlation (Figure 2.5). This phenomenon makes sense both quantitatively and biologically as during points of increased disease activity, the amount of blood in the stool would increase, particularly for UC patients, as the disease tends to be situated closer to the anus. The disease activity metrics are derived from patient-driven surveys where one of the questions is related to whether the patient experienced blood in their stool.

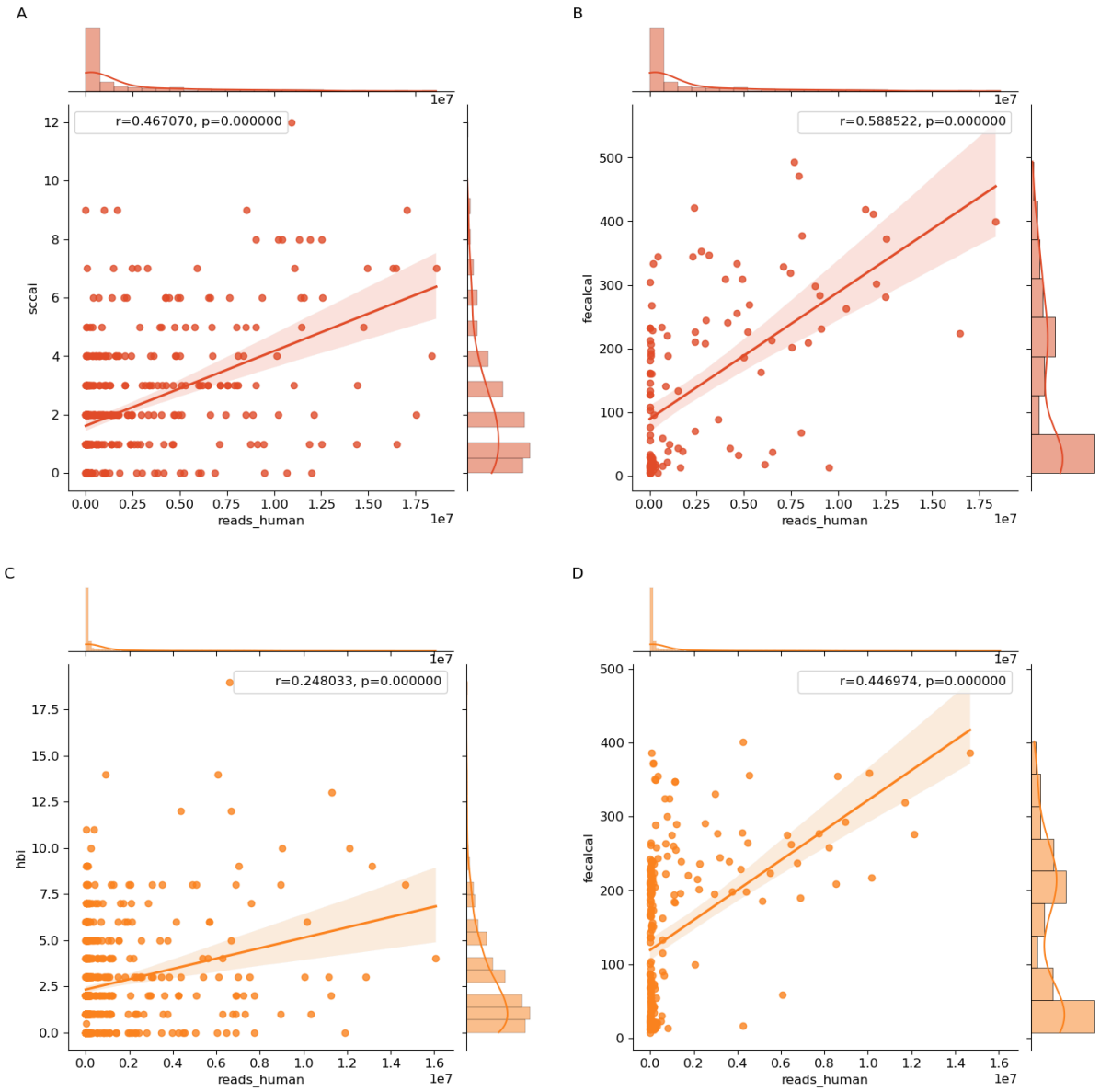


Figure 2.4. Regression analysis between the number of human reads extracted from the faecal samples and paired disease activity metric from IBD patients. (A) Disease activity in UC (SCCAI) against the number of reads mapped to the human genome in the sample. (B) UC samples Fecal Calprotectin scores against the number of reads mapped to the human genome. (C) Disease activity in CD (HBI) against the number of reads mapped to the human genome. (D) CD samples Fecal Calprotectin scores against the number of reads mapped to the human genome.

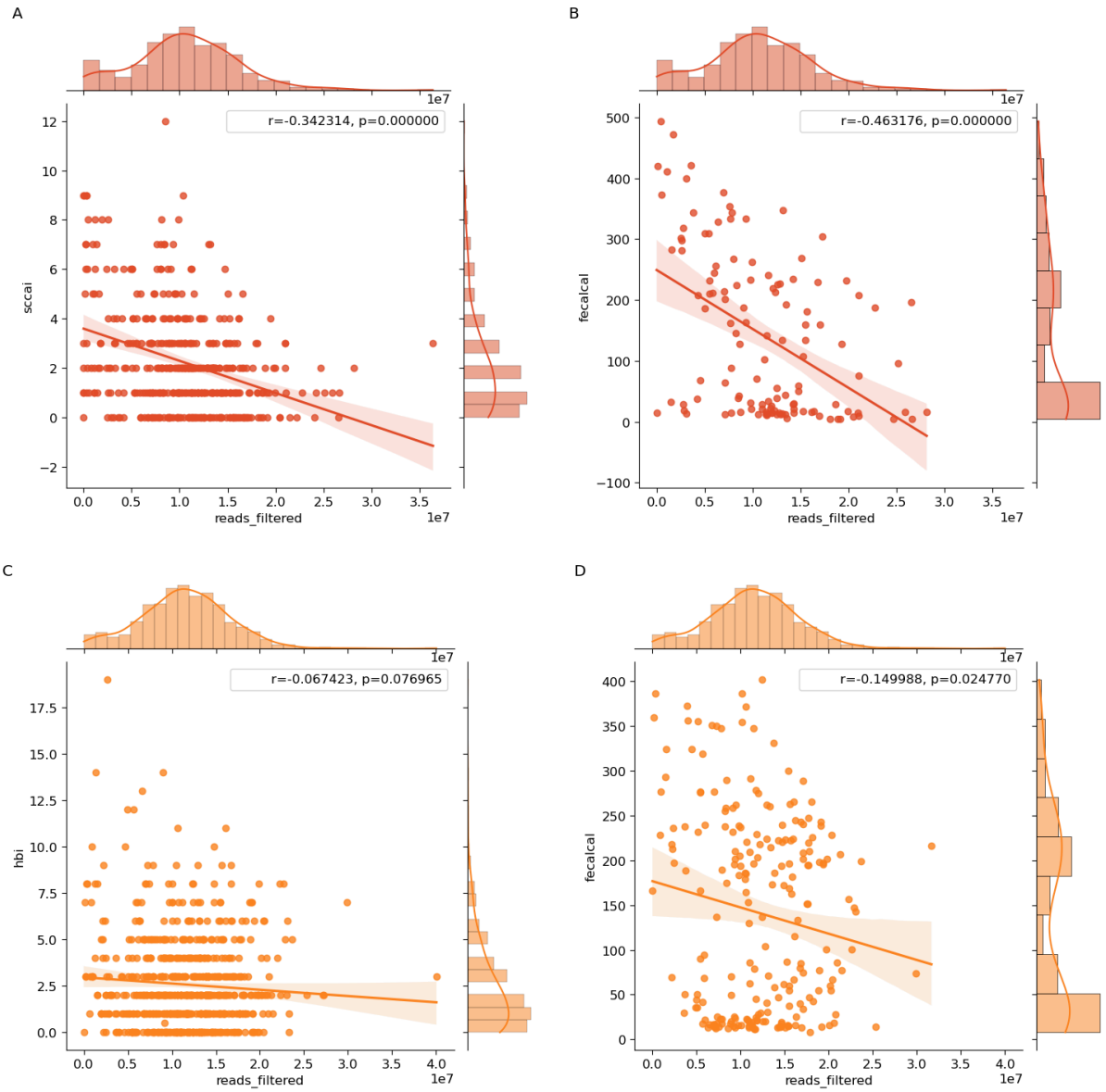


Figure 2.5. Regression analysis between the number of bacterial genes extracted from the faecal samples against disease activity metric from IBD patients. (A) Disease activity in UC (SCCAI) against the number of reads mapped to a bacterial genome in the sample. (B) UC samples Fecal Calprotectin scores against the number of reads mapped to a bacterial genome. (C) Disease activity in CD (HBI) against the number of reads mapped to a bacterial. (D) CD samples Fecal Calprotectin scores against the number of reads mapped to a bacterial genome.

2.3.2. Ordination

All samples that passed quality control were visualised using PCoA ordination of beta diversity calculated by Bray Curtis dissimilarity. There was no clear separation between the diagnosis, locations or sex of the individual. Thus, this demonstrates that the data cannot be separated by ordination alone.

Most of the variation captured from the PCoA was either from *Bacteroidetes* or *Firmicutes* phylum. Alpha diversity between the groups showed a decrease in diversity in both UC and CD compared to non-IBD patients (Kruskal-Wallis test p=7.888e-07 Stat=2.439e+01; p=8.305e-15 Stat=6.026e+01 respectively) (Figure 2.6). This further confirms the microbiome's increased instability in IBD patients compared to non-IBD patients. There was no statistically significant difference between UC and CD patients (p=2.651e-01, stat=1.242e+00).

The effects of repeated measures, in this instance the longitudinal nature of the study, can clearly be seen on the PCoA plot in Figure 2.7. The variation between subjects is greater than the variation between time points, resulting in the localisation of samples. The location of the patient cohort and the sex of the patient seems to show a reasonable mixing; however, a better mixing could be achieved through batch effect correction.

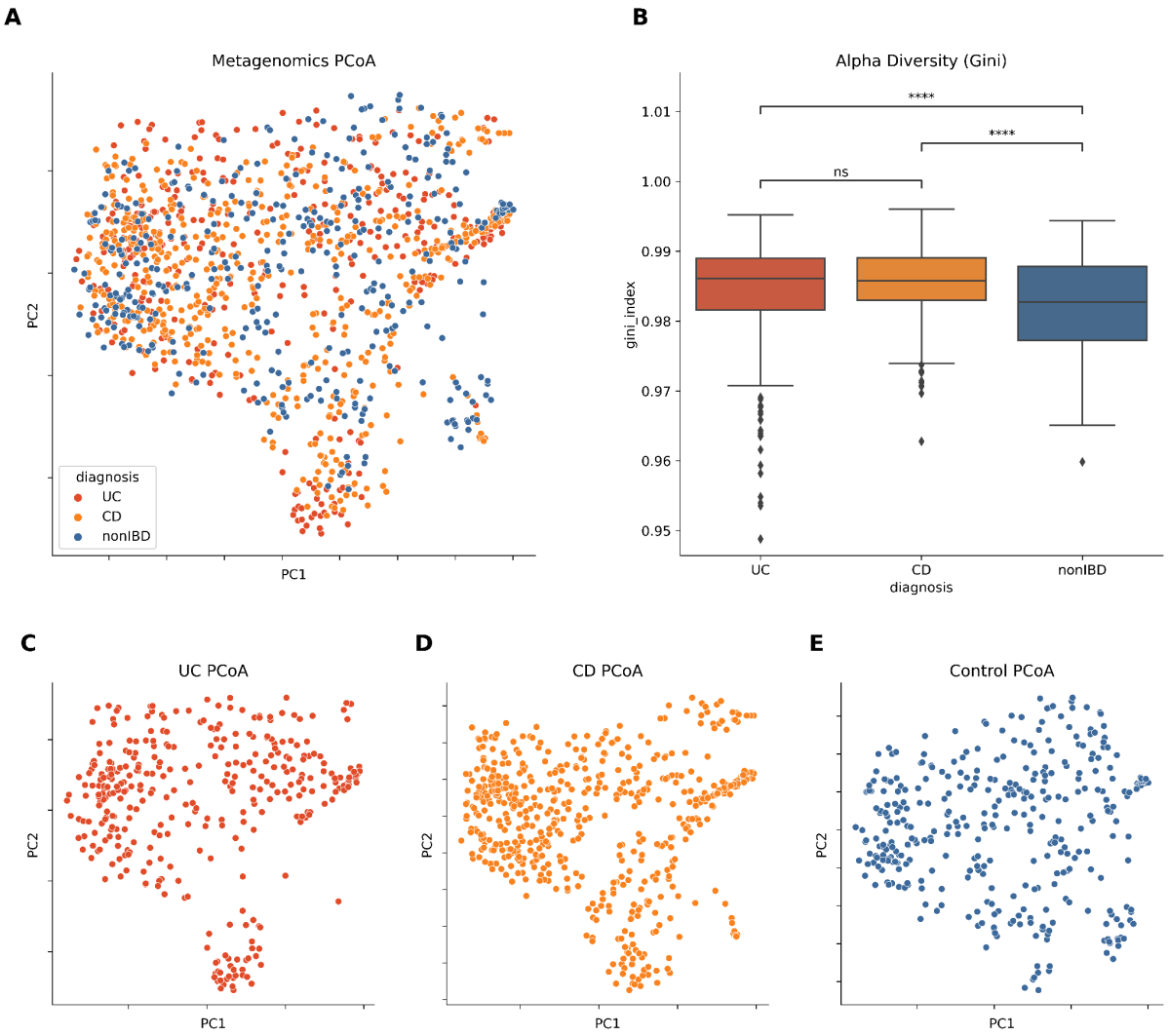


Figure 2.6. Assessment of alpha and beta diversity from all metagenomics samples. (A) PCoA ordination of beta diversity calculated by Bray Curtis dissimilarity. (B) Comparison of the alpha diversity, Gini-Simpson, between UC, CD and Controls. The Kruskal-Wallis test was performed between samples. Alpha diversity was not significant between UC vs CD (p=2.651e-01, $stat$=1.242e+00), but was significant between UC vs controls (p=7.888e-07 $Stat$=2.439e+01), and CD vs Controls (p=8.305e-15 $Stat$=6.026e+01). (C, D, E) Individual ordinations of UC, CD and Controls. (*ns*: p <= 1.00e+00 ****: p <= 1.00e-04)

Clinical metadata was overlaid as well to determine at a high level if the samples clustered by their disease activity. SCCAI, HBI and faecal calprotectin levels for each patient. Note the reduction in the number of samples due to the lower sampling rate of the data and the removal of data points not relating to disease activity markers (i.e. non-IBD patients cannot have SCCAI or HBI scores). There was no obvious clustering of the data by the disease

activity of the patients; however, a slight gradient of activity can be seen moving from the bottom to the top of the PCoA plots (Figure 2.7 D,E,F).

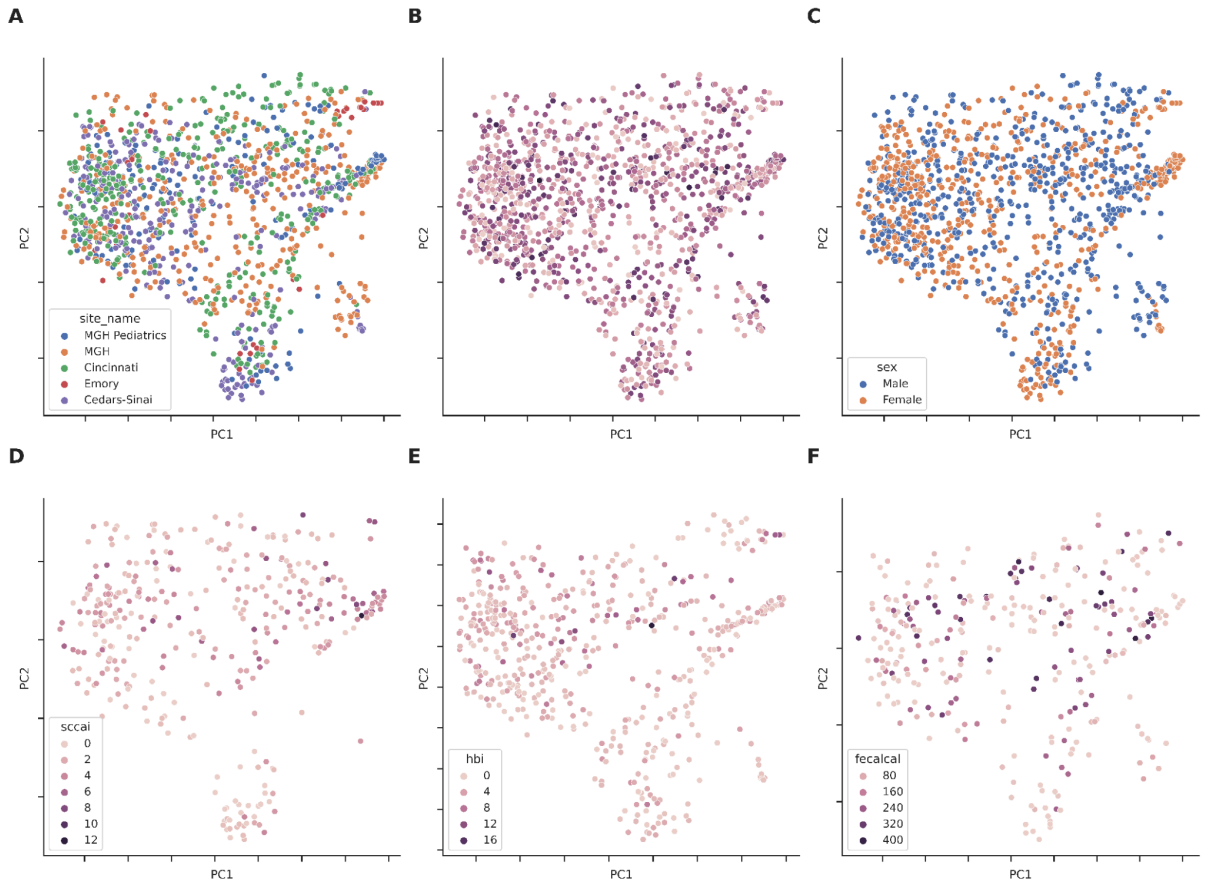


Figure 2.7. Ordination overlaid with metadata. To investigate potential batch effects and the effect of repeated measures the metadata was overlaid onto the PCoA projection. (A) Shows the cohort sites. (B) To investigate the effect of time/repeated measures. (C) Ordination of the sex of the patient. (D) SCCAI as a gradient overlayed onto the ordination on UC samples. (E) HBI as a gradient overlayed onto the ordination on CD samples. (F) Faecal calprotectin measurements as a gradient.

To extend the findings of the ordination analysis the composition and frequency of species in each diagnosis group were summarised. Figure 2.8 shows the average composition of each stacked bar plot of the composition at phylum, genus and species-level. This again further confirmed that Bacteroidetes or Firmicutes phylum were the largest taxa represented in the patients' microbiome. On average *Bacteroides vulgatus* and *Bacteroides uniformis* were the most abundant species, with a larger abundance level in both UC and CD

when compared to the non-IBD cohort. However, across the entire cohort, inclusive of IBD and healthy controls, the most abundant microbial species were *Subdoligranulum unclassified*, *Faecalibacterium prausnitzii*, *Ruminococcus torques*, *Bacteroides vulgatus*, and *Bacteroides uniformis* (Supplementary Fig 2.1). Conversely, *Faecalibacterium prausnitzii* and *Prevotella copri* are more abundant in non-IBD patients. The histogram of the shared species within the diagnosis group shows that the majority of the species are shared across all samples seen from the skew to the left in the histogram.

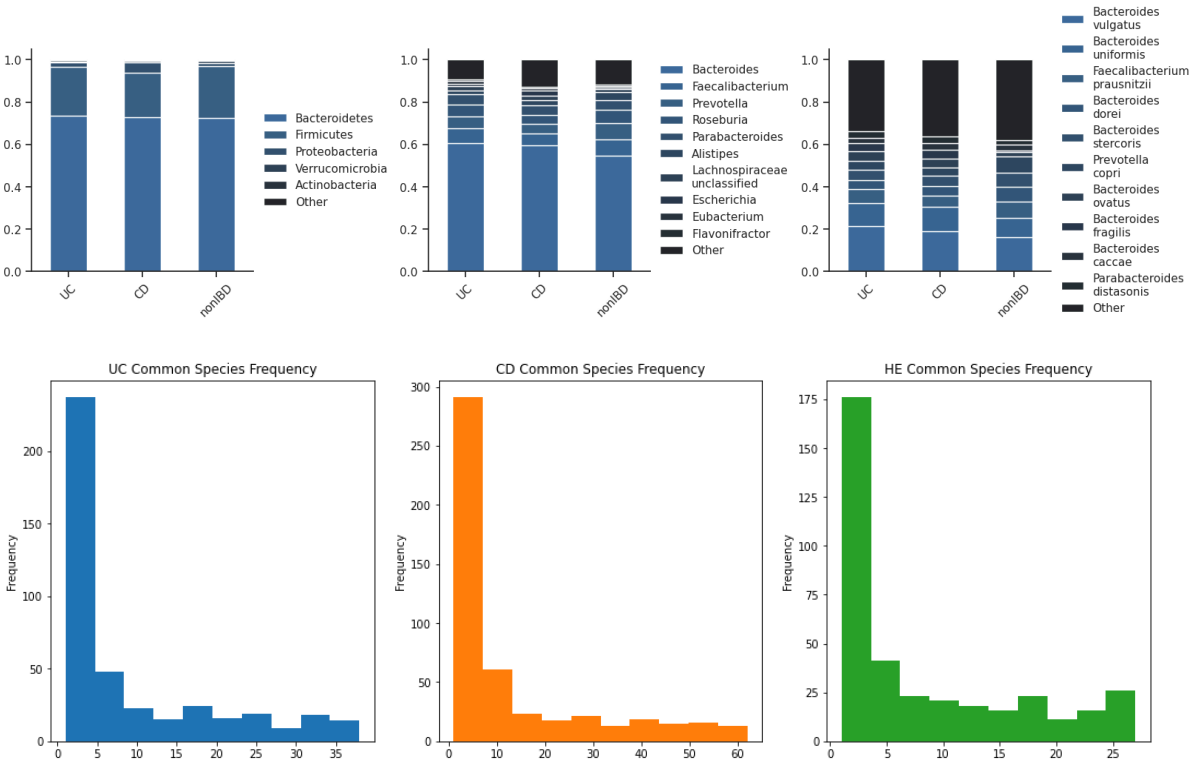


Figure 2.8. Compositional analysis between each diagnosis. The mean relative abundance across conditions at Phylum, Genus and Species levels respectively. The topmost abundance taxa were selected and the remainder were grouped into the 'Other' category. Histograms representing the frequency of shared species within the diagnosis. The most frequently observed species (leftmost bin) and rarest (rightmost bin) in each species between diagnoses.

2.2.3 Differential abundance analysis between non-IBD and IBD patients

To get a baseline of what existing methods identified as biomarkers, three commonly used methodologies for biomarker identification were implemented. Each method's top markers were extracted and then collected together to assess the overlap between each method. Depending on the algorithm used, the resulting data is presented in a different way. For ANCOM the data is presented as the centre-log ratio of the abundance after the top 10 features are extracted from the ranking for plotting. For dbRDA, the top 20 absolute highest coefficients are extracted and plotted on a bar plot with the vector value representing the weighting of the coefficient.

For ANCOM UC vs non-IBD found a total of 18 differentially abundant taxa which rejected the null hypothesis (Figure 2.9.). Of which *Alistipes* genus is found to be higher in the non-IBD cohort with both *Alistipes putredinis*, *Alistipes shahii* and *Alistipes finegoldii* being found in the top ten (Figure 2.9.). Odoribacter splanchnicus Only one *Bacteroides* species was found in the top 10 differential abundant species (Figure 2.9.).

CD vs non-IBD obtained the greatest number of differentially abundant taxa between the three analyses. In total 73 species were able to reject the null hypothesis. Again *Alistipes putredinis* was found to be the most differentially abundant species. In CD, however, *Bacteroides fragilis*, *Clostridium bolteae*, *Flavonifractor plautii*, and *Ruminococcus gnavus* were all significantly more abundant compared to non-IBD controls. Finally, when looking between the two IBD conditions, ANCOM identified 13 species as differentially abundant. *Bacteroides* and *Roseburia* genus made up the majority of the 13 species, with *Odoribacter splanchnicus* again being the second most abundant species (Figure 2.9.).

In the non-IBD controls, *Prevotella copri* was noted as the species to observe the most shifts over the study but only dbRDA identified it (Figure 2.10). dbRDA also showed *Faecalibacterium prausnitzii* and *Roseburia* genera being more influential for non-IBD when compared to CD (Figure 2.11.). Interestingly, when comparing UC and CD dbRDA suggested *Alistipes putredinis* as more relevant to CD suggesting it has both a protective and harmful effect. When compared with AMCON and the mixed effects model, dbRDA ranked *Escherichia coli* much higher up in both CD and UC.

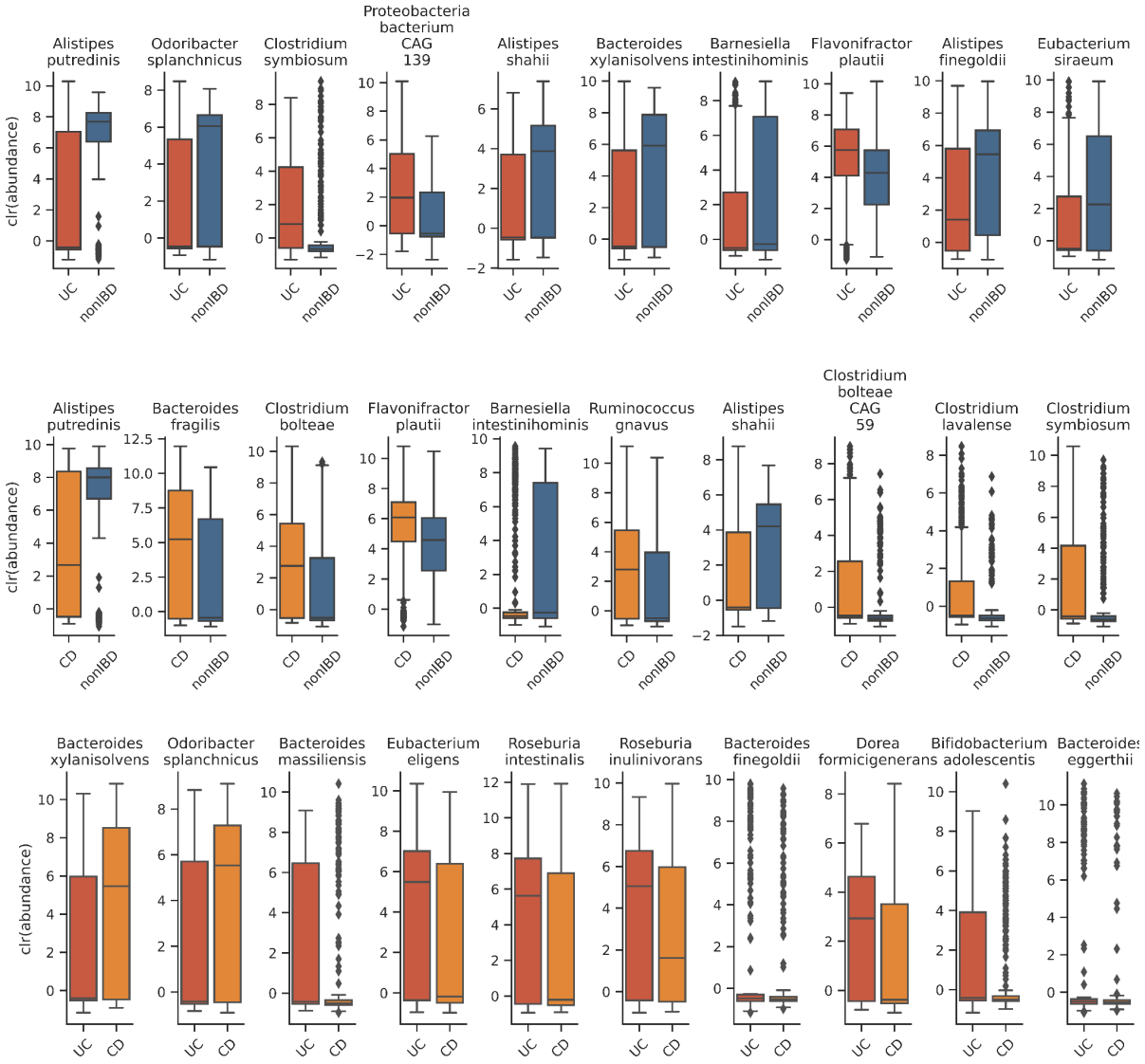
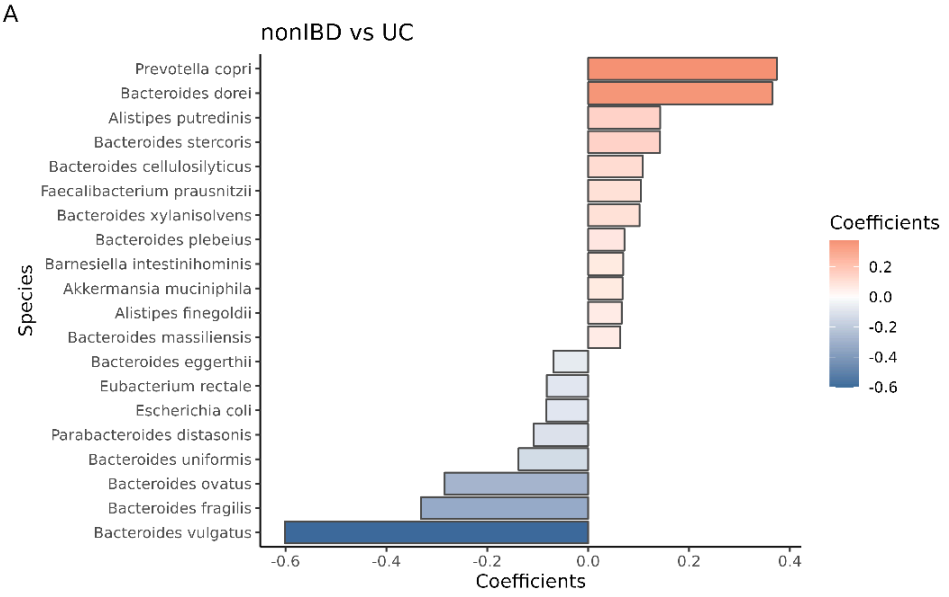
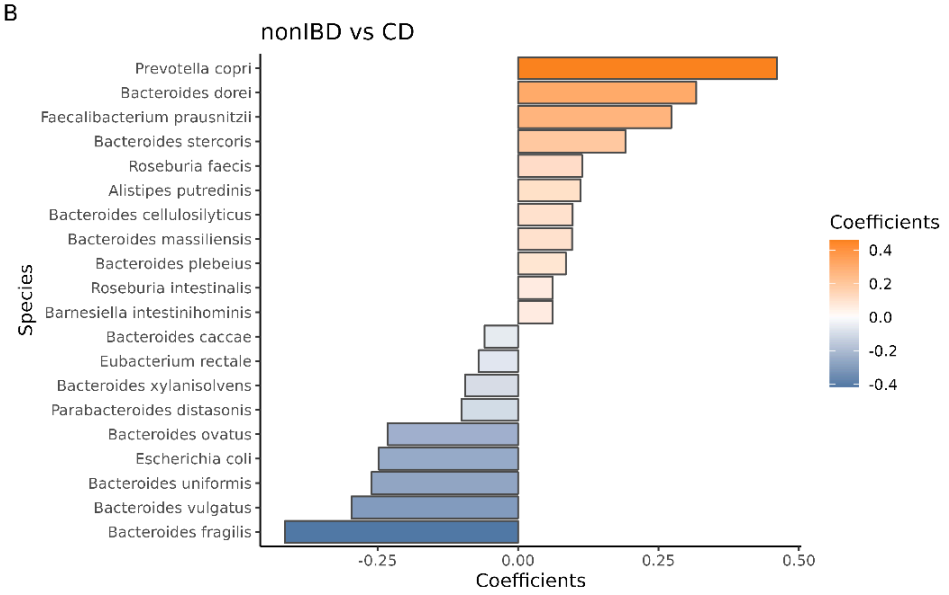


Figure 2.9. Comparison of differential abundance analysis using ANCOM between conditions. (A) Top 10 (of 18) most differentially abundant taxa between UC and Control group. (B) Top 10 (of 73) most differentially abundant taxa between CD and Control group. (C) Top 10 (of 13) most differentially abundant taxa between UC and CD.

A


nonIBD vs UC
Prevotella copri
Bacteroides dorei
Alistipes putredinis
Bacteroides stercoris
Bacteroides cellulosilyticus
Faecalibacterium prausnitzii
Bacteroides xylanisolvens
Bacteroides plebeius
Barnesiella intestinihominis
Akkermansia muciniphila
Alistipes finegoldii
Bacteroides massiliensis
Bacteroides eggerthii
Eubacterium rectale
Escherichia coli
Parabacteroides distasonis
Bacteroides uniformis
Bacteroides ovatus
Bacteroides fragilis
Bacteroides vulgatus
Species
-0.6
-0.4
-0.2
0.0
0.2
0.4
Coefficients
Coefficients
0.2
0.0
-0.2
-0.4
-0.6

B


nonIBD vs CD
Prevotella copri
Bacteroides dorei
Faecalibacterium prausnitzii
Bacteroides stercoris
Roseburia faecis
Alistipes putredinis
Bacteroides cellulosilyticus
Bacteroides massiliensis
Bacteroides plebeius
Roseburia intestinalis
Barnesiella intestinihominis
Bacteroides caccae
Eubacterium rectale
Bacteroides xylanisolvens
Parabacteroides distasonis
Bacteroides ovatus
Escherichia coli
Bacteroides uniformis
Bacteroides vulgatus
Bacteroides fragilis
Species
-0.25
0.00
0.25
0.50
Coefficients
Coefficients
0.4
0.2
0.0
-0.2
-0.4

C

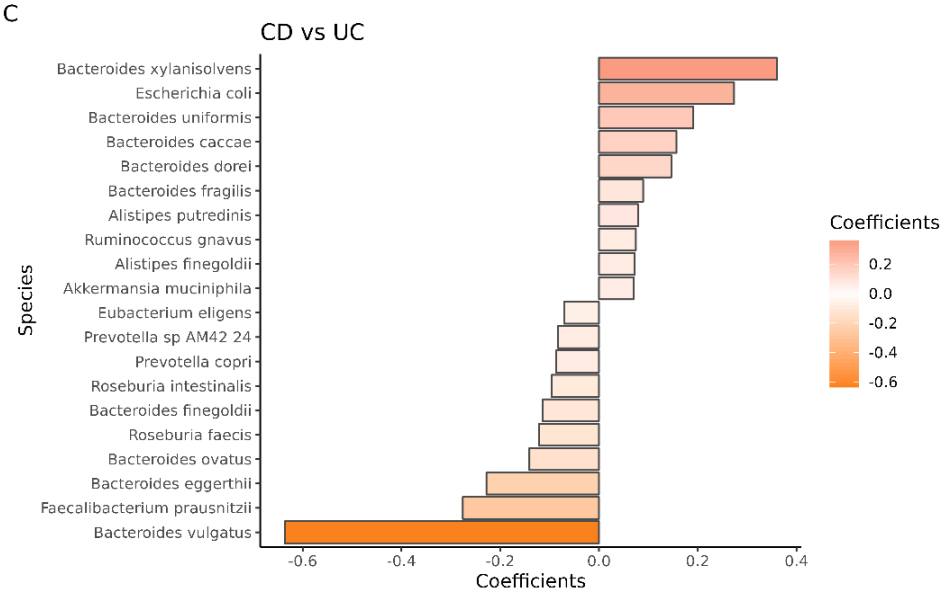

CD vs UC
Bacteroides xylanisolvens
Escherichia coli
Bacteroides uniformis
Bacteroides caccae
Bacteroides dorei
Bacteroides fragilis
Alistipes putredinis
Ruminococcus gnavus
Alistipes finegoldii
Akkermansia muciniphila
Eubacterium eligens
Prevotella sp AM42 24
Prevotella copri
Roseburia intestinalis
Bacteroides finegoldii
Roseburia faecis
Bacteroides ovatus
Bacteroides eggerthii
Faecalibacterium prausnitzii
Bacteroides vulgatus
Species
-0.6
-0.4
-0.2
0.0
0.2
0.4
Coefficients
Coefficients
0.2
0.0
-0.2
-0.4
-0.6

Figure 2.10. (Previous page) Comparison of species abundance between conditions using dbRDA. (A) The top most differing abundance of species between non-IBD and UC patients. (B) The top most differing abundance of species between non-IBD and CD patients. (C) The top most differing abundance of species between CD and UC patients.

Finally, after multiple testing, the mixed effect model showed no differential abundant taxa. There were no values with a q-value < 0.286. This demonstrates the limitations of the other models and the importance of accounting for the environmental and longitudinal data effects of the data. After comparing the non-IBD to CD and non-IBD to UC, the top most differential abundant but not statically significant taxa are shown in Figure 2.11.

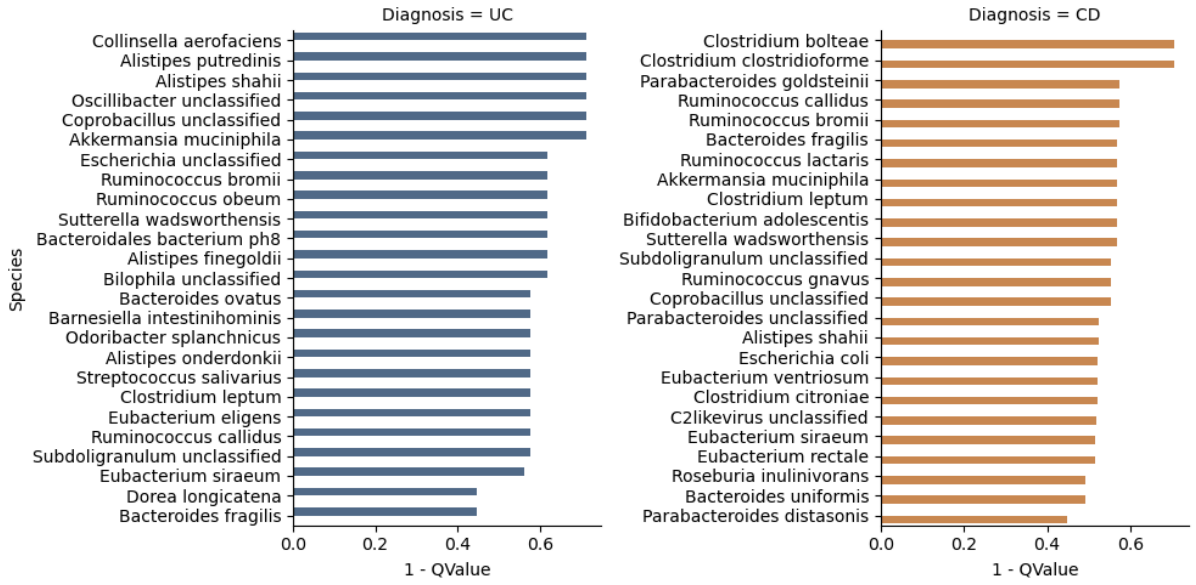


Figure 2.11. Comparison of species abundance between conditions using mixed effects models. (A) The top differing abundance of species between non-IBD and UC patients. (B) The top differing abundance of species between non-IBD and CD patients.

2.4 Discussion

The human gut microbiome plays a vital role in the pathogenesis and prognosis of IBD patients. Despite this knowledge and other studies investigating its role, no global IBD microbiome signature has been defined. The lack of a clear IBD microbiome signature is consistent across all patients and, more broadly, across multiple patient cohorts and locations. Studies have shown how each individual presents with a relatively unique microbial fingerprint adding to the complexity of obtaining biomarkers or prognostic indicators from IBD samples (Ley et al., 2006; Clemente et al., 2012; Lloyd-Price et al., 2019). This demonstrates the complexity of microbiome data and the effect the environment and the host can have on the microbial communities in the gut. That being said, some bacterial communities are linked or correlated with IBD microbiome.

The temporal component of the microbiome is an added complexity, particularly for computational modelling. It has been observed that the microbiome is present with both autoregressive and non-autoregressive factors (Gibbons et al., 2017; Integrative HMP (iHMP) Research Network Consortium, 2014; Liu, 2023). The combination of autoregressive and non-autoregressive time series makes extracting prognostic indicators even more challenging. When the data you are presented with displays such a large amount of complexity it suggests a more complex model might be required. However, in this case, and many other clinical studies the most powerful techniques struggle to leverage the inconsistencies and overall lack of data. For example, two powerful machine learning models for time series analysis are Rocket and Long short-term memory (LSTM) neural networks. Rocket is a state-of-the-art linear algorithm that leverages the power of random convolutional kernels to achieve fast and accurate time series classification (Dempster, Petitjean and Webb, 2020). Even though Rocket is designed for small datasets with irregular time points, missing time points and short time series mean it is unsuitable for this data set. The other end of the spectrum is the non-linear deep architecture LSTM neural network (Hochreiter and Schmidhuber, 1997). These models, however, require a hundreds of times orders of magnitude larger dataset than is currently available.

Another clear challenge with longitudinal studies is the collection of patient metadata. Often these meta data are missing from studies, either because they require additional steps to obtain the data or because of ethical and privacy reasons. Even with a large-scale study

such as that conducted by Llyod-Price et al without this information, it is difficult to extract biological meaning. For example, when comparing the meta data of the IBD patients a clear association between the UC disease activity metric of SCCAI and faecal calprotectin was observed (r=0.32). Interestingly, this was not the case for HBI (r=−0.012) even though it had been observed previously. Moreover, the clear correlation between the disease activity metric and the number of human reads found in a stool sample does not mean a reduction in microbiome diversity but instead with an increase in disease activity a patient is more likely to experience complications resulting in blood or tissue being passed with the faecal matter.

A key area of work research in microbiology is ordination. Multiple studies have applied numerous ordination methods to try and cluster patients into groups. In my analysis, using PCoA there were no naturally forming clusters. This suggests the importance of non-linear dimensionality reduction methods for the projection and clustering of microbiome data. Two methods currently being used extensively in single-cell RNA-seq analysis are t-distributed stochastic neighbour embedding (t-SNE) (van der Maaten and Hinton, 2008) and Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018). These have both been hypothesised as good alternatives to linear models such as PCoA (Armstrong et al., 2021, 2022). However, they are both less suited to compositional data and are difficult to interpret without additional models (i.e. differential expression/abundance analysis between clusters).

After assessing the ordination and the stability of the microbiome, the next step is to determine what is causing those differences. There are a large number of totals to perform differential abundance or biomarker identification. By testing four commonly used but methodologically very different tools, the aim was to identify areas which could be improved on. The models identified bacterial species that were implicated in IBD already. Other species have been implicated in IBD more recently like, such as *Odoribacter splanchnicus* (*Lima et al.*, 2022) and *Clostridium symbiosum* (*He et al.*, 2017; *Hassouneh*, *Loftus and Yooseph*, 2021), *Faecalibacterium prausnitzii* and *Bacteroides vulgatus* (Mills et al., 2022). *Bacteroides* species were dominant across all methods but particularly with dbRDA. Methods such as dbRDA were also able to identify microbial shifts in *Prevotella copri* occurring in the healthy condition which the other methods missed which again has been implicated by other studies (Bajer et al., 2017; Lloyd-Price et al., 2019). All methods pointed towards the role of

Alistipes genera in the healthy condition which has been described as having protective and harmful properties across multiple inflammatory diseases (Moschen et al., 2016; Zuo et al., 2019; Bangsgaard Bendtsen et al., 2012; Parker et al., 2020). Finally, both the mixed effects model and ANCOM strongly suggested the role of *Clostridium bolteae* between non-IBD and CD patients. More recently, studies have shown the metacommunities in Crohn's disease in which *Clostridium bolteae* cluster with other harmful bacterial species like *Escherichia coli*, *Klebsiella pneumoniae* and *Streptococcus salivarius* (*He et al.*, 2017)

This chapter has identified multiple areas of research to address. From a technical point of view, the longitudinal nature of the data needs to be accounted for and is an area where models such as the mixed effects model are most performant. However, these models work best when you have the meta data to build into the model's covariates and interaction terms. Most studies lack this fine metadata detailing and a constraint method like dbRDA could help discover the difference between the phenotypes. AMCON is a popular tool for differential abundance analysis, which shows in its citation numbers, but also has been found to be sensitive while simultaneously controlling the FDR effectively. However, benchmarking studies have shown that AMCON struggles when there are less than 20 samples in each group, meaning larger study sizes are most appropriate for AMCON (Weiss et al., 2017).

In the next chapter, I will address some of these issues by developing a model to fill this gap and the lack of longitudinal microbiome tools. This model should be able to be used either to detect shifts in the microbiome composition between phenotypes, as a feature selection step or in an unsupervised manner to explore the dynamics of the microbiome composition.

Chapter 3: Exploring the temporal dynamics of the microbiome in inflammatory bowel disease

3.1 Introduction

The traditional approaches to microbiome analysis begin to fall apart with the introduction of a temporal component. The precise nature of the temporal dynamics of the microbiome remains unclear (Li, Shen and Li, 2021; Glassner, Abraham and Quigley, 2020). This is particularly the case for the microbiome composition during dysbiosis, which results in an unbalanced or abnormal microbiome, for example, during times of an increase in disease activity.

A typical time series can be defined as a series of data points obtained at successive time points with equal intervals between them. The aim of a time series analysis is to measure the overall change in the data points over time. Most publicly available data from microbiome studies are not time series data and are instead longitudinal data. Though similar to time series data, longitudinal data tends to have fewer time points and is normally taken at different intervals. The advantage of longitudinal data, over particularly cross-sectional data, is that the increased sampling allows for distinction between an actual signal and noise within an individual. Thus, longitudinal studies are more precise and informative as they help account for any sampling or technical errors which are difficult to detect in cross-sectional analysis. More specifically, in the case of longitudinal microbiome analysis, there still remains a key area of research (Kodikara, Ellul and Lê Cao, 2022):

1. Differential abundance over time (e.g. the difference between external/clinical factors)
2. Clustering of microorganisms evolving concomitantly across time
3. Identification of temporal relationships between microorganisms

Although whole genome sequencing (WGS) from a faecal sample has the advantage of being non-invasive, it also has limitations with respect to assessing the microbiome composition in patients (Hildebrand, 2021). Multiple studies have demonstrated that microbial communities in the gut are spatially organised (Sheth et al., 2019; Duncan, Carey-Ewend and Vaishnava, 2021; Mark Welch et al., 2017). This means that a species of bacteria is more likely to be located next to the same species than it is a different species. It is thought that this disruption in the spatial organisation of the gut microbiome is a contributing factor to disease pathogenesis. Furthermore, this means that detecting the composition from a faecal sample can be challenging in that you are only sub-sampling the population and therefore, you may need to account for unobserved species of bacteria.

To address these issues outlined above many models have been developed to take into account covariates such as age, location, sex and even correlation between species which are known to be co-expressed (Martin, Witten and Willis, 2020). The limitations of these approaches are that they rely on having a wealth of high-quality metadata and that the assumptions you are making about the biological prior are correct. These models are mostly discriminative, meaning that they are trying to separate or predict the changes based on the observed data to determine decision boundaries which best separate the data.

An alternative approach is generative models. A generative model differs from a discriminative model as it attempts to model how data is placed in space rather than draw decision boundaries between the data in this space. Instead, it models $P(X \mid Y = y)$, the conditional probability of observing X given the target variable Y. Then by sampling from this distribution of the input and output data, it creates new synthetic data in the input space. Loosely, a generative model can be placed in three classes; autoregressive models, generative adversarial models, and latent variable models (Jebara, 2004).

3.1.1 Aims

In this chapter, in collaboration with my industrial partner, BenevolentAI, I explore the dynamics of the microbiome by performing exploratory data analysis on the longitudinal metagenomic dataset. I then develop a Bayesian generative model to infer the dispersion on

a patient-level, species-level and then a pooled model which accounts for the patient's individual microbiome over time.

This chapter's aims were as follows:

- Develop a Bayesian model to infer the dispersion of an individual's microbiome between UC, CD and healthy controls
- Identify similarities or differences in microbial dispersion within patients with increased disease activity compared with those with lower disease activity
- Identify microbiome difference between patients with flare compared to patients who remain in remission over time

3.2 Methods

3.2.1 Bayesian Models

In this section I explore intra- and inter-patient microbiome variability over time. To achieve this two models are developed. Precision model (PM) which focuses on inter-patient variability and Species Precision Model (SPM) which focuses on modelling the individual taxa from each patient over time. The model is designed for analysis of microbiome count data and is modelled from a beta-binomial distribution. The model includes patient-specific baselines and scaling factors, allowing for flexibility in how different taxa are represented across different patients.

3.2.1.1 Model definition for inferring species dispersion per patient

The intra-patient variability was explored by exploring the dispersion of their microbiome composition over all their sampling points. To account for the steady-state or patient-specific baseline the parameter μ was generated for each patient. μ represents the underlying baseline probability for each patient and is calculated by the mean composition of the patient over all time points. The model then infers the parameter s (dispersion) for all the species in each sample within each patient (e.g. one value for s for each sample). This was modelled over a beta-binomial distribution.

Let $X \in \mathbb{R}^{NxD}$ be a count matrix where $N \in \mathbb{N}$ is the number of samples for $n = 1,..., N$, and $D \in \mathbb{N}$ is the number of taxa for $d = 1,..., D$. μ is a vector that represents a patient baseline as the mean proportion across samples such that $\mu \in \mathbb{R}^{D}$.

To calculate the patient-specific baseline a mask was created for each of the patient's repeated measures. Let $X_k \in \mathbb{R}^{NxD}$ be a matrix which contains observations relating to patient k. For each patient k, k=1,...,K we have a set of matrices

$$\{X_k\}_{k=1}^{K}, \qquad \text{(Equation.3.1)}$$

where n_k are observations relating to patient k,

$$N = \sum_{k=1}^{k} n_k. \qquad \text{(Equation.3.2)}$$

Define $\mu(X_k) = [\mu((x_k)_1), \mu((x_k)_2), ..., \mu((x_k)_D)]$ to be a vector of column means. The i-th element of $\mu(x_k)$ as

$$\mu((x_k)_i) = \sum_{j=1}^{n_k} x_{ij} \qquad \text{(Equation.3.3)}$$

For all X_k, we have a set of column means for each patient $k = 1,..., K$,

$$\{\mu(X_k)\}_{k=1}^{K} \qquad \text{(Equation.3.4)}$$

t to be an integer vector of total counts, $t \in \mathbb{N}^D$, where

$$t_n = \sum_{d=1}^{D} X_{n,d}, \; for\, n = \{1,..., N\}, \qquad \text{(Equation.3.5)}$$

$s \in \mathbb{R}^N$ is a vector of length N which is a sample-specific scaling factor. The transformation factors $\alpha \in \mathbb{R}^{NxD}$ and $\beta \in \mathbb{R}^{NxD}$ for each sample $n = 1,..., N$ and taxa $d = 1,..., D$,

$$\alpha_{n,d} = s_n \cdot \mu_d, \qquad \text{(Equation.3.6)}$$

$$\beta_{n,d} = s_n \cdot (1 - \mu_d). \qquad \text{(Equation.3.7)}$$

For each sample $n = 1,..., N$, the scaling factor s_n follows a normal distribution defining the prior as,

$$s_n \sim NORMAL(0, 10000). \qquad \text{(Equation.3.8)}$$

For each sample $n = 1,..., N$, the count data $X_{n,}$ follows a beta-binomial distribution,

$$X_{n,} \sim BETABINOMIAL(t_{n,}, \alpha_{n,}, \beta_{n,}). \qquad \text{(Equation.3.9)}$$

The simulation was done using *Stan* (https://mc-stan.org) and Markov chain Monte Carlo (MCMC). *Stan* is a Turing-complete probabilistic programming language used for performing statistical inference of Bayesian models. In particular, STAN solves MCMC using

a variant of the Hamiltonian Monte Carlo (HMC) algorithm called the No-U-Trun sampler (NUTS) (Hoffman and Gelman, 2014). The model was run for 2000 iterations with a warmup (burnin) of 1000. A model was built per patient; therefore, each patient had their own microbiome modelled. Once s_n had been inferred for each patient, the results were inspected for each patient to determine if patients with an increased disease activity had a higher level of dispersion in their microbiome signatures.

3.2.1.2 Model definition for inferring species dispersion

To extend the model defined in 3.2.1.1, a pooled model was then created. This time as well as accounting for the patient's steady-state and patient dispersion, we also inferred the species dispersion. Again, this was modelled from a beta-binomial distribution, but this time the parameter s (dispersion) infers a dispersion for each species rather than each sample given, while still accounting for the patient's baseline. The baseline for each patient was calculated as in Equations 3.1-3.4.

Let $X \in \mathbb{R}^{NxD}$ be a count matrix where $N \in \mathbb{N}$ is the number of samples for $n = 1,..., N$, and $D \in \mathbb{N}$ is the number of taxa for $d = 1,..., D$. μ is a vector that represents a patient baseline as the mean proportion across samples such that $\mu \in \mathbb{R}^D$. t to be an integer vector of total counts, $t \in \mathbb{N}^D$, where

$$t_n = \sum_{d=1}^{D} X_{n,d}, \; for\, n = 1,..., N, \quad \text{(Equation.3.10)}$$

$K \in \mathbb{N}$ is the total number of patients . m is a sample map which is a vector of length N, such that it indicates which sample belongs to which patient for $k = 1,..., K$. Define two matrices $\alpha \in \mathbb{R}^{NxD}$ and $\beta \in \mathbb{R}^{NxD}$ for each sample $n = 1,..., N$ and taxa $d = 1,..., D$ where

$$\alpha_{n,d} = s_d \cdot \mu_{m_{n,d}}, \quad \text{(Equation.3.11)}$$

$$\beta_{n,d} = s_d \cdot \left(1 - \mu_{m_{n,d}}\right). \quad \text{(Equation.3.12)}$$

$s \in \mathbb{R}^D$ is a vector of length D which is a taxa-specific scaling factor. For each taxon d, s_d is assumed to follow a normal distribution defining the prior as,

$$s_d \sim NORMAL(0,\ 1000). \qquad \text{(Equation.3.13)}$$

And for each sample n and taxon d, $if\ \mu_{m_{n,d}} > 0$, then $X_{n,d}$ is assumed to follow a beta-binomial distribution:

$$X_{n,d} \sim BETABINOMIAL\left(t_n,\ \alpha_{n,d},\ \beta_{n,d}\right) \quad \text{(Equation.3.14)}$$

Like with the previous model, the simulation was done using *Stan* (https://mc-stan.org), which performed full Bayesian statistical inference with MCMC sampling, solved using the HMC algorithm. The model was run for 2000 iterations with a warmup (burnin) of 1000.

Differing from the first model, we now extend the model to take account for each taxa's dispersion within that patient over the repeated measures. The parameter vector s in the model plays a crucial role. It is a vector of length D (the number of taxa), with each element s_d representing a non-negative scale factor for each taxon d. The purpose of s is to modulate the influence of the patient-specific baseline μ on the observed counts X. For each taxon d, the element sd scales the baseline proportion, $\mu_{m_{n,d}}$ for the corresponding patient mapped by m_n. This scaling results in the parameters $\alpha_{n,d}$ and $\beta_{n,d}$ which are then used as parameters for the beta-binomial distribution of the observed count $X_{n,d}$. The interpretation of s_d can be seen as a measure of the dispersion or variability of each taxon across the samples, relative to the baseline μ. A higher value of s_d indicates greater variability or influence of taxon d in the counts observed across different patients. The prior distribution for each sd, a normal distribution with a mean of 0 and a large variance, allows for a wide range of values, reflecting the potential for significant differences in the taxa's prevalence and variability across the samples. This prior also implies a regularisation effect, preventing overfitting by penalising large values of sd unless supported by the data.

3.2.3 Benchmarking

To evaluate how well the model performed at identifying and extracting the most variable species of bacteria in each disease or within the disease, the models above were compared against a more standard approach. A typical approach would be to perform differential abundance analysis on the data between the two subsets. In differential abundance, the raw count data is normalised, and a statistical test is used to discover quantitative changes in abundance levels between groups (Li, Shen and Li, 2021). In so, the idea is to uncover the directionality of features within the data to identify up-regulated or down-regulated features between conditions.

3.3 Results

To evaluate the models described in 3.2.1 and probe the dispersion of the microbiome over time in IBD patients, I used the largest publicly available longitudinal metagenomic study available, created by *Lloyd-Price et al.* This patient cohort consists of 132 individuals who were recruited as part of the Human Microbiome Project (HMP) (Human Microbiome Project Consortium, 2012; Lloyd-Price et al., 2019). The patients were from four US hospitals and comprised three paediatric and two adult cohorts. In total, the authors collected 1,785 stool samples along with various metadata, including disease activity metrics, diet, therapy, disease age and more. For more information about the data and preprocessing steps, see Chapter 2.

3.3.1 Inferring dispersion between active and inactive disease states

To explore how the microbiome varies at a patient-specific level in UC and CD, the first step was to infer the dispersion of the complete patient microbiome with respect to that patient's baseline microbiome composition. As the data is longitudinal and therefore doesn't have the same properties of time series, by defining a patient-specific baseline, the model can infer the changes over time within a patient. The model was run for n=30 UC and n=50 CD patients who passed the quality control. Only samples had a minimum total read count > $1e10^6$, a minimum feature count > 100, and a minimum prevalence of 10% of all samples. Each model was fit with a prior normal distribution of 1000.0.

A trace plot displays the sampled values of a parameter (or parameters) over each iteration of the MCMC simulation. After an initial "burn-in" period, where the chain might show non-representative behaviour, the plot should ideally display a "fuzzy caterpillar" pattern. This indicates that the chain is exploring the parameter space effectively without getting stuck in any particular region. If the trace plot shows clear, systematic patterns or drifts, it's a sign that the chain might not have converged. Figures 3.1 and Figure 3.2 show the mixing of the patient *s* dispersion parameter for all species within that patient's microbiome for CD and UC, respectively. Chains are considered healthy when they are well-mixed and stationery. An unhealthy chain can be an indication of a poorly specified model. The mixings shown in the trace plot suggest that the chains mixed well and look satisfactory for both CD and UC models.

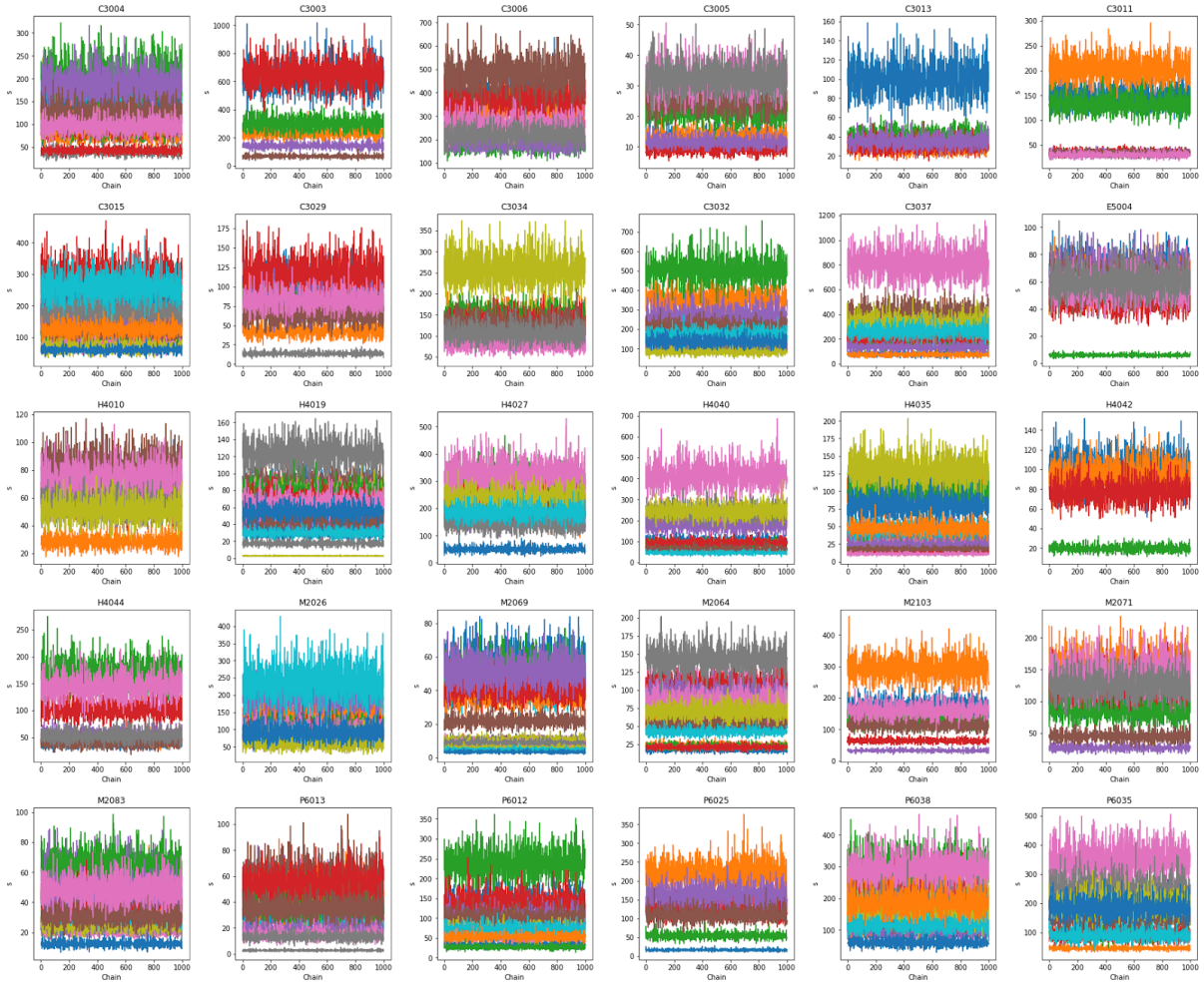


Figure 3.1. Mixing of the patient-dispersion model for each species shows the convergence of parameter S for each UC patient. Each plot represents an individual patient's trace for their inferred microbiome dispersion. Each colour presents a single chain which is a sample from that patient over the 52 weeks of the study.

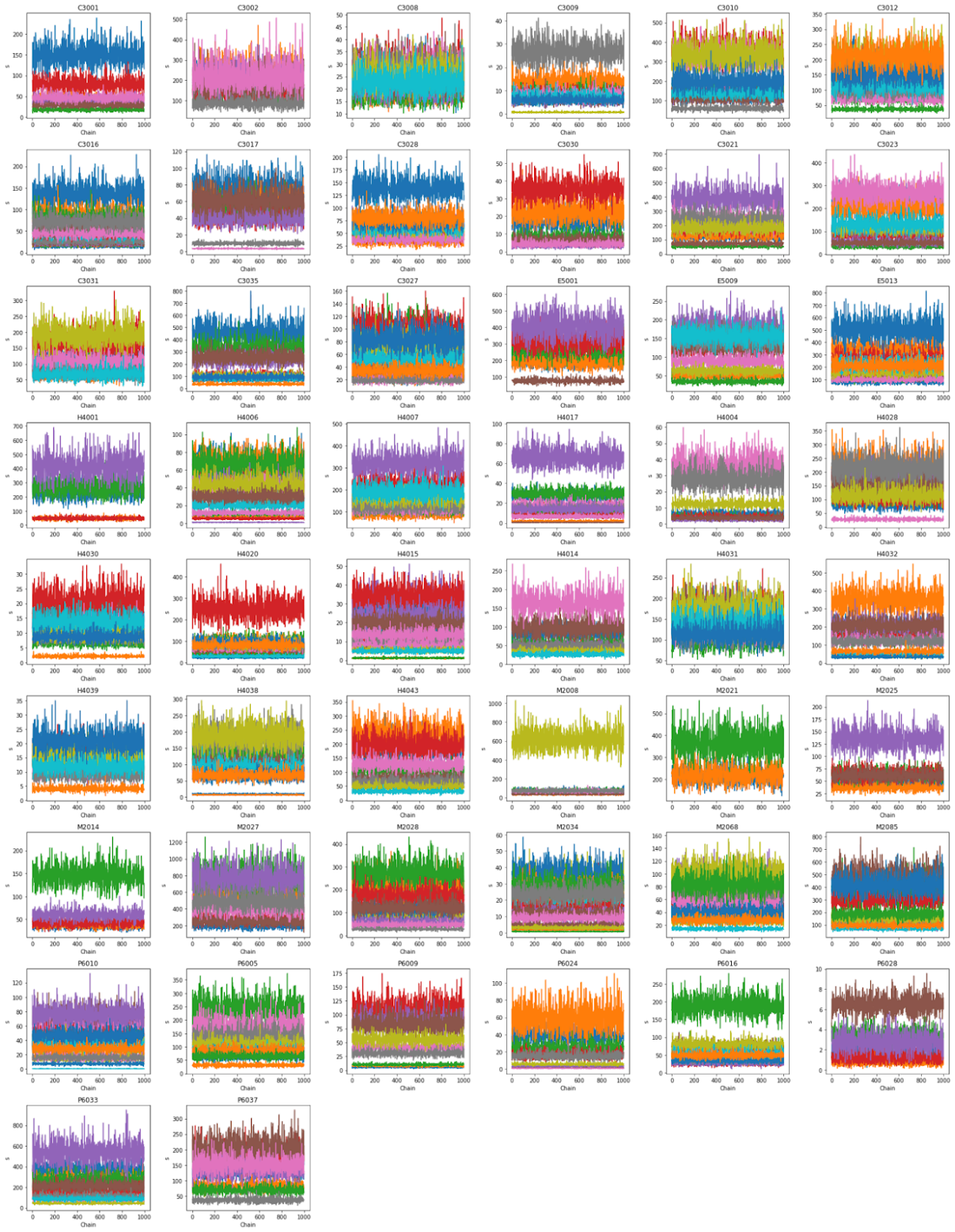


Figure 3.2. Mixing of the patient-dispersion model for each species shows the convergence of parameter S for each CD patient. Each plot represents an individual patient's trace for their inferred microbiome dispersion. Each colour presents a single chain which is a sample from that patient over the 52 weeks of the study.

To interpret the model results. The fitted model, input data, mapping file and metadata were exported and held in compressed files. The patient baseline μ, and the dispersion parameter *S* was also extracted from the model. As each model was built on an individual patient these results were then mapped back to the original metadata using the mapping file.

When accounting for the patient's baseline composition, the model shows a decrease in stability in the microbiome composition of UC patients who have a high level of disease activity as defined by an SCCAI score ≤ 2.5 (Figure 3.3. A and 3.3. B). Though a downward trend was seen, it should be noted there were some outliers and crossovers between the two distributions (Figure 3.3. A and 3.3. B). To check that the total read count and metadata were not biassing the model's results, disease activity metrics (Figure 3.3. C,D,E) and cohort location (Figure 3.3 F), which were labelled on a plot of the dispersion parameter *s* plotted against the total counts captured for each sample (Figure 6F). This shows that a good mixture between the total reads and the cohort of each patient suggests the baseline regression approach has not induced further biases in the model.

The same investigation was conducted on CD patients, and again a decrease in compositional stability was observed (Figure 3.4. A and 3.4. B). Interestingly, the difference between patients with inactive or active disease was not as striking as seen in UC patients. Again the model's output dispersion parameter *s* is plotted against the total counts captured for each sample annotated with the disease activity metrics and sample cohort location. This again did not suggest any clear bias towards higher or lower read counts.

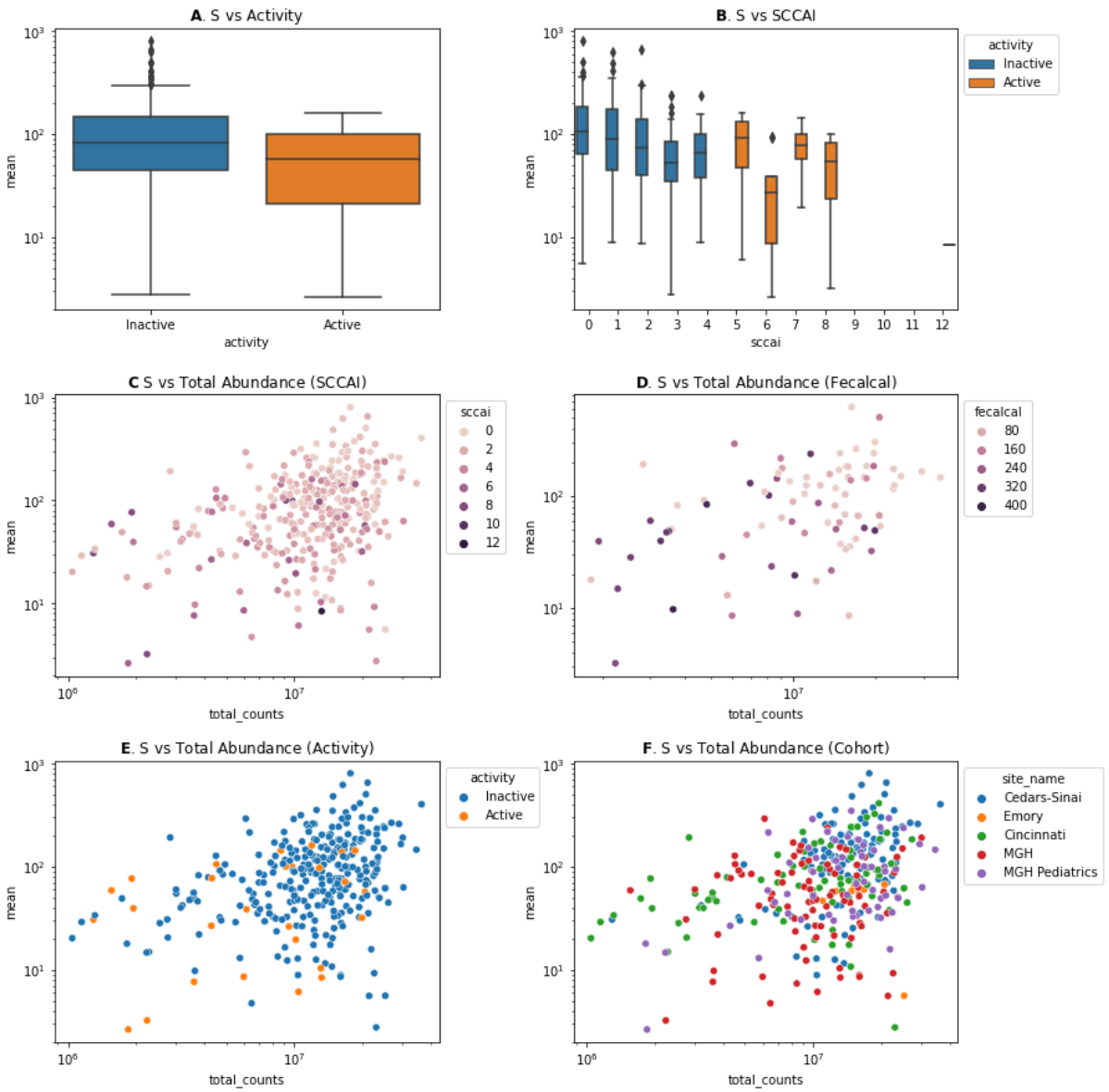


Figure 3.3. Patient dispersion compared to disease activity in UC patients. Aggregated models for each patient sample microbiome dispersion accounting for the patient's baseline composition. (A) The log dispersion (s) against the thresholded SCCAI score, inactive < 2.5 and active >= 2.5. (B) The log dispersion (s) against the SCCAI for each sample's score. (C, D, E, F) The log dispersion (s) against the sum of the total abundance coloured by the SCCAI, faecal calprotectin, disease activity and cohort location, respectively.

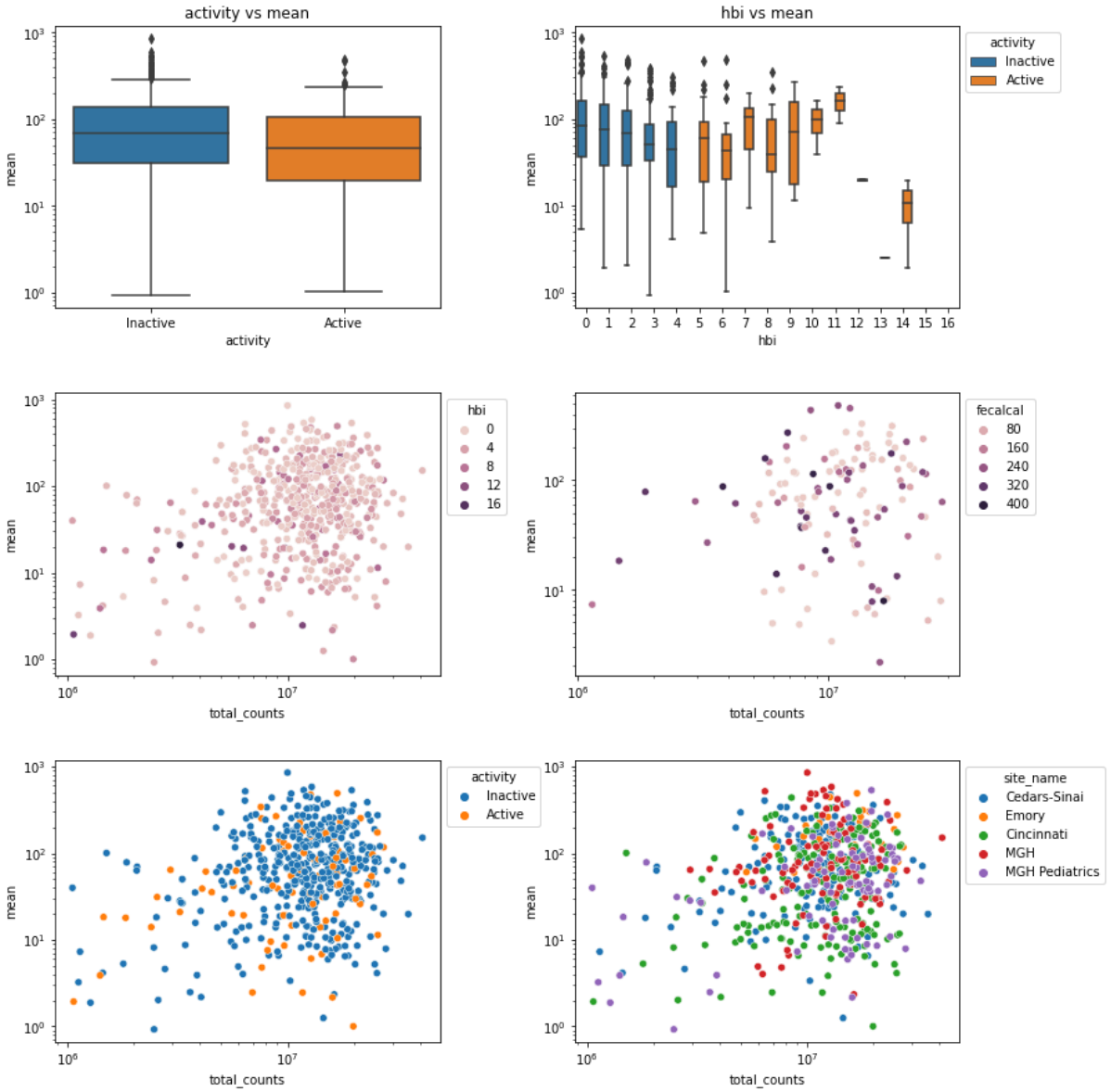


Figure 3.4. Patient dispersion compared to disease activity in CD patients. Aggregated models for each patient sample microbiome dispersion accounting for the patient's baseline composition. (A) The log dispersion (s) against the thresholded HBI score, inactive < 5 and active >= 5. (B) The log dispersion (s) against the HBI for each sample's score. (C, D, E, F) The log dispersion (s) against the sum of the total abundance coloured by the HBI, faecal calprotectin, disease activity and cohort location, respectively.

3.3.2 Inferring species dispersion between UC, CD and healthy controls

Differing from the patient dispersion model above, where the entire patient microbiome was accounted for, the model was then extended to account for features (species) making up the microbiome composition. Furthermore, rather than modelling the patient samples individually, this model was pooled across all the patients before the data was passed to the model. This model was termed the Species Precision Model (SPM) as it was able to capture both the dispersion between groups and also give feature-level information. The same basic QC steps were taken. Samples had to meet the following criteria; (1) total reads > $1x10^6$ reads, (2) had at least 3-time points (i.e. n samples > 3), (3) features with absolute zero abundance were removed, and (4) features (species) present in at least 10% of the cohort.

To compare the species-level dispersion across conditions, the metagenomic data for UC, CD and healthy controls were all pooled together. A model for each condition was built accounting for the patient's baseline to combat the inter- and intra-patient biases. Across all conditions, *Bacteroides* and *Roseburia* genus were very dynamic. Interestingly, when looking into the individual abundance for each species within these genera, groups of patients who saw the shifts in one species would not see the shift in other species. This shows the ecosystem within the microbiome and demonstrates the model can determine patient-specific changes using the baseline regression approach.

3.3.2.1 Species dispersion in ulcerative colitis patients

In UC, the *Bacteroides* genus was highly unstable across all patients. *Bacteroides* have been shown to be very transcriptionally active at the mucosal surfaces, pointing to a functional potential in a dysbiotic microbiome (Rehman et al., 2010). The model highlighted *Bacteroides fragilis* as having the largest dispersion across all patients. *Bacteroides fragilis* is a common bacteria found in the human colon and has been reported to play a role in disease development and is one of the most common causes of anaerobic infections in humans. More specifically, in UC certain strains of *Bacteroides fragilis* are enterotoxigenic and, therefore, produce toxins which result in vomiting, diarrhoea and an increase in inflammation, contributing to disease development and progression in both mouse dextran sodium sulphate (DSS) models and UC patients (Rabizadeh et al., 2007; Zamani et al., 2017).

3.3.2.2 Species dispersion in Crohn's disease

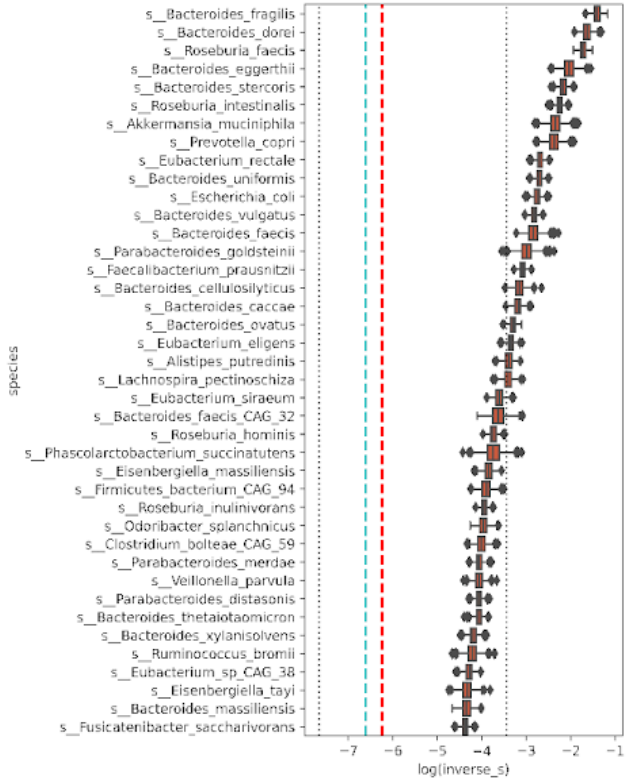
The species with the most dispersion in CD was *Klebsiella pneumonia*, which has been connected to CD disease development (Rashid, Ebringer and Wilson, 2013; Rashid and Ebringer, 2011; Garrett et al., 2010). Another species with a large amount of dispersion was *Escherichia coli*, which is an invasive pathogen. *Escherichia coli*, under the correct conditions, can colonise the intestinal mucosa by adhering to intestinal epithelial cells (Palmela et al., 2018; Roda et al., 2020; Barnich and Darfeuille-Michaud, 2007). In particular, this is the case for a pathotype called adherent-invasive *Escherichia coli (Palmela et al., 2018)*. Moreover, specifically in CD patients, neutrophil cells' antimicrobial defence system is defective and therefore their inflammatory responses to kill *Escherichia coli* are reduced (Segal, 2018). (Figure 3.5 and Figure 3.6)

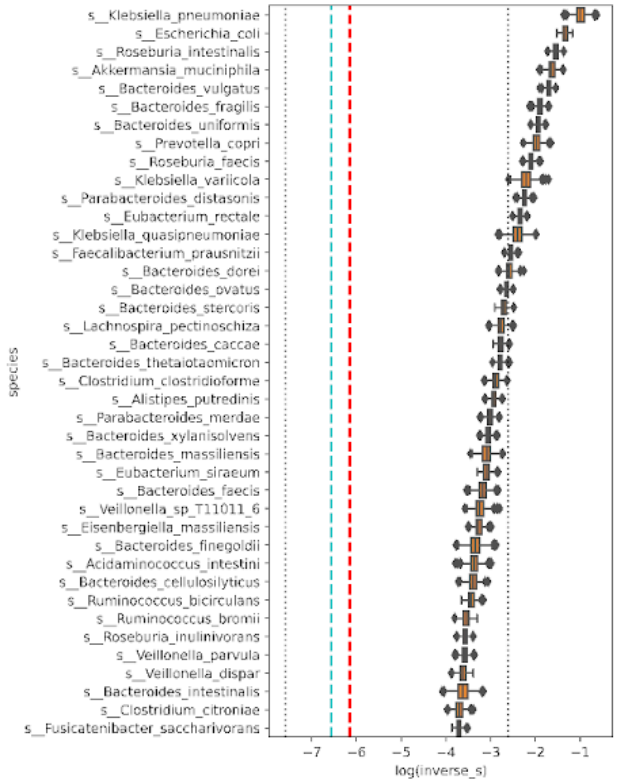
3.3.2.3 Species dispersion in healthy controls

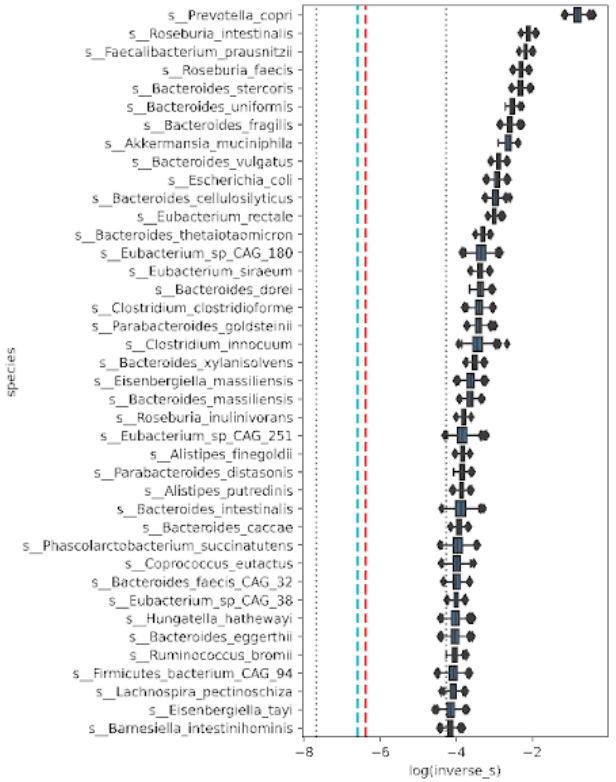
Finally, as a control group, we also assessed the species-level dispersion of healthy controls over time. In this group, *Prevotella copri* was exceptionally dynamic in the healthy controls. This was also reported by the authors, who used a different approach to identify shifts (Lloyd-Price et al., 2019) (Figure 3.5 and Figure 3.6).

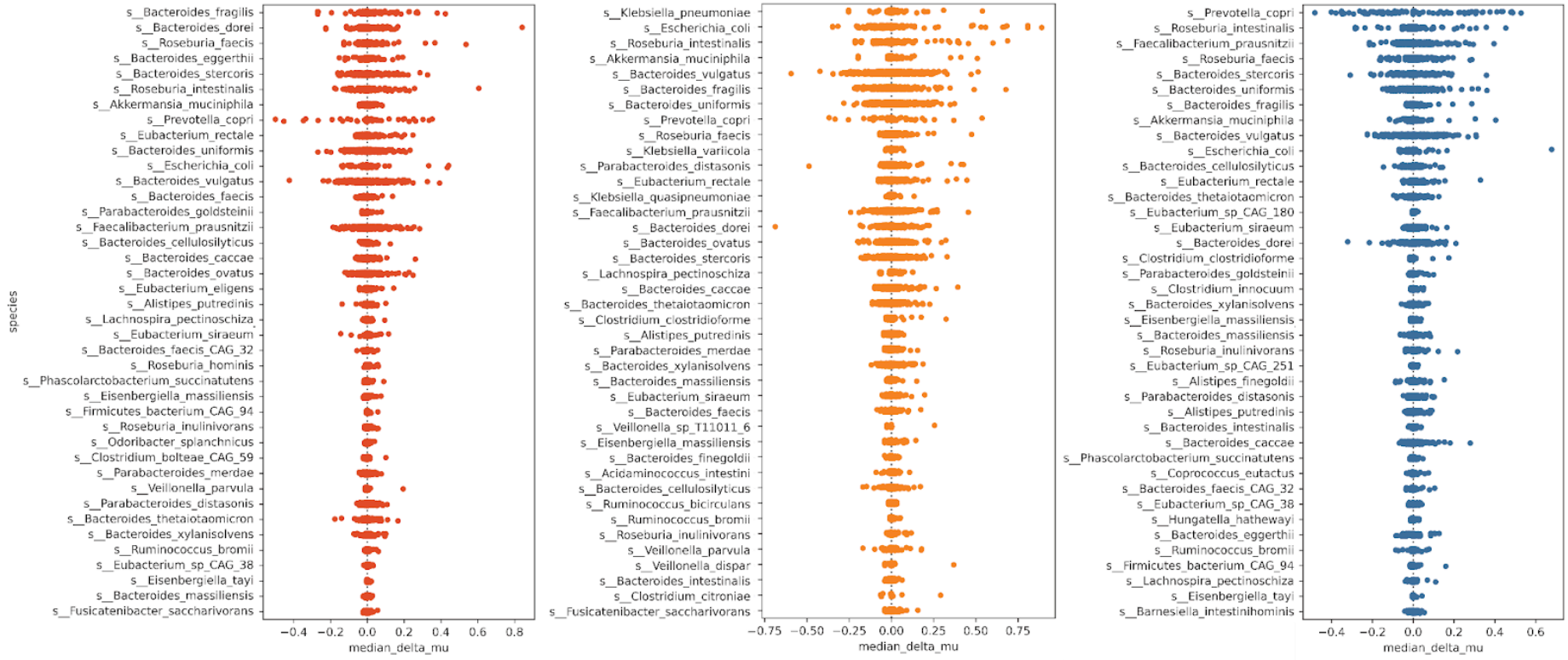
Figure 3.5. (Next page) Microbial species with the largest dispersion (s) in each condition. The microbes with the largest dispersion were UC (red), CD (orange) and controls (blue). Dispersion is represented by the log inverse of *s*. This means the greater the value, the more dispersion is captured by the model. The boxes represent the CI for each species extracted from the model (the smaller the interval, the higher the confidence). The prior is presented as the grey dashed lines, which are 5% and 95% percentiles; the green dashed is the median; the red dashed line is the mean.

Figure 3.6. (Page after the next) Microbial species change compared to the baseline across all patients (Δμ). The top 40 microbes with the largest dispersion in UC (red), CD (orange), and controls (blue) with the matching Δμ. The larger the absolute value of Δμ the more dispersion of that species in that sample comparatively to the baseline at that time point.


s__Bacteroides_fragilis
s__Bacteroides_dorei
s__Roseburia_faecis
s__Bacteroides_eggerthii
s__Bacteroides_stercoris
s__Roseburia_intestinalis
s__Akkermansia_muciniphila
s__Prevotella_copri
s__Eubacterium_rectale
s__Bacteroides_uniformis
s__Escherichia_coli
s__Bacteroides_vulgatus
s__Bacteroides_faecis
s__Parabacteroides_goldsteinii
s__Faecalibacterium_prausnitzii
s__Bacteroides_cellulosilyticus
s__Bacteroides_caccae
s__Bacteroides_ovatus
s__Eubacterium_eligens
s__Alistipes_putredinis
s__Lachnospira_pectinoschiza
s__Eubacterium_siraeum
s__Bacteroides_faecis_CAG_32
s__Roseburia_hominis
s__Phascolarctobacterium_succinatutens
s__Eisenbergiella_massiliensis
s__Firmicutes_bacterium_CAG_94
s__Roseburia_inulinivorans
s__Odoribacter_splanchnicus
s__Clostridium_bolteae_CAG_59
s__Parabacteroides_merdae
s__Veillonella_parvula
s__Parabacteroides_distasonis
s__Bacteroides_thetaiotaomicron
s__Bacteroides_xylanisolvens
s__Ruminococcus_bromii
s__Eubacterium_sp_CAG_38
s__Eisenbergiella_tayi
s__Bacteroides_massiliensis
s__Fusicatenibacter_saccharivorans
species
−7
−6
−5
−4
−3
−2
−1
log(inverse_s)


s__Klebsiella_pneumoniae
s__Escherichia_coli
s__Roseburia_intestinalis
s__Akkermansia_muciniphila
s__Bacteroides_vulgatus
s__Bacteroides_fragilis
s__Bacteroides_uniformis
s__Prevotella_copri
s__Roseburia_faecis
s__Klebsiella_variicola
s__Parabacteroides_distasonis
s__Eubacterium_rectale
s__Klebsiella_quasipneumoniae
s__Faecalibacterium_prausnitzii
s__Bacteroides_dorei
s__Bacteroides_ovatus
s__Bacteroides_stercoris
s__Lachnospira_pectinoschiza
s__Bacteroides_caccae
s__Bacteroides_thetaiotaomicron
s__Clostridium_clostridioforme
s__Alistipes_putredinis
s__Parabacteroides_merdae
s__Bacteroides_xylanisolvens
s__Bacteroides_massiliensis
s__Eubacterium_siraeum
s__Bacteroides_faecis
s__Veillonella_sp_T11011_6
s__Eisenbergiella_massiliensis
s__Bacteroides_finegoldii
s__Acidaminococcus_intestini
s__Bacteroides_cellulosilyticus
s__Ruminococcus_bicirculans
s__Ruminococcus_bromii
s__Roseburia_inulinivorans
s__Veillonella_parvula
s__Veillonella_dispar
s__Bacteroides_intestinalis
s__Clostridium_citroniae
s__Fusicatenibacter_saccharivorans
species
−7
−6
−5
−4
−3
−2
−1
log(inverse_s)


s__Prevotella_copri
s__Roseburia_intestinalis
s__Faecalibacterium_prausnitzii
s__Roseburia_faecis
s__Bacteroides_stercoris
s__Bacteroides_uniformis
s__Bacteroides_fragilis
s__Akkermansia_muciniphila
s__Bacteroides_vulgatus
s__Escherichia_coli
s__Bacteroides_cellulosilyticus
s__Eubacterium_rectale
s__Bacteroides_thetaiotaomicron
s__Eubacterium_sp_CAG_180
s__Eubacterium_siraeum
s__Bacteroides_dorei
s__Clostridium_clostridioforme
s__Parabacteroides_goldsteinii
s__Clostridium_innocuum
s__Bacteroides_xylanisolvens
s__Eisenbergiella_massiliensis
s__Bacteroides_massiliensis
s__Roseburia_inulinivorans
s__Eubacterium_sp_CAG_251
s__Alistipes_finegoldii
s__Parabacteroides_distasonis
s__Alistipes_putredinis
s__Bacteroides_intestinalis
s__Bacteroides_caccae
s__Phascolarctobacterium_succinatutens
s__Coprococcus_eutactus
s__Bacteroides_faecis_CAG_32
s__Eubacterium_sp_CAG_38
s__Hungatella_hathewayi
s__Bacteroides_eggerthii
s__Ruminococcus_bromii
s__Firmicutes_bacterium_CAG_94
s__Lachnospira_pectinoschiza
s__Eisenbergiella_tayi
s__Barnesiella_intestinihominis
species
−8
−6
−4
−2
log(inverse_s)

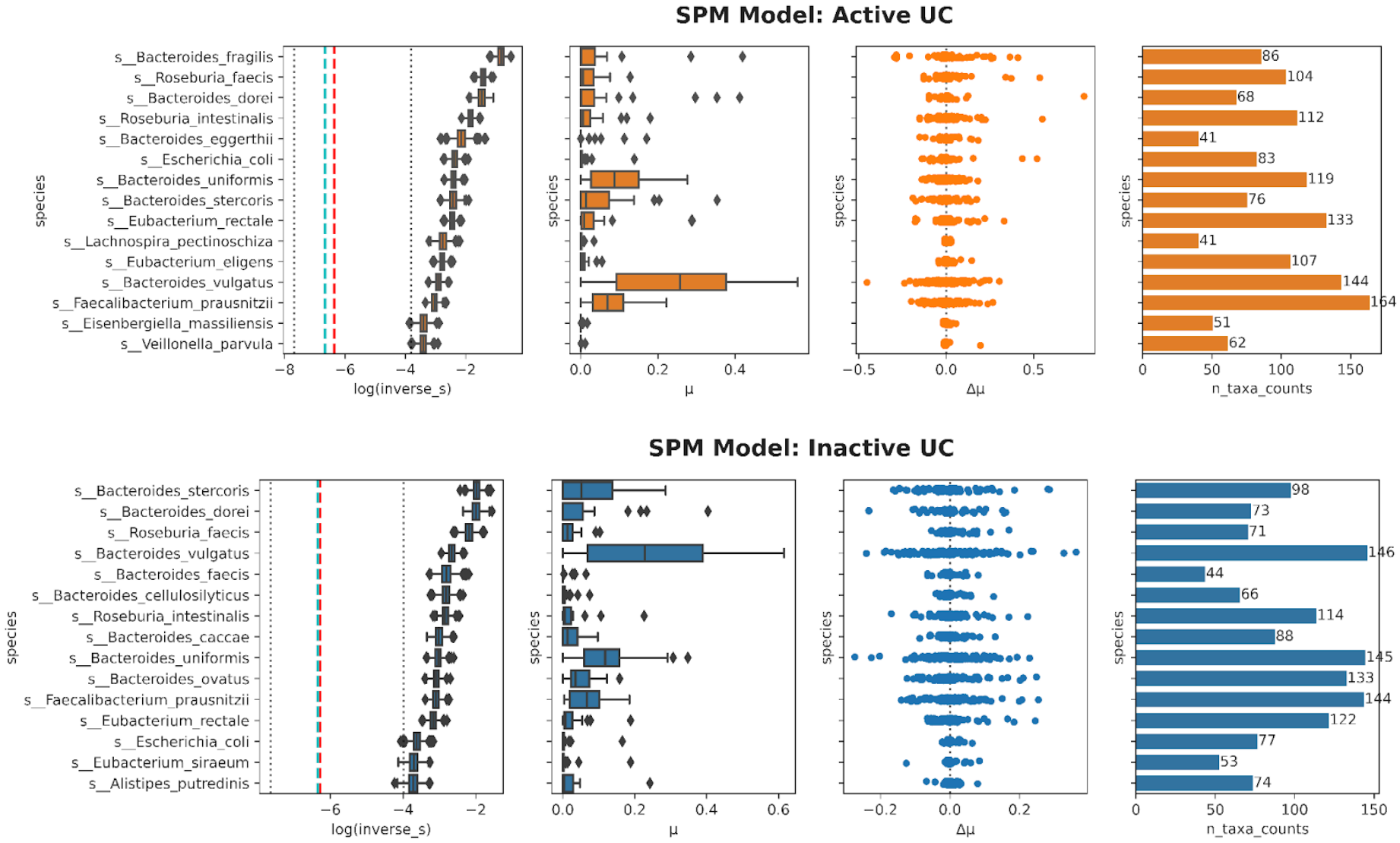

species
s__Bacteroides_fragilis
s__Bacteroides_dorei
s__Roseburia_faecis
s__Bacteroides_eggerthii
s__Bacteroides_stercoris
s__Roseburia_intestinalis
s__Akkermansia_muciniphila
s__Prevotella_copri
s__Eubacterium_rectale
s__Bacteroides_uniformis
s__Escherichia_coli
s__Bacteroides_vulgatus
s__Bacteroides_faecis
s__Parabacteroides_goldsteinii
s__Faecalibacterium_prausnitzii
s__Bacteroides_cellulosilyticus
s__Bacteroides_caccae
s__Bacteroides_ovatus
s__Eubacterium_eligens
s__Alistipes_putredinis
s__Lachnospira_pectinoschiza
s__Eubacterium_siraeum
s__Bacteroides_faecis_CAG_32
s__Roseburia_hominis
s__Phascolarctobacterium_succinatutens
s__Eisenbergiella_massiliensis
s__Firmicutes_bacterium_CAG_94
s__Roseburia_inulinivorans
s__Odoribacter_splanchnicus
s__Clostridium_bolteae_CAG_59
s__Parabacteroides_merdae
s__Veillonella_parvula
s__Parabacteroides_distasonis
s__Bacteroides_thetaiotaomicron
s__Bacteroides_xylanisolvens
s__Ruminococcus_bromii
s__Eubacterium_sp_CAG_38
s__Eisenbergiella_tayi
s__Bacteroides_massiliensis
s__Fusicatenibacter_saccharivorans
−0.4 −0.2 0.0 0.2 0.4 0.6 0.8
median_delta_mu
s__Klebsiella_pneumoniae
s__Escherichia_coli
s__Roseburia_intestinalis
s__Akkermansia_muciniphila
s__Bacteroides_vulgatus
s__Bacteroides_fragilis
s__Bacteroides_uniformis
s__Prevotella_copri
s__Roseburia_faecis
s__Klebsiella_variicola
s__Parabacteroides_distasonis
s__Eubacterium_rectale
s__Klebsiella_quasipneumoniae
s__Faecalibacterium_prausnitzii
s__Bacteroides_dorei
s__Bacteroides_ovatus
s__Bacteroides_stercoris
s__Lachnospira_pectinoschiza
s__Bacteroides_caccae
s__Bacteroides_thetaiotaomicron
s__Clostridium_clostridioforme
s__Alistipes_putredinis
s__Parabacteroides_merdae
s__Bacteroides_xylanisolvens
s__Bacteroides_massiliensis
s__Eubacterium_siraeum
s__Bacteroides_faecis
s__Veillonella_sp_T11011_6
s__Eisenbergiella_massiliensis
s__Bacteroides_finegoldii
s__Acidaminococcus_intestini
s__Bacteroides_cellulosilyticus
s__Ruminococcus_bicirculans
s__Ruminococcus_bromii
s__Roseburia_inulinivorans
s__Veillonella_parvula
s__Veillonella_dispar
s__Bacteroides_intestinalis
s__Clostridium_citroniae
s__Fusicatenibacter_saccharivorans
−0.75 −0.50 −0.25 0.00 0.25 0.50 0.75
median_delta_mu
s__Prevotella_copri
s__Roseburia_intestinalis
s__Faecalibacterium_prausnitzii
s__Roseburia_faecis
s__Bacteroides_stercoris
s__Bacteroides_uniformis
s__Bacteroides_fragilis
s__Akkermansia_muciniphila
s__Bacteroides_vulgatus
s__Escherichia_coli
s__Bacteroides_cellulosilyticus
s__Eubacterium_rectale
s__Bacteroides_thetaiotaomicron
s__Eubacterium_sp_CAG_180
s__Eubacterium_siraeum
s__Bacteroides_dorei
s__Clostridium_clostridioforme
s__Parabacteroides_goldsteinii
s__Clostridium_innocuum
s__Bacteroides_xylanisolvens
s__Eisenbergiella_massiliensis
s__Bacteroides_massiliensis
s__Roseburia_inulinivorans
s__Eubacterium_sp_CAG_251
s__Alistipes_finegoldii
s__Parabacteroides_distasonis
s__Alistipes_putredinis
s__Bacteroides_intestinalis
s__Bacteroides_caccae
s__Phascolarctobacterium_succinatutens
s__Coprococcus_eutactus
s__Bacteroides_faecis_CAG_32
s__Eubacterium_sp_CAG_38
s__Hungatella_hathewayi
s__Bacteroides_eggerthii
s__Ruminococcus_bromii
s__Firmicutes_bacterium_CAG_94
s__Lachnospira_pectinoschiza
s__Eisenbergiella_tayi
s__Barnesiella_intestinihominis
−0.4 −0.2 0.0 0.2 0.4 0.6
median_delta_mu

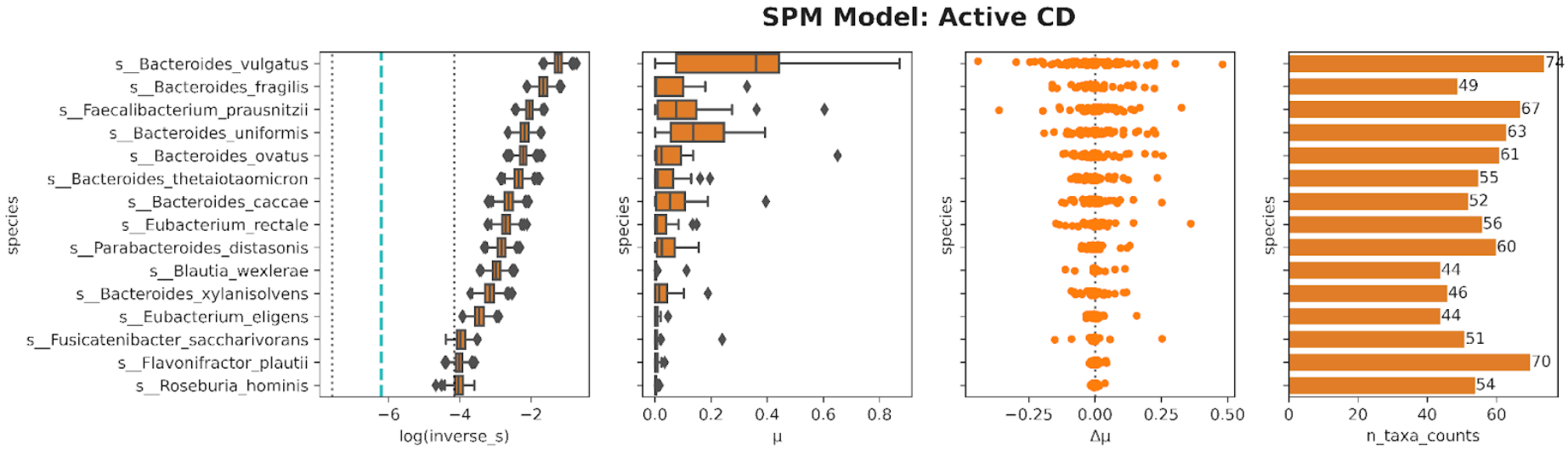
3.3.2 Inferring species dispersion between inactive IBD and active IBD states

To investigate the role species-level dispersion plays in increased disease activity, the IBD cohorts were split into two sub-groups within the disease; patients which remained in an inactive disease state (in UC a SCCAI < 2.5 and CD HBI < 5) and patients who did experience an increase in disease activity leading to an active disease state (UC a SCCAI ≥ 2.5 and CD HBI ≥ 5) during the 52 weeks of the study. Comparatively to the previous study, I am now looking within the disease and, therefore also interested in moderate disease activity. I.e. patients who are sitting around the threshold of SCCAI and HBI. There is a slight terminology change from the notion outlined in the first experiments (3.3.1) where we define sub-groups as inactive or active. As this model only considered and inferred the entire microbiome composition into a single parameter. Therefore, the previous model (1) cannot infer a species *s* for each patient but instead defines a global *s* for that patient, and (2) it would not enable the visualisation of what happens leading up to the flare point.

When comparing active and inactive UC, there was some overlap between the top-ranked species. In particular, *Roseburia faecis*, *Bacteroides dorei* and *Bacteroides faecis* (Figure 3.7 and Figure 3.8). These were also seen to be highly unstable when comparing disease and healthy individuals as well. Again, *Bacteroides vulgatus* displays a large amount of dispersion in both the active and inactive sstates in both CD and UC (Figure 3.7 and 3.8). Interestingly, it was ranked the most unstable species in active CD (Figure 3.8). Moreover, *Faecalibacterium prausnitzii* was consistently placed in the most dispersed species of bacteria when comparing the active and inactive states of both UC and CD. It was particularly high in CD. Other notable species include *Escherichia coli*, which was also seen to be highly dispersed in active UC (Figure 3.7 and Figure 3.8).

Figure 3.7. (Next page) SPM model dispersion difference between active and inactive IBD. Active and inactive UC defined by SCCAI being inactive < 2.5 and >= 2.5 being active. The top species were selected based on their distributions being the furthest from the prior (dashed lines) and ranked by their dispersion parameter S. Active disease is shown in orange and inactive disease is shown in blue in both UC and CD.


SPM Model: Active UC
s__Bacteroides_fragilis
s__Roseburia_faecis
s__Bacteroides_dorei
s__Roseburia_intestinalis
s__Bacteroides_eggerthii
s__Escherichia_coli
s__Bacteroides_uniformis
s__Bacteroides_stercoris
s__Eubacterium_rectale
s__Lachnospira_pectinoschiza
s__Eubacterium_eligens
s__Bacteroides_vulgatus
s__Faecalibacterium_prausnitzii
s__Eisenbergiella_massiliensis
s__Veillonella_parvula
species
log(inverse_s)
μ
Δμ
n_taxa_counts
86
104
68
112
41
83
119
76
133
41
107
144
164
51
62
SPM Model: Inactive UC
s__Bacteroides_stercoris
s__Bacteroides_dorei
s__Roseburia_faecis
s__Bacteroides_vulgatus
s__Bacteroides_faecis
s__Bacteroides_cellulosilyticus
s__Roseburia_intestinalis
s__Bacteroides_caccae
s__Bacteroides_uniformis
s__Bacteroides_ovatus
s__Faecalibacterium_prausnitzii
s__Eubacterium_rectale
s__Escherichia_coli
s__Eubacterium_siraeum
s__Alistipes_putredinis
species
log(inverse_s)
μ
Δμ
n_taxa_counts
98
73
71
146
44
66
114
88
145
133
144
122
77
53
74


SPM Model: Active CD
s__Bacteroides_vulgatus
s__Bacteroides_fragilis
s__Faecalibacterium_prausnitzii
s__Bacteroides_uniformis
s__Bacteroides_ovatus
s__Bacteroides_thetaiotaomicron
s__Bacteroides_caccae
s__Eubacterium_rectale
s__Parabacteroides_distasonis
s__Blautia_wexlerae
s__Bacteroides_xylanisolvens
s__Eubacterium_eligens
s__Fusicatenibacter_saccharivorans
s__Flavonifractor_plautii
s__Roseburia_hominis
species
log(inverse_s)
μ
Δμ
n_taxa_counts
74
49
67
63
61
55
52
56
60
44
46
44
51
70
54

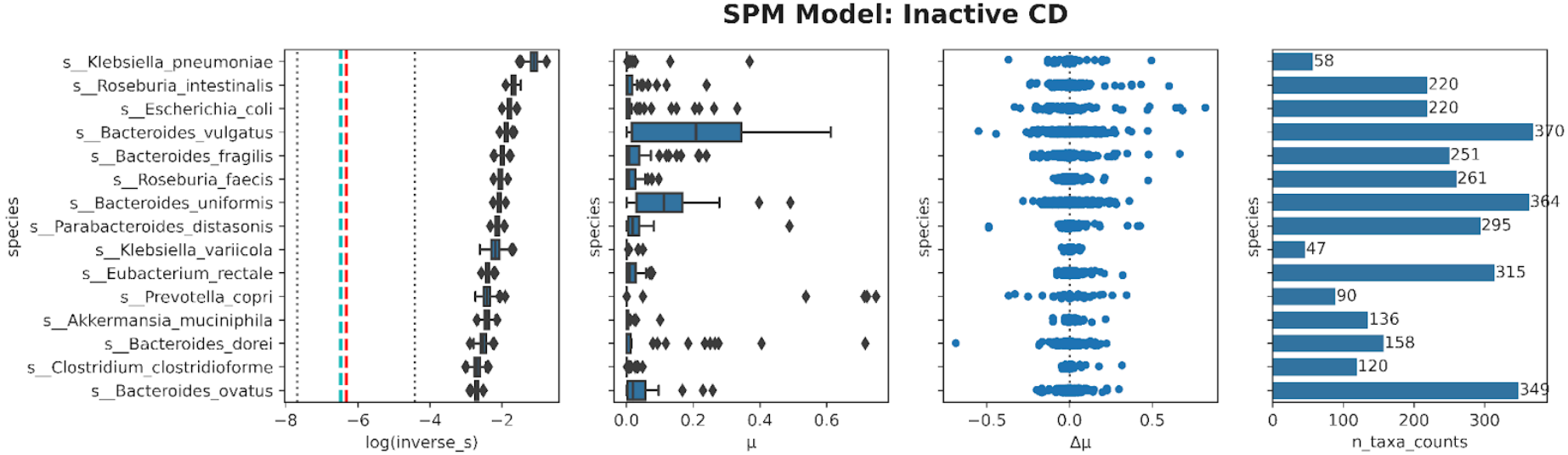

SPM Model: Inactive CD
s__Klebsiella_pneumoniae
s__Roseburia_intestinalis
s__Escherichia_coli
s__Bacteroides_vulgatus
s__Bacteroides_fragilis
s__Roseburia_faecis
s__Bacteroides_uniformis
s__Parabacteroides_distasonis
s__Klebsiella_variicola
s__Eubacterium_rectale
s__Prevotella_copri
s__Akkermansia_muciniphila
s__Bacteroides_dorei
s__Clostridium_clostridioforme
s__Bacteroides_ovatus
species
log(inverse_s)
μ
Δμ
n_taxa_counts
58
220
220
370
251
261
364
295
47
315
90
136
158
120
349

Figure 3.8. SPM model dispersion difference between active and inactive IBD. Active and inactive CD defined by HBI being inactive < 5 and >= 5 being active. The top species were selected based on their distributions being the furthest from the prior (dashed lines) and ranked by their dispersion parameter S. Active disease is shown in orange and inactive disease is shown in blue in both UC and CD.

To investigate the co-abundance of each of these identified species from the SPM model a regression analysis between the disease activity and relative abundance was conducted. This showed that when looking at just the relative abundance of these species against the disease activity, it did not display any significant correlation. This further demonstrates the importance of implementing a patient-specific baseline (Supplementary Figure 3.2-3.5). Furthermore, the correlation between the set of unique species was plotted in a correlation matrix to assess the co-abundance of these species.

The bacterial species were identified using the inferred dispersion parameter *S* produced by the model defined in 3.2.1.2. The top 15 most unstable species were identified. Then the inferred μ (patient-specific baseline) was used and extracted from the model. The difference from the patient baseline was calculated to the current sample at that time point and then plotted over the time course of the study (Supplementary Figure 3.1).

3.4 Discussion

An important concept to note here is that dispersion (*s*) and diversity (e.g. alpha or beta diversity) are two different notions of biological variation. Particularly, in this case, dispersion focuses on the exploration of variable differences among individuals or later on, between sub-diseases, while diversity is an exploration of numbers of distinct types found within a sample (Gregorius and Kosman, 2017).

High levels of variability in the microbiome can be an indicator of perturbations caused by a number of different factors. These could include individual variation, environmental influences, and cross-talk between microbes and could be an indicator of change in health. The SPM model tries to mitigate the individual variation by accounting for that patient's baseline microbiome signature and then seeing if the other patients across the groups share

the same dispersion of species. The species precision model showed when comparing inactive CD and active CD that *Faecalibacterium prausnitzi* was one of the most dispersed species. *Faecalibacterium prausnitzi* is an important regulator of intestinal inflammation (Cao, Shen and Ran, 2014; Lopez-Siles et al., 2017; Sokol et al., 2008) and has been shown to anti-inflammatory effects in cellular and TNBS colitis models. The authors demonstrated this was partly because of metabolites that were secreted which have the ability to inhibit NF-κB activation and IL-8 production (Sokol et al., 2008). *Faecalibacterium prausnitzi* has been shown previously to be decreased in IBD patients when compared to healthy controls (Sokol et al., 2008) and also in other longitudinal studies have been shown to correlate with changes in faecal calprotectin which is used as marker of inflammation (Björkqvist et al., 2019; Cao, Shen and Ran, 2014). The authors found an inverse correlation between *Faecalibacterium prausnitzi* and faecal calprotectin levels. However, it is still unclear where this decrease is casual or a reflection of the dysbiotic microbiome of CD patients.

When comparing inactive UC and active UC, the stand-out species was *Escherichia coli*. *Escherichia coli* is a bacteria that normally lives within the human gut and can be completely harmless, but a few strains of Escherichia coli are pathogenic. Pathogenic strains of Escherichia coli have been linked to IBD, with Adherent invasive *Escherichia coli* in CD and diffusely adherent *Escherichia coli* in UC (Kotlowski et al., 2007; Petersen et al., 2009). Singh et al. have shown that during an inflammatory response in the gut, even the commensal bacterial strains of *Escherichia coli* can contribute to disease (Singh et al., 2015), importantly demonstrating the mechanism at which *Escherichia coli* could inhibit the host innate immune response through the release of siderophore.

By comparing the remission and flaring patient dispersion and accounting for the patient's microbiome and, therefore, inter-patient variation, the models have identified microbes that are seen to be more variable between disease states. Interestingly, there seems to be a larger difference in the microbiome dispersion in UC inactive vs active than in CD inactive vs active disease. However, one commonality when comparing disease states independently was that the most dispersion was seen in *Bacteroides* genera. The Human Microbiome Project found that in the healthy human gut *Bacteroides* were one of the abundant genera. Since then, a number of studies have shown *Bacteroides* genera are reduced in IBD (Zhou and Zhi, 2016; Conte et al., 2006). Interestingly, *Bacteroides vulgatus* has been implicated in both the decrease (Zhou and Zhi, 2016) and increase (Mills et al., 2022) of disease severity,

suggesting that these predominant bacterial species should be investigated further in the case of IBD.

In UC, n=24 patients remained in remission throughout the course of the study, with n=6 patients experiencing a flare. Meanwhile, in CD, n=32 patients remained in remission compared to n=18 patients who experienced a flare. This ratio of remission to flare is not unsurprising, as flares in IBD patients with correct treatment can go months or even years without experiencing any symptoms or only experiencing mild symptoms. The small sample size of the cohort evaluated here means that further work would be required to validate the findings and conduct a more robust overall evaluation of the model.

In most other longitudinal microbiome studies and statistical methods, each patient had a shared baseline or start point. For example, *Velten et al.* developed MEFISTO to integrate multi-modal longitudinal data with the aim to disentangle the sources of variation that either change slowly compared with the covariate and those which are independent of the covariate (Velten et al., 2022). In the application to the microbiome, they applied MEFISTO to investigate how the infant microbiome develops after birth, exploring the effects of delivery methods, diets and months after birth (Bokulich et al., 2016; Velten et al., 2022; Martino et al., 2021). This meant that all individuals had the same starting point e.g. birth. This means that you can use methods such as dynamic time warping (DTW) or imputation based on similar time points.

The models developed in this chapter demonstrate a new method for analysing longitudinal microbiome data with no respect to their starting point by regressing the individual microbiome baseline. The reasoning behind this model design choice is that many complex diseases progress in a patient-specific way, and most clinical studies are unable to have individuals who all share a common starting point.

The model demonstrates that the shifts in the microbiome over time occur in a patient-individual manner. However, the model is still vulnerable to noise and does not account well for associations between bacterial species. This means further work is needed to extract the underlying associations in the data. This could be achieved by the addition of a hierarchical model to try to model the complexities of microbiome composition and potentially the addition of zero inflation to handle structural zeros (Sankaran and Holmes,

2019). These additional features have been implemented by others and therefore since the development of this model, several Bayesian latent variable models have shown to be performant compared to traditional microbiome methods to stratify patients (Sankaran and Holmes, 2019). An example of this is the zero-inflated Latent Dirichlet Allocation model (zinLDA) developed by Deek et al. The authors' model is a flexible implementation of the Latent Dirichlet Allocation model that accounted for both the sparsity of microbiome data, while also allowing for zero-inflated observations in microbial counts data (Deek and Li, 2020). However, it should be noted that this model was developed for application to cross-sectional data not longitudinal data and although zero inflation was not accounted for within the SPM model, it was mitigated by the removal of microbial features whose prevalence was below 10% across all samples.

In this chapter, I developed and demonstrated a method for exploring, detecting shifts, and as a potential unsupervised feature selection step for downstream analysis or prediction. Future work for this model could include incorporating statistical tests within the model to determine the distributions that are most different from the prior. This could be done using a Kolmogorov–Smirnov test for example. Furthermore, using the SPM model as a basic hierarchical model could be implemented. In this model, each SPM pooled model would be created as a sub-model just for the specific patient and these models would then be integrated together to form the Hierarchical model. This in turn would capture the individual's temporal trajectory better and create an overall more robust model, as Bayesian hierarchical models posterior distribution is less sensitive to flexible hierarchical priors.

Chapter 4: Predicting healthy and unhealthy status in inflammatory bowel disease from multi-omic microbiome data

4.1 Introduction

Dimensionality reduction is a powerful technique widely used in biomarker discovery to identify and isolate relevant signals from complex biological datasets (Velliangiri, Alagumuthukrishnan and Thankumar joseph, 2019; Velten et al., 2022; Hira and Gillies, 2015; Bhadra et al., 2022; Argelaguet et al., 2018; Dong and Bacher, 2022). The goal of dimensionality reduction is to reduce the input data set into a new lower dimensional space. This aims to identify patterns, signals and trends which would not have been detectable from the raw data. This is done by taking the high-dimensional input data and identifying a representation of that data in a lower-dimensional space (Xu et al., 2018). Notably, this lower-dimensional space remains faithful to the original input data and can be reconstructed back into the original input.

Dimensionality reduction methods have the added advantage of enabling the visualisation of the data to understand the structure of the datasets. Some other advantages of dimensionality reduction include (this list was updated from (Xu et al., 2018; Velliangiri, Alagumuthukrishnan and Thankumar joseph, 2019) :

- As the number of dimensions decreases, storage/memory requirements decrease.
- Reduces computational (time) complexity
- Removal of redundant, irrelevant, and noisy data from the original dataset.
- It can improve the quality of the original data (for example, denoising).
- It is challenging to visualise data in higher dimensions. So, reducing the dimension may allow us to design and examine patterns more clearly.
- It simplifies the process of classification and also improves efficiency.

However, dimensionality reduction also has limitations. Firstly, all dimensionality reduction techniques result in some loss of information when the data is “squeezed” from its high-dimensionality state to a low-dimensionality state. This means that when applying these methods, one must balance the tradeoff between information loss and the improved interpretability gained from dimensionality reduction (Xu et al., 2018; Armstrong et al., 2022). Another known limitation of dimensionality reduction methods is the misinterpretation of the projection and the potential display of structures that may not be present in the original input data. Finally, as dimensionality reduction methods tend to be unsupervised, it can be challenging to determine whether the embedding accurately represents the original dataset (Velliangiri, Alagumuthukrishnan and Thankumar Joseph, 2019). This is particularly true when there is a lack of the original labelled data, and instead, annotations are derived from clusters or communities found in the embeddings.

One of the most commonly used dimensionality reduction methods is principal component analysis (PCA), which projects the data onto a lower-dimensional space while preserving as much of the variance in the data as possible (Ma and Dai, 2011). PCA has been applied to various biological data, including proteomics, genomics, and metabolomics, as either a preprocessing step, quality control, an exploratory step or a feature extraction step (Ma and Dai, 2011). Another popular method for dimensionality reduction is independent component analysis (ICA), which seeks to identify and isolate independent signals in the data. ICA has been applied to multiple different Omics datasets, including; metabolomics (Liu et al., 2016; Krumsiek et al., 2012), metaproteomics (Sompairac et al., 2019), microarray (Engreitz et al., 2010) and transcriptomics (Engreitz et al., 2010; Cantini et al., 2019).

4.1.1 Aims

In this chapter, in collaboration with my industrial partner BenevolentAI, we develop machine learning (ML) methods for predicting disease activity of inflammatory bowel disease patients (IBD) based on microbiome omics data (metagenomics, metaproteomics and metabolomics data). Furthermore, this chapter will utilise both unsupervised and supervised models together. The aim here is to use the unsupervised models to be two fold, 1) as a feature extraction step and 2) an exploratory analysis to find latents which describe the biological signal of interest (in this case healthy vs unhealthy conditions). Then to

evaluate the effectiveness of these discovered latents at determining the condition, they will then be used as the input of the supervised models.

This chapter's aims were as follows:

- Implement and evaluate machine learning methods and apply them to the microbiome and metabolome
- Evaluate the performance of baseline transformation methods in predicting the difference disease states
- Extract features of interest using interpretable machine learning efficient methods
- Identify subsets of features that can be used to explain the differences between conditions (e.g. give biological context to the findings of the models)
- Identify potential prognostic indicators from metabolome and microbiome between IBD and healthy controls

4.2 Methods

4.2.1 Data preprocessing

4.2.1.1 Metagenomics

Metagenomics data was taken from Lloyd-Price et al., 2019 study to compare the gut microbial ecosystem in inflammatory bowel diseases. Raw reads were downloaded from SRA BioPorject PRJNA398089. MetaPhlAn3 was used as it is widely viewed as the industry standard approach for shotgun metagenomics preprocessing in addition to a large amount of support for the pipeline. This means the pipeline is reliable and robust for use in a production setting. For further information on the preprocessing pipeline and dataset size see Chapter 2 section 2.2.1 and Table 2.1.

4.2.1.2 Metabolomics

The data used in this study is from the HMP2 project. The metabolomics data was acquired from Workbench (http://www.metabolomicsworkbench.org), Project ID PR000639. The

authors used the following steps to process the raw LC–MS. Nontargeted data were processed using Progenesis QIsoftware, which is a software suite to measure the levels of small molecules, lipids, and proteins in a sample. Unknown peaks were labelled by their method, m/z and retention time. To identify non-target metabolites LC-MS peaks were matched based on the RT and masses or by mapping to the author's own internal database of compounds. This resulted in 551 metabolites from 546 samples, derived from 106 subjects. Of the 106 patients (CD=50, UC=30, non-IBD=26).

4.2.2 Normalisation and Transformation methods

4.2.2.1 Normalisation methods

4.2.2.1.1 Relative abundance normalisation

One of the most common normalisation methods used for compositional data is relative abundance. Essentially, relative abundance provides a measure of how frequent a species is in a sample relative to the other species found in the sample. A key strength of relative abundance is how simple the method is both conceptually and to implement. However, as abundances within a given sample are not truly independent of each other, normalising using relative abundance makes downstream inference more challenging.

4.2.2.1.2 Probabilistic quotient normalisation

Probabilistic quotient normalisation (PQN) was introduced by Dieterle *et al* as a robust normalisation method to account for the complexities found in biological datasets. The approach of PQN assumes that changes in the concentrations of single analytes only influence parts of the spectra, whereas changes in the overall concentration of a sample influence the complete spectrum (Dieterle et al., 2006). In brief, PQN can be thought of as a normalisation of the sample data by the median fold change of all samples as the reference.

PQN is calculated using the following steps as defined by Dieterle *et al*:

1. Perform an integral normalisation (typically a constant integral of 100 is used).
2. Choose/calculate the reference spectrum (the best approach is the calculation of the median spectrum of control samples).
3. Calculate the quotients of all variables of interest of the test spectrum with those of the reference spectrum.

4. Calculate the median of these quotients.
5. Divide all variables of the test spectrum by this median.

These steps were implemented using a custom Python function utilising the Numpy (Harris et al., 2020) for speed of calculations.

4.2.2.2 Transformation methods

Compositional data are data in which the relative abundances of different components or parts add up to a constant, such as microbiome data (metagenomics, metabolomics or metaproteomics data), which consist of the relative abundances of different microbial taxa, proteins or metabolites. Because compositional data have unique statistical properties, such as closure (the sum of the relative abundances is constant), they require special treatment in statistical analysis. One of the most common approaches is transforming the data to alleviate these properties. However, the choice of transformation should depend on the model, the question being asked and the type of data it's being applied to.

4.2.2.2.1 Centred log-ratio transformation (CLR)

As mentioned in Chapter 2, CLR is a transformation method that can be used to remove the constraint that is present on compositional data. This enables the data to be used by statistical methods and other downstream approaches and is a fundamental tool used by researchers to explore the complexities of compositional data (Faith, 2015). This approach would be robust if microbiome data were not sparse. The sparsity of the data is problematic for these transformation algorithms as they cannot compute the geometric mean if the vector they are being applied to is 0 (Gloor and Reid, 2016; Mandal et al., 2015). This means a pseudocount must be added to any zero values before applying this transformation.

Centred log-ratio (CLR) transformation (Aitchison, 1982) is defined as:

$$clr(X) = (log(x_1/g_x),\ log(x_2/g_x) \ldots,\ log(x_D/g_x))$$

(Equation.4.1)

4.2.2.2.2 Longitudinal patient-baseline transformation

Following up on the work conducted in Chapter 3 by defining a patient-specific baseline, in this chapter the work is extended to create two new methods. These methods are (1) Log

Fold Change baseline transformation (FCBT) and (2) Subtracted baseline transformation (SBT). These methods were applied either as relative abundance or counts depending on the input data. They could also be extended to group-specific transformation for intra-patient rather than inter-patient transformation. These methods were implemented as a custom Python function.

For both implementations, the baseline is calculated using the method defined in chapter 3 (Equations.3.1-3.14).

For the Log Fold Change baseline transformation (FCBT), each patient-specific baseline was created by generating the mean composition of that patient over all their measures. Let $X \in \mathbb{R}^{NxD}$ and $FCBF$ be the transformed matrix of X such that $FCBF \in \mathbb{R}^{NxD}$. The following formula describes the implementation to calculate FCBT where X_{k_i} is the k-th patient for 1,..., K, at the i-th feature for 1,..., D in X. Finally, μ_{k_i} is the baseline for that patient k for the i -th feature. It can be described as seen by Equation 4.2:

$$FCBT\left(X_{k_i}\right) = log_2\left(\frac{X_{k_i}}{\mu_{k_i}}\right).$$

(Equation 4.2)

Similarly, for subtracted baseline transformation (SBT), each patient-specific baseline was created by generating the mean composition of that patient and this time was subtracted from the other time point of that patient. Let $SBT \in \mathbb{R}^{NxD}$ be a matrix of transformed X where the baseline of patient k for 1,..., K, at the i-th feature for 1,..., D is subtracted from the match feature at X_i. It can be described as seen in Equation 4.3:

$$SBT\left(X_{k_i}\right) = X_{k_i} - \mu_{k_i}.$$

(Equation 4.3)

4.2.3 Matrix factorisation methods

4.2.3.1 Principal Component Analysis

The most popular method for dimensionality reduction is principal component analysis (PCA) (Pearson, 1901). PCA is a linear dimensionality reduction method used to project data

into a lower dimensional space. There are multiple different implementations of PCA, but this work uses the singular value decomposition (SVD) interpretation which results in a latent factor interpretation. SVD can be thought of as matrix factorisation that takes the input X and decomposes it into three matrices, $X = USV^T$. In the latent factor model these matrices are then rearranged and summed using linear contributions resulting in a weights matrix W. Therefore, we can rewrite decomposition as $X = WV^T$.

The input data, X, needs to be centred but not scaled for each feature before applying the SVD. The implementation from Scikit-learn V1.2.0 (Alex et al., n.d.) uses the LAPACK implementation of the full SVD or a randomised truncated SVD by Halko et al if the maximum dimensions of X > 500.

4.2.3.2 Independent Component Analysis

Independent component analysis (ICA) is a method which is typically applied to blind source separation problems (Herault and Jutten, 1986; Hyvärinen and Oja, 2000; Moldakarimov and Sejnowski, 2017). The process of blind source separation refers to an input dataset that is only mixed data with both the original sources or the mixing coefficients not being observed.

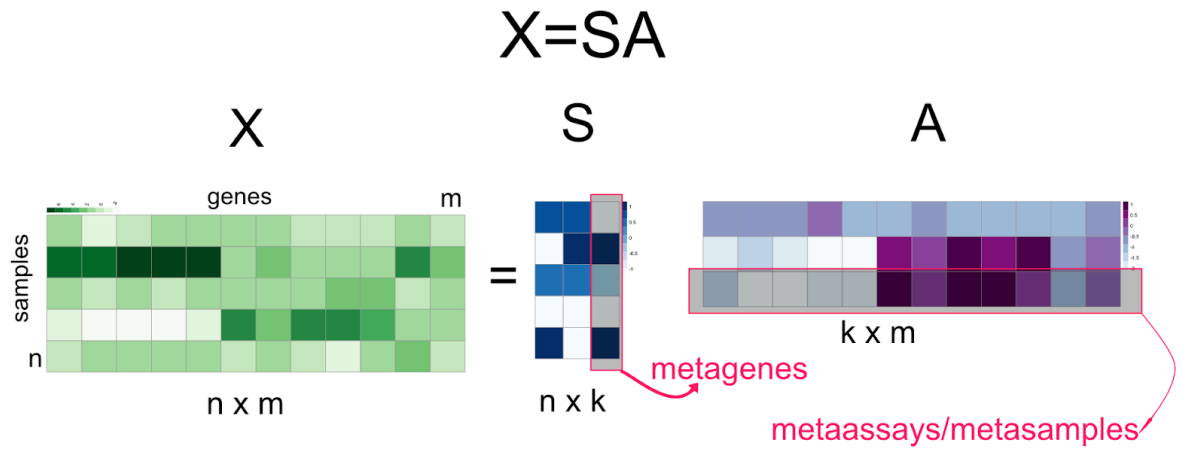


Figure 4.1. Schematic representing the matrix factorisation used within FastICA. The labels show where the features (genes, microbes, proteins, etc.) and samples after ICA are run on the raw input dataset. *S* represents the factors and *A* represents the loadings. Depending on the orientation of the input data. The factors (*S*) and loadings (*A*) can either represent meta-features or meta-samples depending on the objective of the analysis. Figure from (Akalin, 2020).

In the case of temporal data, treating each sample's time point as *i.i.d.* creates a generative model resulting in two latent variables representing the unobserved amplitude of the signal and two other latent variables that signal from the samples. The observed variables are described by the linear combination of latent variables, with the latent being the joint distribution that factorised variables described in equation 5.1, where $p(z)$ is the joint probability distribution.

$$p(z) = \prod_{j=1}^{M} p(z_j).$$

(Equation 4.4)

4.2.2.3 Factor Analysis

Factor analysis (FA) is closely related to PCA and often considered to be an extension of it. The objective of FA differs as it reconstructs the correlations and covariances between variables. Therefore, FA is a latent variable model where the observed variables and their covariance structure is modelled in terms of unobserved variables (i.e. latent), but these latent variables cannot be directly measured. There are two main forms of factor analysis, confirmatory factor analysis and exploratory factor analysis. In confirmatory factor analysis the number of factors is specified beforehand and which feature is related to a specific latent. While in exploratory factor analysis, all data points are related to every late variable. Like with other dimensionality reduction methods factor analysis provides valuable insights into underlying relationships in the data and has the added advantage of being highly interpretable.

4.2.2.4 Orthogonal Projection to Latent Structures Discriminant Analysis

Orthogonal Projection to Latent Structures (OPLS) Discriminant Analysis (DA) is a variant of the Partial Least Squares (PLS) algorithm (Trygg and Wold, 2002). Although OPLS is a multivariate regression model, it can easily be modified for binary classification problems. The biggest advantage of OPLS is its ability to further reduce the number of components needed in a dimension. In short, it achieves this by regressing variation in the data and therefore pushing the move of informative features together into a single component (Trygg and Wold, 2002; Stenlund et al., 2008; Biagioni et al., 2011).

4.2.4 Classification

Some of the models described above already have a supervised component to them, namely OPLS-DA. Other methods such as PCA, FA and ICA do not have an intrinsic classifier built into the algorithm. Although the performance of the unsupervised algorithms can be assessed using clustering methods, such as Lovivan clustering and HDBSCAN, the goal here is to predict class labels rather than investigate the underlying structure of the data. In this case, we are framing our problem set as a binary classification problem.

PCA, FA and ICA were used as feature engineering steps and the resulting data were passed to two different classifiers; Logistic Regression (LR) or Random Forest (RF). LR is an extension to the linear regression model for the application of classification. In this case, rather than fitting a hyperplane, LR squeezes the output of the linear equation into a logistic function. This in turns obtains a value between 0 and 1 which represents the probability between classes. It is generally considered to be one of the best approaches for low-dimensional and relatively noisy data (i.e. where the number of explanatory variables is equal to less than the number of noise variables). Furthermore, in addition to LR performance, it is also very interpretable as coefficients show the influence of a feature. Although this should be noted this differs from linear regression as this is not a linear contribution but instead a probability.

In comparison, RF is a more complex algorithm which is less sensitive to noise, can be applied to high dimensional data, and is less prone to overfitting. RF extends the standard Decision Tree using two additional approaches; bootstrapping and feature subsetting. In doing so, RF builds a large number of decision trees where each tree is trained on a random subset of the original training data and a random subset of features. These trees determine splits by the best subset of features and continue to grow until the maximum depth is reached (predefined by the user). The resulting models are then pooled together using majority voting to determine the class of predictive labels. Finally, like with LR, the resulting model has a high degree of interpretability. Feature importance scores measure how much each feature contributes to the overall accuracy of the model. There are several ways to calculate feature importance in random forests, but one common method is to use mean decrease impurity. This method computes the total reduction of the impurity measure (such as Gini index or entropy) of the decision tree due to a feature, averaged over all trees in the forest. A model that combines both dimensionality reduction and classification is the

Rotation Forest (RTF) (Juez-Gil et al., 2021). The model combines the benefits of feature extraction in PCA and tree-based ensemble methods to generate a highly versatile classifier. Although computationally extensive, it has been shown to perform very well on multiple datasets in different domains (Bagnall et al., 2018; Juez-Gil et al., 2021).

4.2.5 Model optimisation and evaluation

To evaluate the models the following experimental design was implemented. Each model was trained in parallel with the same random seed (starting seed was set to 42) and training set. The Scikit-learn Pipeline class was used to orchestrate each model's experiment. The advantage of using this architecture is that it enables multiple normalisation, transformation, and hyperparameters to be evaluated on the same data in a more efficient way. Normalisation and transformation methods were implemented as an extension of the Scikit-learn's TransformerMixin and BaseEstimator classes and therefore also built into the Pipeline.

To find the optimal number Horn's Parallel analysis is used. Briefly, it works by comparing the eigenvalues derived from the actual data with those obtained from randomly generated data sets of the same size and number of variables. The idea is that the actual data should have larger eigenvalues for the components that are meaningful. The optimal number of components is typically identified at the point where the actual data's eigenvalues begin to be smaller than those from the random data (Glorfeld, 1995; Gently Clarifying the Application of Horn's Parallel Analysis to Principal Component Analysis Versus Factor Analysis, 2014). This method is considered more accurate and reliable than the scree test, as it accounts for the chance that factors that might inflate the eigenvalues.

To ensure a robust evaluation of the model, this was then repeated 100 times each time changing the starting random seed by incrementing up by 1 each time, with the dataset permuted before training each time to ensure groups of patients were not together. For reproducibility a global random seed was set prior to analysis as stated above.

To assess the performance of the model several different metrics are calculated. This is due to each metric assessing a different measure of performance. The approach taken here aims to rigorously evaluate the performance of the algorithm, in contrast to other studies which tend to focus on specific metrics which might be misleading (e.g. just using accuracy alone).

The F1 score is a measure of accuracy which is calculated using the following formula:

$$F1 = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)},$$

(Equation 4.2)

where Precision is the count of correct positives results over the total number of positive results, either true positives (TP) or false positives (FP) predicted. Precision is also known as the False Positive Rate (FPR). It is calculated as:

$$Precision\ (FPR) = \frac{FP}{TN + FP},$$

(Equation 4.3)

where TN is the number of true negatives. Recall is the number of true positives, also known as the True Positive Rate (TPR), is calculated by dividing by the number of all relevant samples, which includes the number of false negatives (FN). It is calculated as follows:

$$Recall\ (TPR) = \frac{TP}{TP + FN}.$$

(Equation 4.4)

The Receiver Operator Curve (ROC) and Area Under the Curve (AUC) can be used together to represent the probability curve and the measure of separability respectively. AUC has the advantage over accuracy as it aggregates all the classification thresholds to produce a performance measure. Therefore, AUC describes the classifier's ability to distinguish between classes. The AUC can have a value between 0 and 1, where the higher the value the better the model's performance. AUC is calculated using the following equation:

$$AUC = \int_0^1 TPR\, d(FPR)$$

(Equation 4.5)

As the brier score is a loss metric, the smaller the resulting value the better. The brier score takes in the predicted probability score of the predicted label and true label and then

calculates the mean squared difference between the two. This results in a value between 0 and 1.

$$brier = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2,$$

(Equation 4.6)

where N is the number of predictions made, f is the predicted probability class and o is the actual outcome of the event.

Matthews correlation coefficient (MMC), also known as the phi coefficient, is the measure of quality of resulting classifications. It has the advantage of being a balanced measure which can be used even with large class imbalances. The metric returns a coefficient between -1 and +1, where -1 is an incorrect prediction, 0 is a random prediction and +1 is a perfect prediction. The metric uses the described as follows:

$$mcc = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}.$$

(Equation 4.7)

Using the metrics defined above, grid search was used to explore the parameter space of each model. Grid Search is an exhaustive search method that systematically goes through multiple combinations of hyperparameter values specified in a pre-defined grid. This approach evaluates each combination for the given model to determine which set of values yields the best performance according to a specified metric, in this case F1-score was used. Grid search was implemented using scikit-learn's *GridSearch*CV (https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html).

Leave-one group out cross validation (LOGOCV) is a technique used for evaluating ML algorithms performance, particularly when the data contains distinct groups or clusters (in this case a group would be a patient with repeated measures). This method is especially useful in situations where the data may have an inherent grouping structure, and it's important to ensure that the model generalises well across these groups. To ensure a robust evaluation of the model, this was then repeated 100 times each time changing the starting random seed by incrementing up by 1 each time, with the dataset permuted before training

each time to ensure groups of patients were not together. For reproducibility a global random seed was set prior to analysis as stated above. LOGOCV was implemented using scikit-learn's *LeaveOneGroupOut* model selection function (https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.LeaveOneGroupOut.html).

Finally, to compare the overall performance of the models the critical distance was calculated between each model. Critical distance is a statistical measure used to determine whether the performance differences between algorithms are significant. If the rank difference between two algorithms is greater than this critical distance, their performance is considered significantly different. It was implemented using the method here https://github.com/hfawaz/cd-diagram (Ismail Fawaz et al., 2019).

4.2.7 Blind source separation between phenotypes

Building on work conducted in Chapter 3, and extending the FastICA model, a context-aware implementation was created (this model was defined in section 4.2.2.2). In this case, using the patient-specific baseline approach time was accounted for in the model. This was either done by rotating the matrix to make the features (in this case either metabolites or microbes) as the sources. In addition to this to account for the time component of the data, both FCBT and SBT were also applied to the data before fitting the model.

Many studies have used ICA on gene expression, micro-array or metabolomic data where they used the samples as the sources (Chiappetta, Roubaud and Torrésani, 2004; Teschendorff et al., 2007; Engreitz et al., 2010; Biton et al., 2014; Nazarov et al., 2018; Cantini et al., 2019; Krumsiek et al., 2012; Liu et al., 2016). In this study, ICA is also used with the features as the sources. This is due to the nature of ICA in signal processing, where the features are normally time points. To achieve this, a wrapper function was used to extend the FastICA (Hyvärinen and Oja, 2000) implementation given by Scikit-learn (Alex et al., n.d.). The wrapper handles not only the data preprocessing in transposing the input matrix but also orientates the factors and loadings such that can be interpreted downstream.

The selection of the number of components is vital and no trivial problem for ICA as it does not have the same orthogonality constraint as PCA (Sompairac et al., 2019; Hyvärinen and Oja, 2000). Consequently, the order of decomposition affects all of the returned factors and therefore the number of components needs to be selected carefully. Horn's parallel analysis is a method for identifying the optimal number of components (Horn, 1965; Glorfeld, 1995; Crawford et al., 2010). Briefly, it simulates a random dataset of the same dimensions as the input data. The matrix factorisation method, in this case, ICA, is then run on both the simulated and actual data with the starting number of components to the maximum number of components. The steps above are repeated a large number of times to create a distribution of eigenvalues for both the simulated and actual data for each different number of components. These distributions are then compared and only factors with eigenvalues that are greater than the values found in stimulated data are kept. The optimal number of components can then be extracted. The reasoning behind this approach is that the eigenvalue which is larger than the resulting eigenvalue from the simulated data set is more likely to be the real underlying factor. As of writing there are no python packages available for this so a custom module and scikit-learn wrapper was implemented.

To identify potentially meaningful latent factors extracted from the ICA model several different methods were implemented. These included supervised methods where the target variable was taken into account and unsupervised methods where the information captured in the factor was used. There are 3 methods for supervised factor selection. Kolmogorov–Smirnov test, Wilcoxon rank-sum and Wilcoxon signed-rank test. Alternatively, in an unsupervised method, the Kurtosis test (Anscombe and Glynn, 1983) can be used to identify which factors captured the most amount of information. Each method is accessible from a custom python module which extends upon the Scipy (Virtanen et al., 2020) implementations of these statistical tests. These methods can all be used to perform tests between the factor distributions to determine which latent captured a signal which was most meaningful, in this case inactive or active IBD (represented by SCCAI or HBI for UC and CD respectively). However, depending on the downstream analysis, care should be taken when using supervised methods to avoid data leaks or biassing downstream models.

Finally, to evaluate the contributions of each factor, the loadings were extracted of the factors which contributed most to the target variable. A thresholding criterion of 2 standard deviations from the mean was applied to select the microbial features in the loadings matrix

to identify features that most contributed to that factor. As the loadings in ICA are an arbitrary value of the sum of contributions, their sign can be ignored. It should be noted that in this case the sign of the loadings is dependent on the input data, normalisation, transformations and scaling applied may not represent up or down-regulation as expected with differential abundance or expression analysis. Therefore, a greater interpretation of these top microbial features was then taken as the absolute value. The resulting weights that are given within the loadings are representative of the top contribution of that feature or that specific factor and are plotted in the form of a bar plot. This was all wrapped into a module as seen in Figure 4.2.

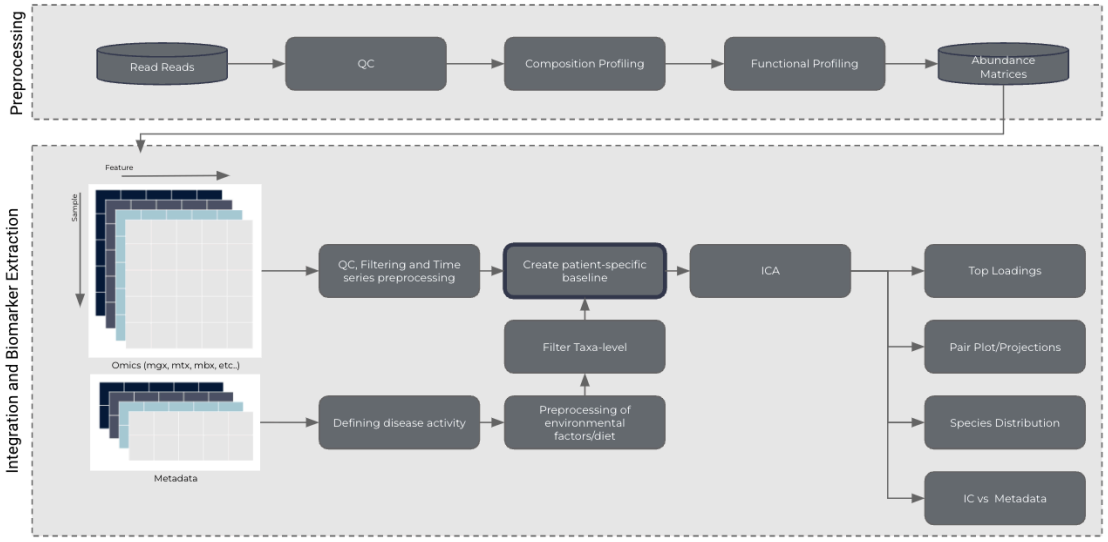


Figure 4.2. Overview of the ICA experimental design with microbes as sources accounting for patient-specific baseline. This framework can be used for any type of omics and for any binary (i.e. two unique class variables) metadata variables. The preprocessing stage can be easily switched to another preprocessing pipeline as this framework was developed in a modular fashion.

4.2.6 Pipeline architecture

The core pipeline developed in this chapter is a machine learning pipeline designed for analysing microbiome data, along with accompanying metadata such as disease conditions. It begins by addressing the high dimensionality nature of these microbiome datasets. To do this pipeline employs dimensionality reduction as its initial step, with Independent Component Analysis (ICA) set as the default method. This approach effectively uncovers latent variables within the microbiome data, which are then meticulously evaluated to

identify those that are most informative with respect to a specific metadata variable. Subsequently, the top loadings – the variables that contribute most significantly to each latent feature – are leveraged in a supervised analysis. This critical phase aims to assess and quantify the predictive power of each latent feature, offering valuable insights into the intricate relationships between the microbiome composition and the associated metadata, such as disease manifestations. This pipeline, therefore, serves as a robust tool for unravelling the complex interplay between microbiome characteristics and various biological and clinical outcomes. The resulting latents can then be used in a supervised analysis to determine how well the extracted latents predict that metadata. The pipeline is intended to be used both as an exploratory tool and as a feature extraction tool by bioinformatics or computational biologists working on microbiome data. It can be interfaced with using a command line interface.

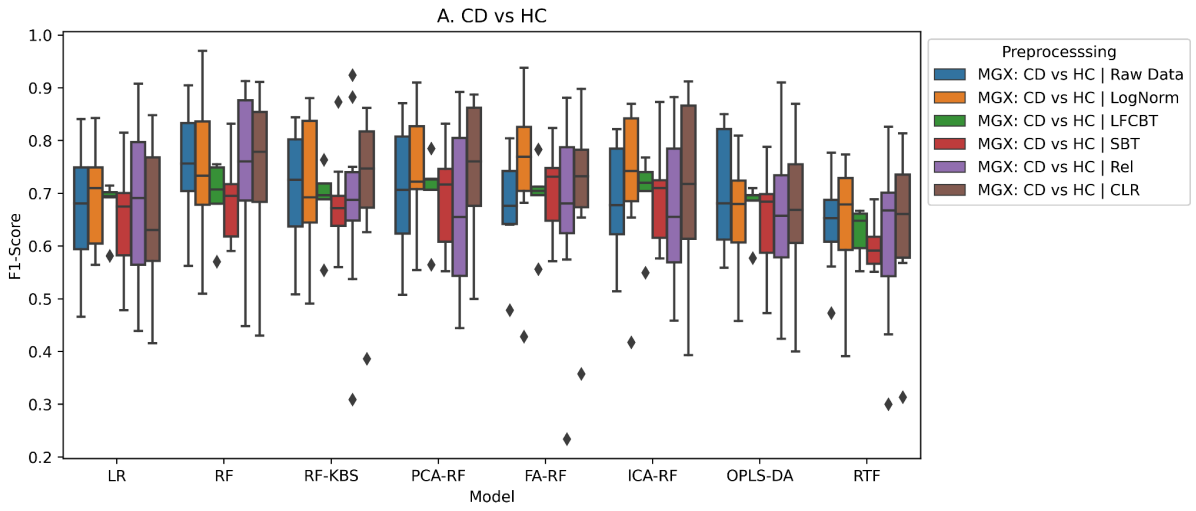
4.3 Results

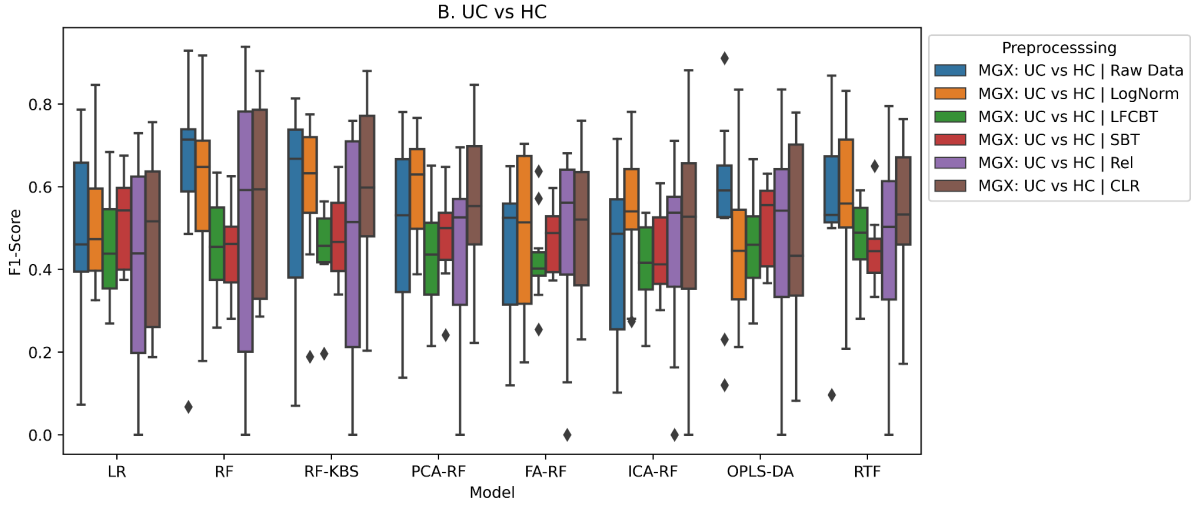
4.3.1 Metagenomics analysis of IBD vs Healthy controls

To explore the difference between the underlying microbiome profiles between IBD and healthy control samples, each of the matrix factorisation approaches was run on the data after different normalisation and transformation stages. For the normalisation stages, the data was either raw taxonomic count data, log normalised, or relative abundance normalisation. Then, for the transformation stage, the data were log-transformed, standardised to a unit-variance, or centre-log transformed. Ultimately the data for ILR transformation was not used due to the loss of interpretability. Finally, to assess the longitudinal nature of the data, both longitudinal patient-baseline transformation, FCBT and SBT, were also applied to the raw counts and the relative abundance of normalised data.

Using Horn's parallel analysis after 1000 iterations, each model's optimal number of components was determined out of a search range of 1 - the total number of features in the input space. The optimal number of components for relative abundance and CLR transformed data. To run this analysis on multiple different datasets is a very computationally expensive approach, and interestingly using this method the number of factors required for each method (i.e. PCA, FA and ICA). For a computational time, this was then used with just PCA (Supplementary Figure 4.1-4.3).

Each normalisation, transformation and a resulting component of the matrix factorisation methods were then used to classify between IBD disease types and health controls. This was done using leave-one group out cross validation (LOGOCV). , LOGOCV was used to avoid any data leaks resulting in the model learning a patient-specific microbiome and, therefore, a misleading performance metric. Using the F1-score and Briers score, each model's performance was ranked. When comparing between methods using critical distance (Supplementary Figure 4.4), no models or methods were statistically significant. The best-performing model was when comparing CD and non-IBD, which was CLR, with RF having the highest F1-score (0.749) and lowest Bier score (0.23). The worst-performing model was UC vs non-IBD. For each method's results, see Figure 4.3 and Figure 4.4.


A. CD vs HC
F1-Score
Model
LR
RF
RF-KBS
PCA-RF
FA-RF
ICA-RF
OPLS-DA
RTF
Preprocesssing
MGX: CD vs HC | Raw Data
MGX: CD vs HC | LogNorm
MGX: CD vs HC | LFCBT
MGX: CD vs HC | SBT
MGX: CD vs HC | Rel
MGX: CD vs HC | CLR


B. UC vs HC
F1-Score
Model
LR
RF
RF-KBS
PCA-RF
FA-RF
ICA-RF
OPLS-DA
RTF
Preprocesssing
MGX: UC vs HC | Raw Data
MGX: UC vs HC | LogNorm
MGX: UC vs HC | LFCBT
MGX: UC vs HC | SBT
MGX: UC vs HC | Rel
MGX: UC vs HC | CLR

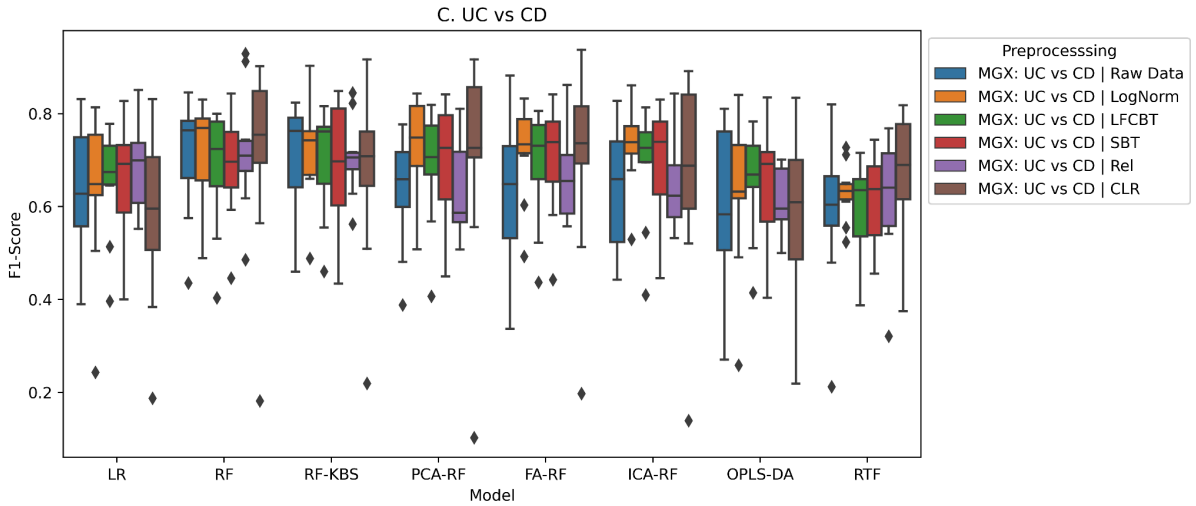
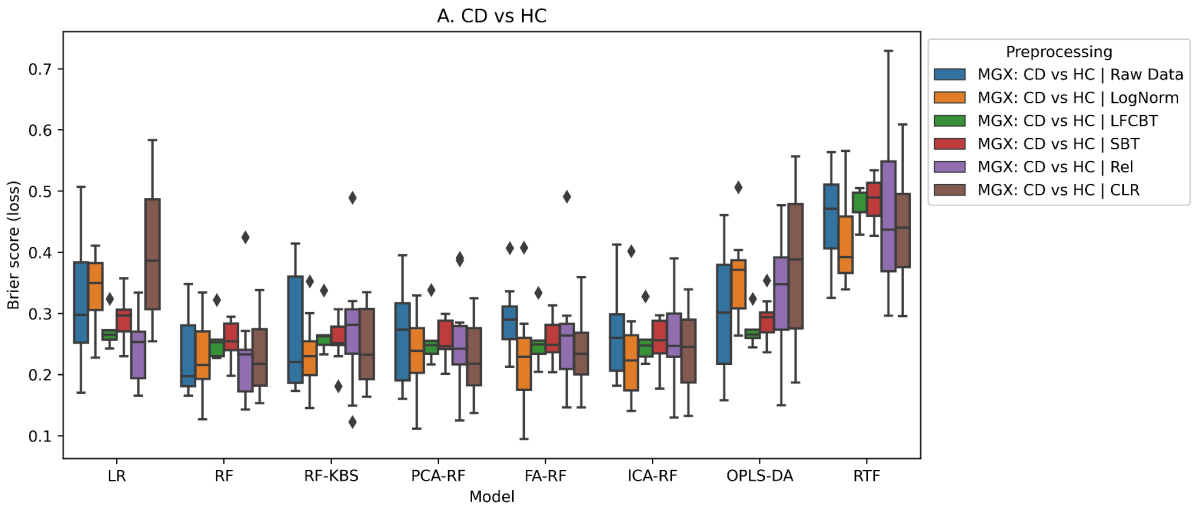
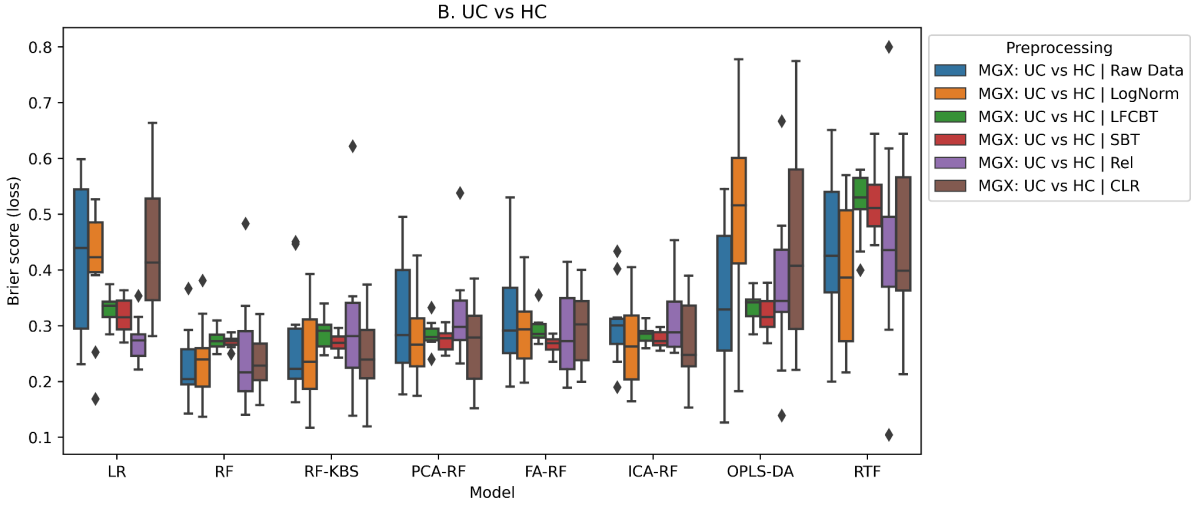

C. UC vs CD
F1-Score
Model
LR
RF
RF-KBS
PCA-RF
FA-RF
ICA-RF
OPLS-DA
RTF
Preprocesssing
MGX: UC vs CD | Raw Data
MGX: UC vs CD | LogNorm
MGX: UC vs CD | LFCBT
MGX: UC vs CD | SBT
MGX: UC vs CD | Rel
MGX: UC vs CD | CLR

Figure 4.3. Model evaluation for prediction of disease phenotypes between IBD patients and healthy controls from metagenomic data evaluated by their F1 score. Each experiment was run 100 times with ten splits with LOGOCV. The boxplot colours represent the normalisation and transformation method used on the data. The x-axis represents the model used. The higher the F1 score the more performant the model.


A. CD vs HC
Brier score (loss)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
LR
RF
RF-KBS
PCA-RF
FA-RF
ICA-RF
OPLS-DA
RTF
Model
Preprocessing
MGX: CD vs HC | Raw Data
MGX: CD vs HC | LogNorm
MGX: CD vs HC | LFCBT
MGX: CD vs HC | SBT
MGX: CD vs HC | Rel
MGX: CD vs HC | CLR


B. UC vs HC
Brier score (loss)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
LR
RF
RF-KBS
PCA-RF
FA-RF
ICA-RF
OPLS-DA
RTF
Model
Preprocessing
MGX: UC vs HC | Raw Data
MGX: UC vs HC | LogNorm
MGX: UC vs HC | LFCBT
MGX: UC vs HC | SBT
MGX: UC vs HC | Rel
MGX: UC vs HC | CLR

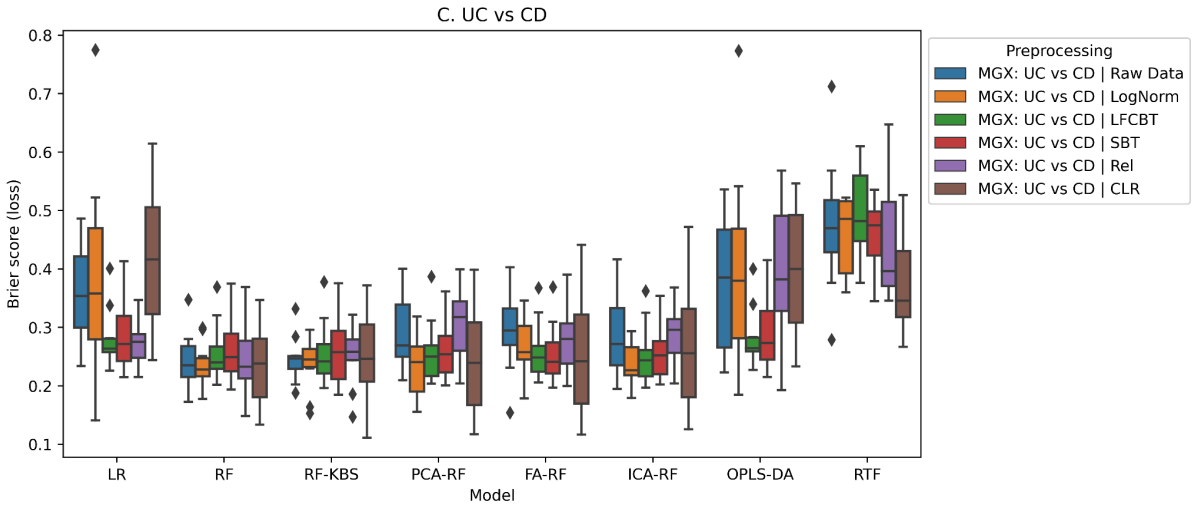
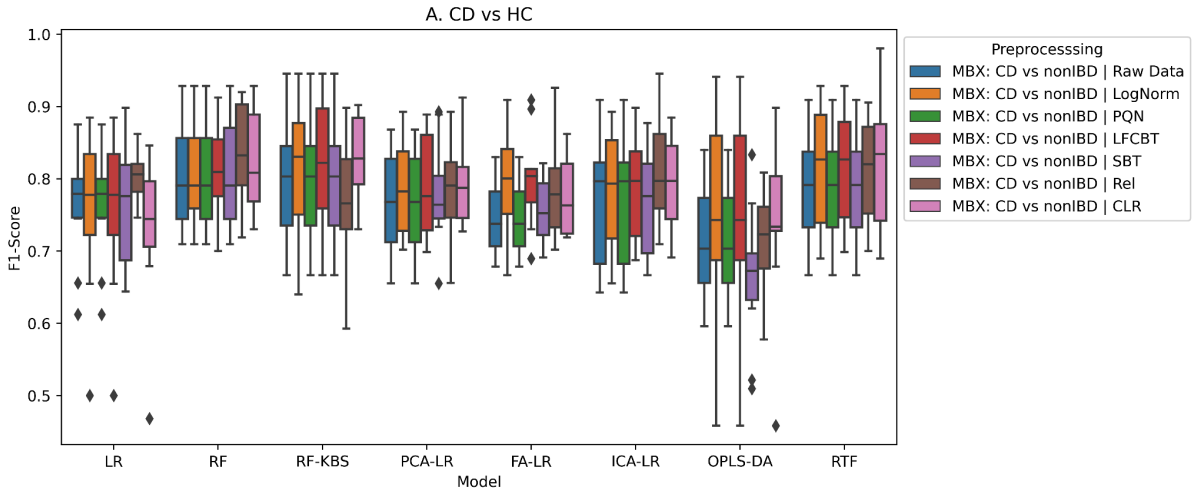

C. UC vs CD
Brier score (loss)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
LR
RF
RF-KBS
PCA-RF
FA-RF
ICA-RF
OPLS-DA
RTF
Model
Preprocessing
MGX: UC vs CD | Raw Data
MGX: UC vs CD | LogNorm
MGX: UC vs CD | LFCBT
MGX: UC vs CD | SBT
MGX: UC vs CD | Rel
MGX: UC vs CD | CLR

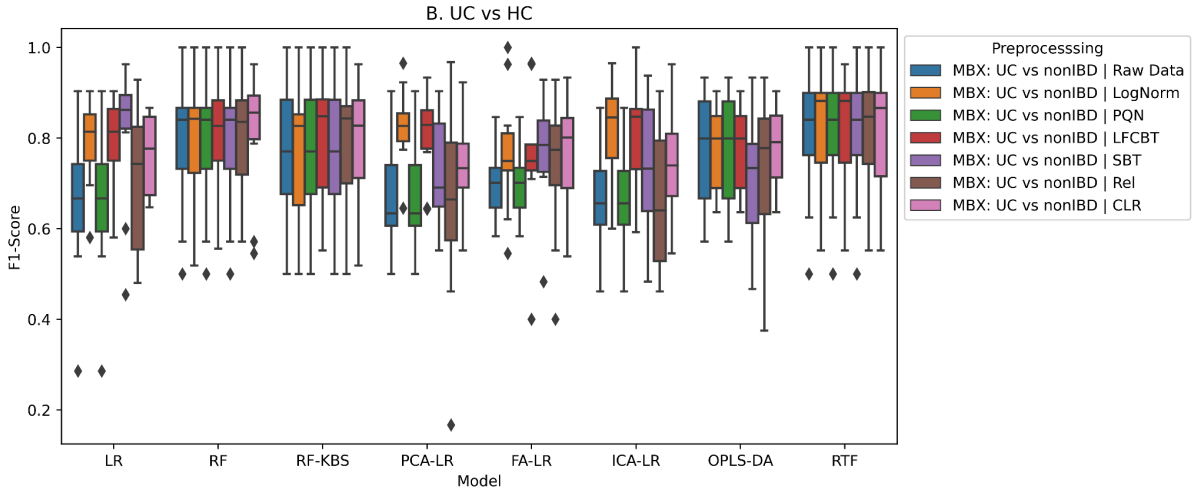
Figure 4.4. Model evaluation for prediction of disease phenotypes between IBD patients and healthy controls from metagenomic data evaluated by their Brier Score. Each experiment was run 100 times with ten splits with LOGOCV. The boxplot colours represent the normalisation and transformation method used on the data. The x-axis represents the model used. The lower the Brier score the more performant the model.

4.3.2 Metabolomics analysis of IBD vs Healthy controls

To explore the difference between the underlying metabolomic profile between the IBD and healthy control samples, the same framework was used with metagenomic data in section 4.2.1. However, due to the differences in metabolomic data, this time, the normalisation stages were, PQN normalised counts and relative abundance normalisation. Then for the transformation stage, the data were log-transformed, standardised to a unit-variance, or centre-log transformed (note for raw data and PQN normalisation, the data was both logs and standardised after normalisation). Once again, to assess the longitudinal nature of the data, both longitudinal patient-baseline transformation, FCBT and SBT, were also applied to the raw counts and to the relative abundance normalised data. Horn's parallel analysis was used again as described in section 4.2.1 and the resulting Scree plots can be seen in Supplementary Figures 4.5-7.

Compared to metagenomic data, metabolomic profiles enabled better stratification of patients between non-IBD and IBD. Again a critical distance analysis was used to compare each model's performance and the results suggested no significant difference after multiple testing. The highest-performing models in UC were the PCA with an RF (F1-score 0.826) and RTF (F1-score 0.820) after log normalisation and then were closely followed by relative abundance normalisation and FCBT. However, the Bier score showed good results from FCBT normalisation with RTF obtaining a score of 0.142. The same was true for CD but RF with KBS had the lowest Brier score and Relative abundance normalisation had the highest F1 score. However, all models seemed to perform equally well. The lowest-scoring models were LR and OPLS-DA.


A. CD vs HC
F1-Score
1.0
0.9
0.8
0.7
0.6
0.5
LR
RF
RF-KBS
PCA-LR
FA-LR
ICA-LR
OPLS-DA
RTF
Model
Preprocesssing
MBX: CD vs nonIBD | Raw Data
MBX: CD vs nonIBD | LogNorm
MBX: CD vs nonIBD | PQN
MBX: CD vs nonIBD | LFCBT
MBX: CD vs nonIBD | SBT
MBX: CD vs nonIBD | Rel
MBX: CD vs nonIBD | CLR


B. UC vs HC
F1-Score
1.0
0.8
0.6
0.4
0.2
LR
RF
RF-KBS
PCA-LR
FA-LR
ICA-LR
OPLS-DA
RTF
Model
Preprocesssing
MBX: UC vs nonIBD | Raw Data
MBX: UC vs nonIBD | LogNorm
MBX: UC vs nonIBD | PQN
MBX: UC vs nonIBD | LFCBT
MBX: UC vs nonIBD | SBT
MBX: UC vs nonIBD | Rel
MBX: UC vs nonIBD | CLR

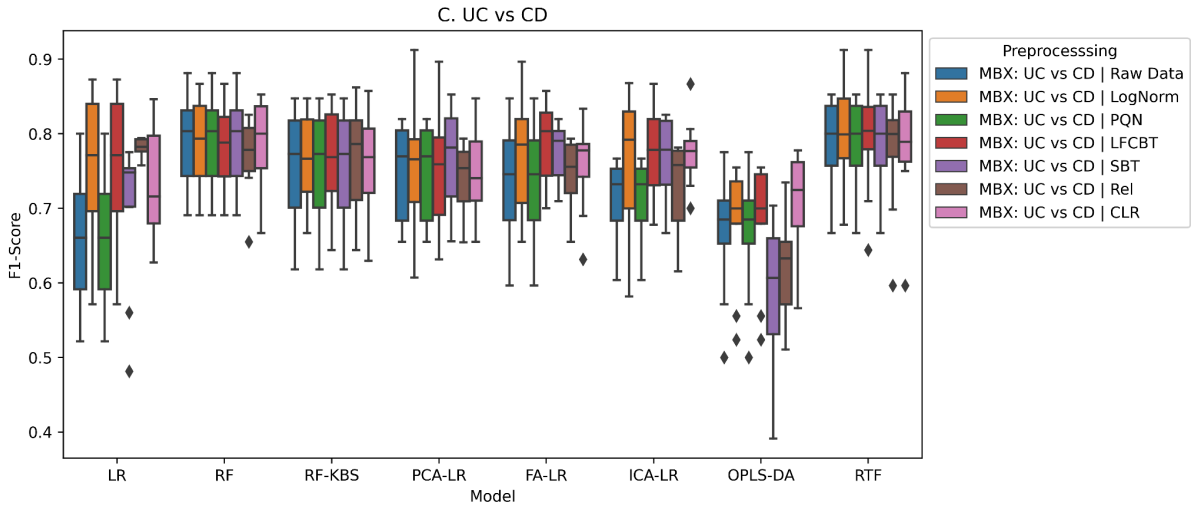
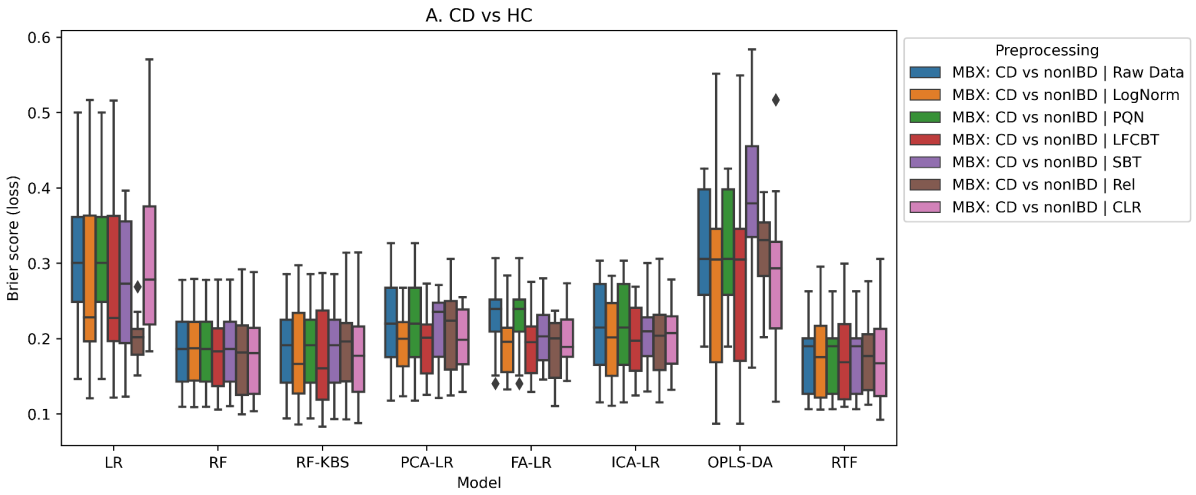
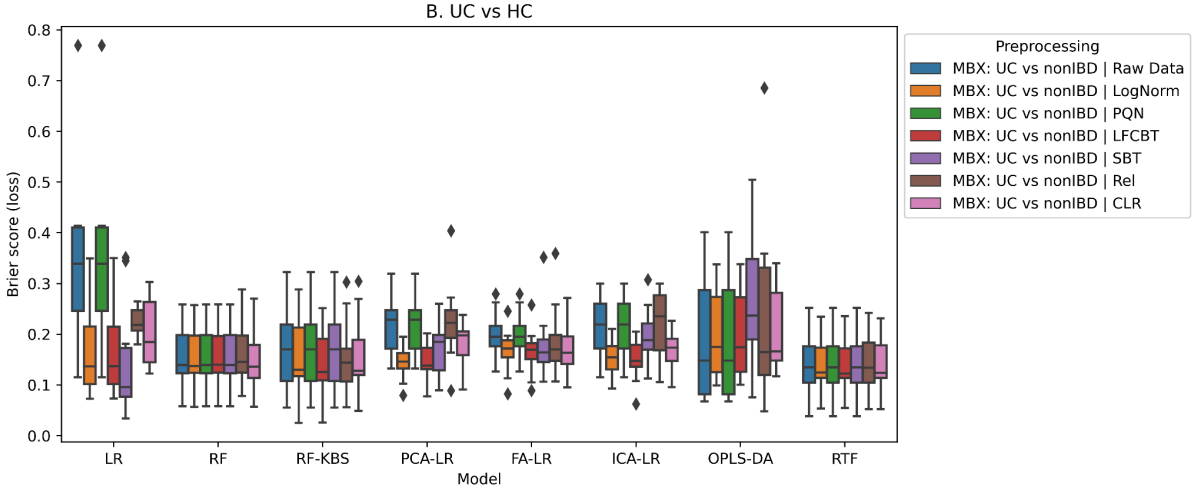

C. UC vs CD
F1-Score
0.9
0.8
0.7
0.6
0.5
0.4
LR
RF
RF-KBS
PCA-LR
FA-LR
ICA-LR
OPLS-DA
RTF
Model
Preprocesssing
MBX: UC vs CD | Raw Data
MBX: UC vs CD | LogNorm
MBX: UC vs CD | PQN
MBX: UC vs CD | LFCBT
MBX: UC vs CD | SBT
MBX: UC vs CD | Rel
MBX: UC vs CD | CLR

Figure 4.5. (Previous page) Model evaluation for prediction of disease phenotypes between IBD patients and healthy controls from metabolomic data evaluated by their F1 score. Each experiment was run 100 times with ten splits with LOGOCV. The boxplot colours represent the normalisation and transformation method used on the data. The x-axis represents the model used. The higher the F1-score, the more performant the model.


A. CD vs HC
Brier score (loss)
0.6
0.5
0.4
0.3
0.2
0.1
LR
RF
RF-KBS
PCA-LR
FA-LR
ICA-LR
OPLS-DA
RTF
Model
Preprocessing
MBX: CD vs nonIBD | Raw Data
MBX: CD vs nonIBD | LogNorm
MBX: CD vs nonIBD | PQN
MBX: CD vs nonIBD | LFCBT
MBX: CD vs nonIBD | SBT
MBX: CD vs nonIBD | Rel
MBX: CD vs nonIBD | CLR


B. UC vs HC
Brier score (loss)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
LR
RF
RF-KBS
PCA-LR
FA-LR
ICA-LR
OPLS-DA
RTF
Model
Preprocessing
MBX: UC vs nonIBD | Raw Data
MBX: UC vs nonIBD | LogNorm
MBX: UC vs nonIBD | PQN
MBX: UC vs nonIBD | LFCBT
MBX: UC vs nonIBD | SBT
MBX: UC vs nonIBD | Rel
MBX: UC vs nonIBD | CLR

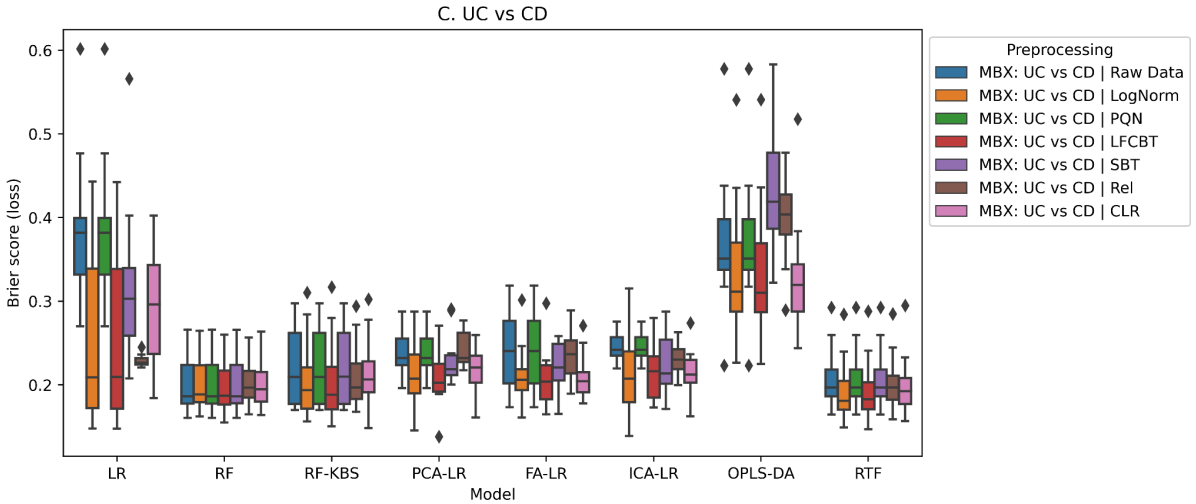

C. UC vs CD
Brier score (loss)
0.6
0.5
0.4
0.3
0.2
LR
RF
RF-KBS
PCA-LR
FA-LR
ICA-LR
OPLS-DA
RTF
Model
Preprocessing
MBX: UC vs CD | Raw Data
MBX: UC vs CD | LogNorm
MBX: UC vs CD | PQN
MBX: UC vs CD | LFCBT
MBX: UC vs CD | SBT
MBX: UC vs CD | Rel
MBX: UC vs CD | CLR

Figure 4.6. (Previous page) Model evaluation for prediction of disease phenotypes between IBD patients and healthy controls from metabolomic data evaluated by their F1 score. Each experiment was run 100 times with ten splits with GKFCV. The boxplot colours represent the normalisation and transformation method used on the data. The x-axis represents the models used. The higher the F1-score the more performant the model.

4.3.3 Unsupervised analysis of metabolomics in IBD vs healthy controls

By transposing the matrix to make the metabolite ICA was able to recover distinct signals between non-IBD and IBD patients. The best-performing normalisation and transformation were the FCBT. This analysis was applied to both UC vs non-IBD and CD vs non-IBD however, as this method requires all of the data and cannot subsequently fix later it was not used in the classification model.

Compared to other methods FCBT with a pseudo count of 1 before the log transformation was able to recover the IBD and non-IBD signals from the data. Other methods were also able to do this, like CLR; however, these had much larger tails suggesting influence from patient or environmental sources or large differences in the values before being fit by ICA. In addition to this FCBT accounts for the patient-specific baseline so can better represent the longitudinal aspect of the data.

The resulting components that captured the most information as ranked by their Kurtios value were then plotted against each other, and their signals from the loadings were extracted as described in 4.2.6. In UC, there was a distinct cluster separation between the phenotypes (Figures 4.7 and 4.8). Between UC and non-IBD both Carnitine (IC18), Bile acids (IC2), amino acids in (IC1) Long-chain fatty acids (IC13) and triacylglycerols (TAGs) dominated most of the ICs (Figure 4.9 and 4.10). The same separation is seen in CD (Figures 4.11 and 4.12). While CD sees even more enrichment of Bile acids (IC2, IC5) (Figure 4.13 and 4.14). Compared to UC, CD also has a large amount of enrichment in triacylglycerols (TAGs) (IC12, IC3) (Figure 4.13 and 4.14).

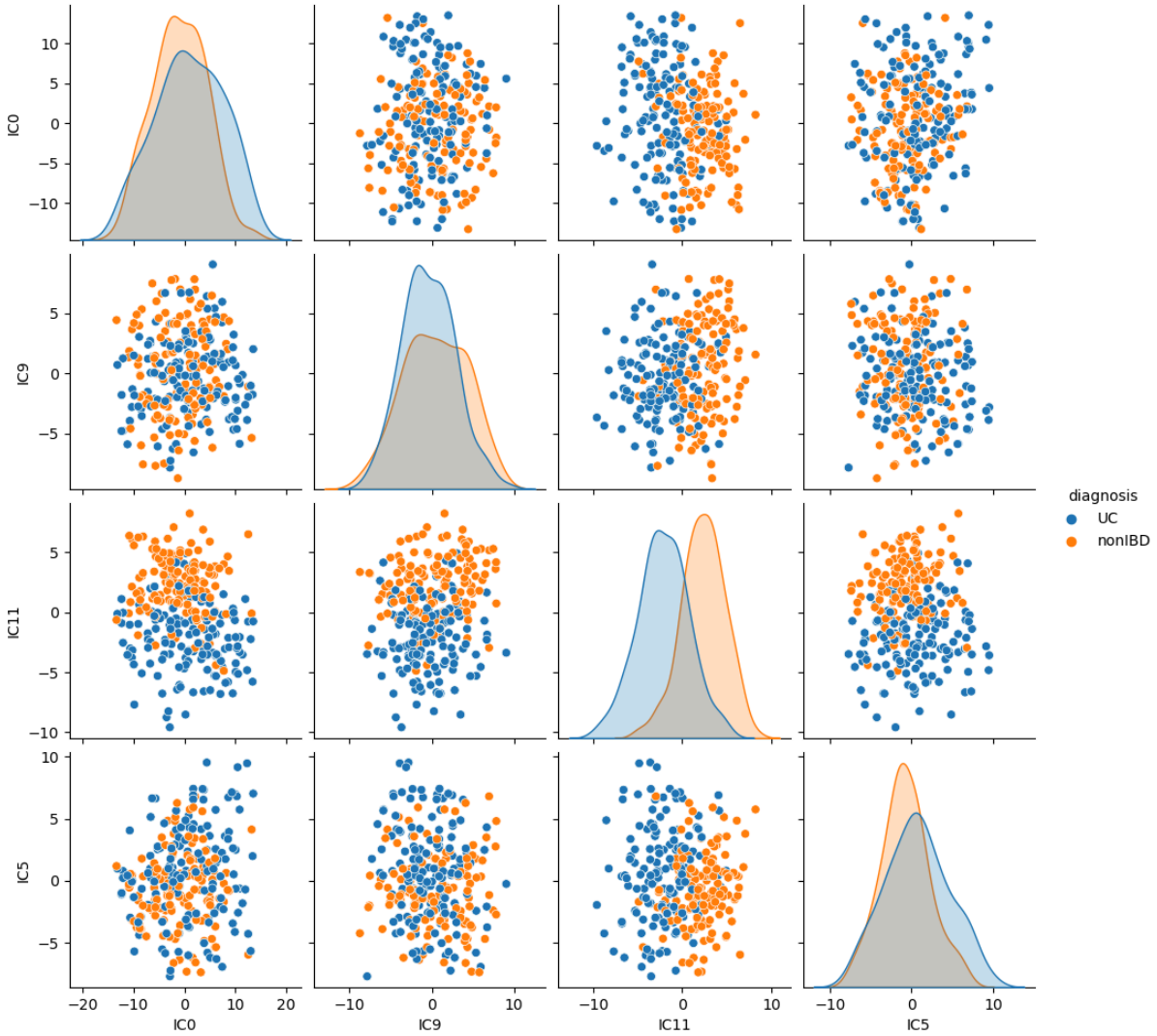


Figure 4.7. Overview of the UC vs healthy controls using ICA with microbes as sources accounting for patient-specific baseline. Each component is ranked by Kurtosis value, and the distributions are split by the target variable. There are several components which begin to show a UC and non-IBD signal difference. The top left-hand corner shows the factor which captures the most information. The total number of components for this model was selected as 18 via Horn's parallel analysis. (Blue; UC and Orange; Healthy Control)

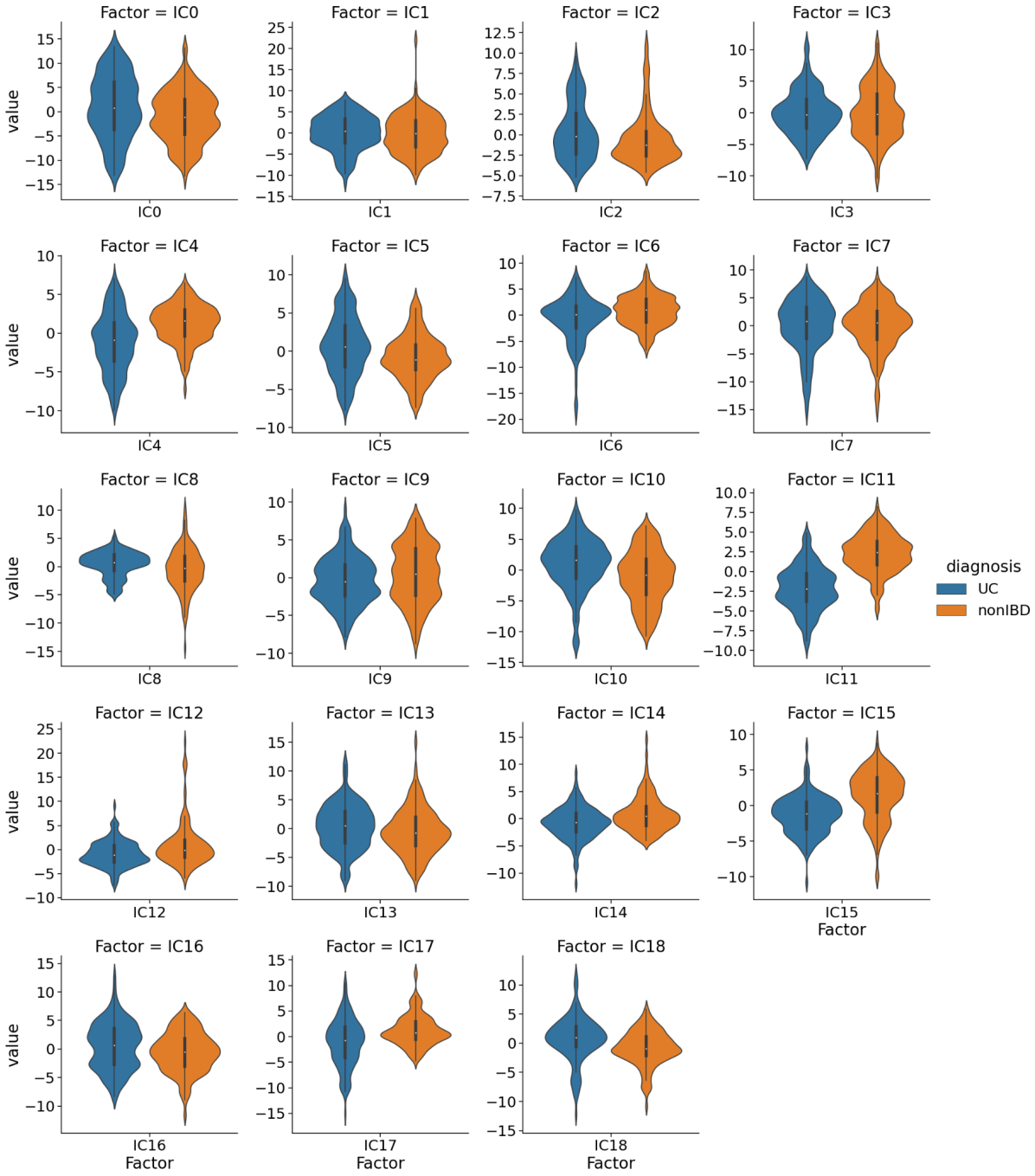


Figure 4.8. Top metabolites extracted from ICs capture a signal that can stratify samples between UC and healthy controls. The weights are thresholded by only selecting weights within the cutoff of 2 standard deviations from the mean. The absolute weights are taken to account for the arbitrary weight values.

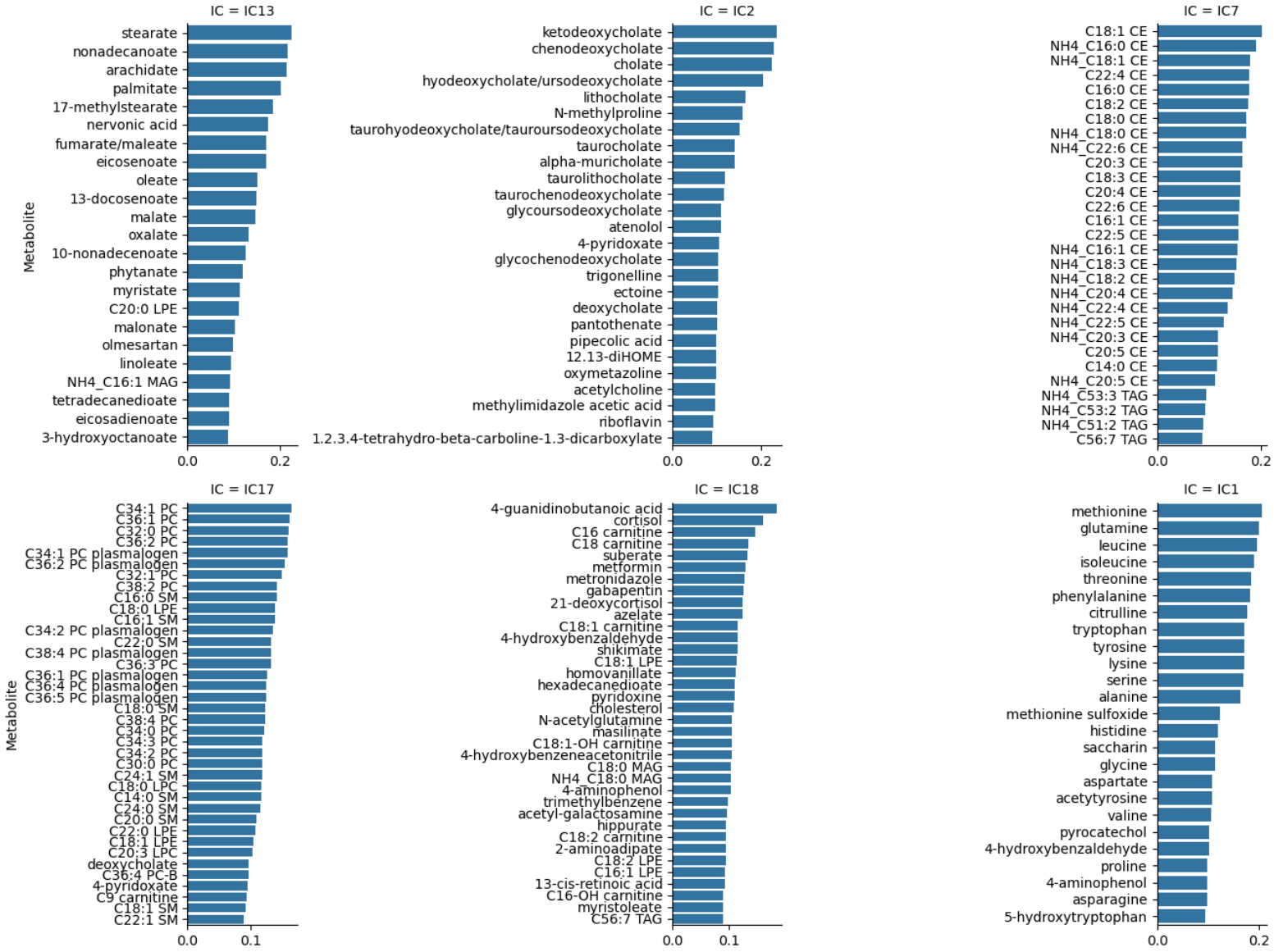
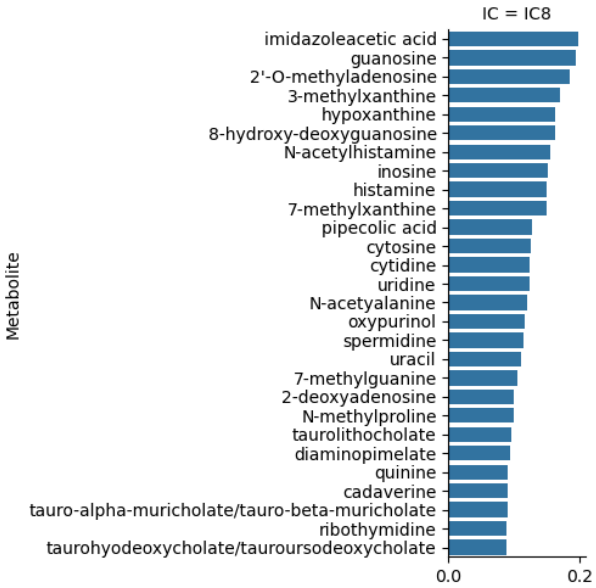
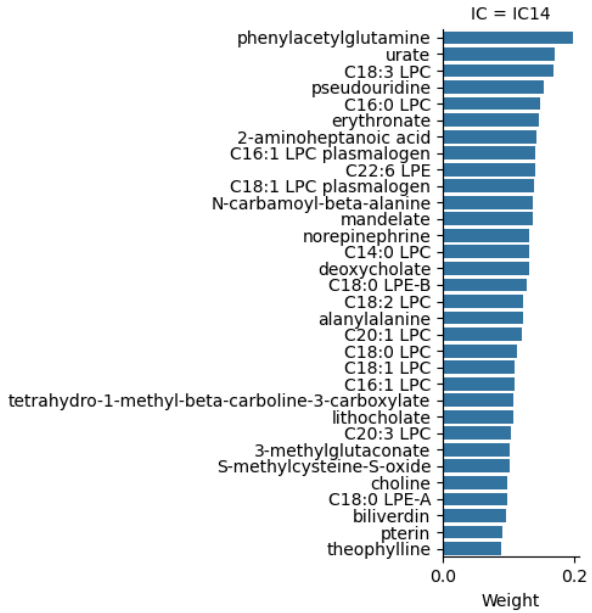
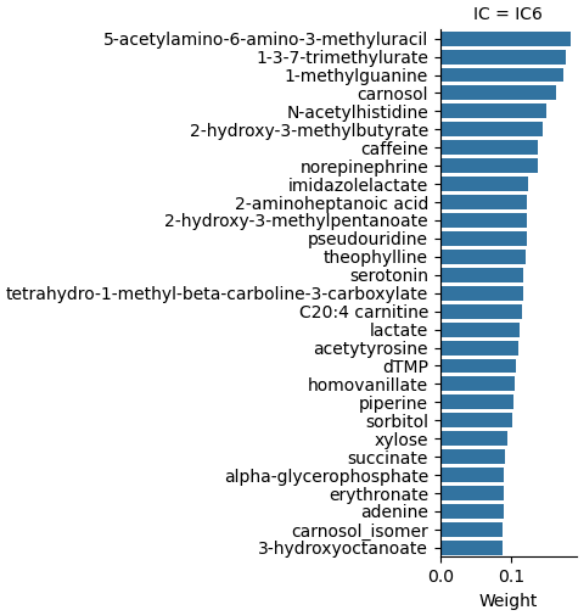

IC = IC13
Metabolite
stearate
nonadecanoate
arachidate
palmitate
17-methylstearate
nervonic acid
fumarate/maleate
eicosenoate
oleate
13-docosenoate
malate
oxalate
10-nonadecenoate
phytanate
myristate
C20:0 LPE
malonate
olmesartan
linoleate
NH4_C16:1 MAG
tetradecanedioate
eicosadienoate
3-hydroxyoctanoate
0.0
0.2
IC = IC2
ketodeoxycholate
chenodeoxycholate
cholate
hyodeoxycholate/ursodeoxycholate
lithocholate
N-methylproline
taurohyodeoxycholate/tauroursodeoxycholate
taurocholate
alpha-muricholate
taurolithocholate
taurochenodeoxycholate
glycoursodeoxycholate
atenolol
4-pyridoxate
glycochenodeoxycholate
trigonelline
ectoine
deoxycholate
pantothenate
pipecolic acid
12.13-diHOME
oxymetazoline
acetylcholine
methylimidazole acetic acid
riboflavin
1.2.3.4-tetrahydro-beta-carboline-1.3-dicarboxylate
0.0
0.2
IC = IC7
C18:1 CE
NH4_C16:0 CE
NH4_C18:1 CE
C22:4 CE
C16:0 CE
C18:2 CE
C18:0 CE
NH4_C18:0 CE
NH4_C22:6 CE
C20:3 CE
C18:3 CE
C20:4 CE
C22:6 CE
C16:1 CE
C22:5 CE
NH4_C16:1 CE
NH4_C18:3 CE
NH4_C18:2 CE
NH4_C20:4 CE
NH4_C22:4 CE
NH4_C22:5 CE
NH4_C20:3 CE
C20:5 CE
C14:0 CE
NH4_C20:5 CE
NH4_C53:3 TAG
NH4_C53:2 TAG
NH4_C51:2 TAG
C56:7 TAG
0.0
0.2
IC = IC17
Metabolite
C34:1 PC
C36:1 PC
C32:0 PC
C36:2 PC
C34:1 PC plasmalogen
C36:2 PC plasmalogen
C32:1 PC
C38:2 PC
C16:0 SM
C18:0 LPE
C16:1 SM
C34:2 PC plasmalogen
C22:0 SM
C38:4 PC plasmalogen
C36:3 PC
C36:1 PC plasmalogen
C36:4 PC plasmalogen
C36:5 PC plasmalogen
C18:0 SM
C38:4 PC
C34:0 PC
C34:3 PC
C34:2 PC
C30:0 PC
C24:1 SM
C18:0 LPC
C14:0 SM
C24:0 SM
C20:0 SM
C22:0 LPE
C18:1 LPE
C20:3 LPC
deoxycholate
C36:4 PC-B
4-pyridoxate
C9 carnitine
C18:1 SM
C22:1 SM
0.0
0.1
IC = IC18
4-guanidinobutanoic acid
cortisol
C16 carnitine
C18 carnitine
suberate
metformin
metronidazole
gabapentin
21-deoxycortisol
azelate
C18:1 carnitine
4-hydroxybenzaldehyde
shikimate
C18:1 LPE
homovanillate
hexadecanedioate
pyridoxine
cholesterol
N-acetylglutamine
masilinate
C18:1-OH carnitine
4-hydroxybenzeneacetonitrile
C18:0 MAG
NH4_C18:0 MAG
4-aminophenol
trimethylbenzene
acetyl-galactosamine
hippurate
C18:2 carnitine
2-aminoadipate
C18:2 LPE
C16:1 LPE
13-cis-retinoic acid
C16-OH carnitine
myristoleate
C56:7 TAG
0.0
0.1
IC = IC1
methionine
glutamine
leucine
isoleucine
threonine
phenylalanine
citrulline
tryptophan
tyrosine
lysine
serine
alanine
methionine sulfoxide
histidine
saccharin
glycine
aspartate
acetytyrosine
valine
pyrocatechol
4-hydroxybenzaldehyde
proline
4-aminophenol
asparagine
5-hydroxytryptophan
0.0
0.2

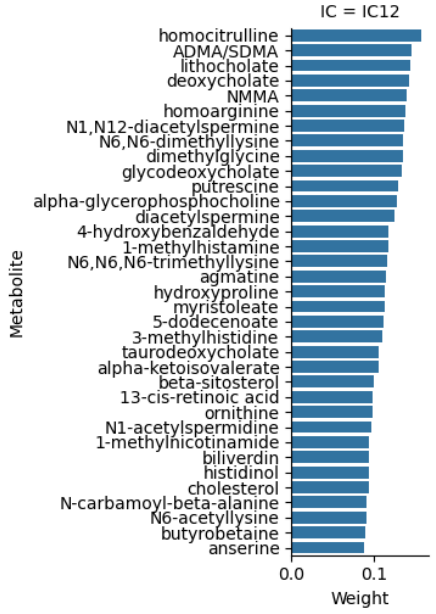
Figure 4.9. Top metabolites extracted from ICs capture a signal that can stratify samples between UC and healthy controls. The weights are thresholded by only selecting weights within the cutoff of 2 standard deviations from the mean. The absolute weights are taken to account for the arbitrary weight values.

Figure 4.10. Top metabolites extracted from ICs capture a signal that can stratify samples between UC and healthy controls continued… (Next page). The weights are thresholded by only selecting weights within the cutoff of 2 standard deviations from the mean. The absolute weights are taken to account for the arbitrary weight values.


IC = IC8
Metabolite
imidazoleacetic acid
guanosine
2'-O-methyladenosine
3-methylxanthine
hypoxanthine
8-hydroxy-deoxyguanosine
N-acetylhistamine
inosine
histamine
7-methylxanthine
pipecolic acid
cytosine
cytidine
uridine
N-acetyalanine
oxypurinol
spermidine
uracil
7-methylguanine
2-deoxyadenosine
N-methylproline
taurolithocholate
diaminopimelate
quinine
cadaverine
tauro-alpha-muricholate/tauro-beta-muricholate
ribothymidine
taurohyodeoxycholate/tauroursodeoxycholate
0.0
0.2


IC = IC14
phenylacetylglutamine
urate
C18:3 LPC
pseudouridine
C16:0 LPC
erythronate
2-aminoheptanoic acid
C16:1 LPC plasmalogen
C22:6 LPE
C18:1 LPC plasmalogen
N-carbamoyl-beta-alanine
mandelate
norepinephrine
C14:0 LPC
deoxycholate
C18:0 LPE-B
C18:2 LPC
alanylalanine
C20:1 LPC
C18:0 LPC
C18:1 LPC
C16:1 LPC
tetrahydro-1-methyl-beta-carboline-3-carboxylate
lithocholate
C20:3 LPC
3-methylglutaconate
S-methylcysteine-S-oxide
choline
C18:0 LPE-A
biliverdin
pterin
theophylline
0.0
0.2
Weight


IC = IC6
5-acetylamino-6-amino-3-methyluracil
1-3-7-trimethylurate
1-methylguanine
carnosol
N-acetylhistidine
2-hydroxy-3-methylbutyrate
caffeine
norepinephrine
imidazolelactate
2-aminoheptanoic acid
2-hydroxy-3-methylpentanoate
pseudouridine
theophylline
serotonin
tetrahydro-1-methyl-beta-carboline-3-carboxylate
C20:4 carnitine
lactate
acetytyrosine
dTMP
homovanillate
piperine
sorbitol
xylose
succinate
alpha-glycerophosphate
erythronate
adenine
carnosol_isomer
3-hydroxyoctanoate
0.0
0.1
Weight


IC = IC12
Metabolite
homocitrulline
ADMA/SDMA
lithocholate
deoxycholate
NMMA
homoarginine
N1,N12-diacetylspermine
N6,N6-dimethyllysine
dimethylglycine
glycodeoxycholate
putrescine
alpha-glycerophosphocholine
diacetylspermine
4-hydroxybenzaldehyde
1-methylhistamine
N6,N6,N6-trimethyllysine
agmatine
hydroxyproline
myristoleate
5-dodecenoate
3-methylhistidine
taurodeoxycholate
alpha-ketoisovalerate
beta-sitosterol
13-cis-retinoic acid
ornithine
N1-acetylspermidine
1-methylnicotinamide
biliverdin
histidinol
cholesterol
N-carbamoyl-beta-alanine
N6-acetyllysine
butyrobetaine
anserine
0.0
0.1
Weight

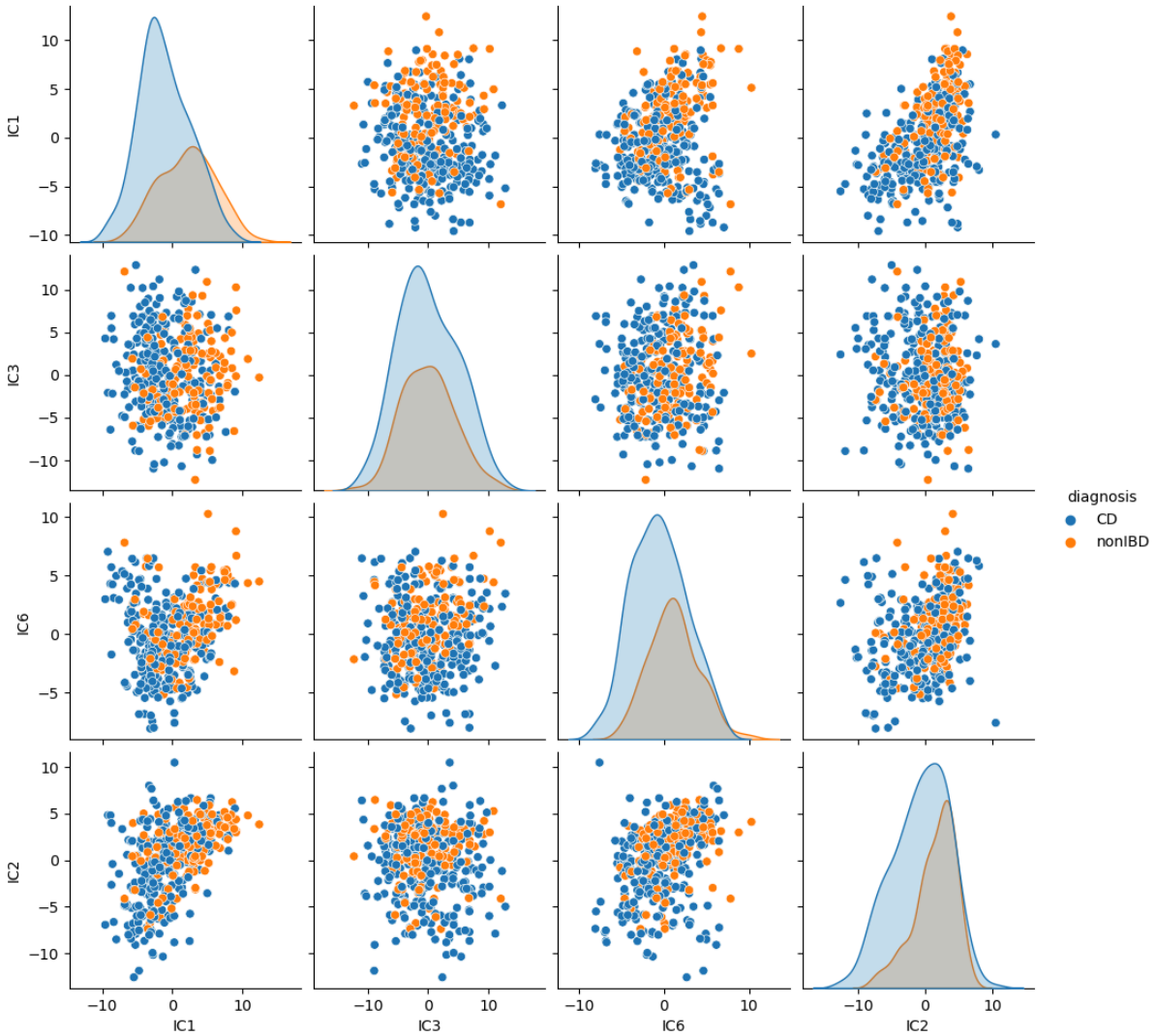


Figure 4.11. Overview of the CD vs healthy controls using ICA with microbes as sources accounting for patient-specific baseline. Each component is ranked by Kurtosis value, and the distributions are split by the target variable. There are several components which begin to show a UC and non-IBD signal difference. The top left-hand corner shows the factor which captures the most information. The total number of components for this model was selected as 18 via Horn's parallel analysis. (Blue; CD and Orange; Healthy Control)

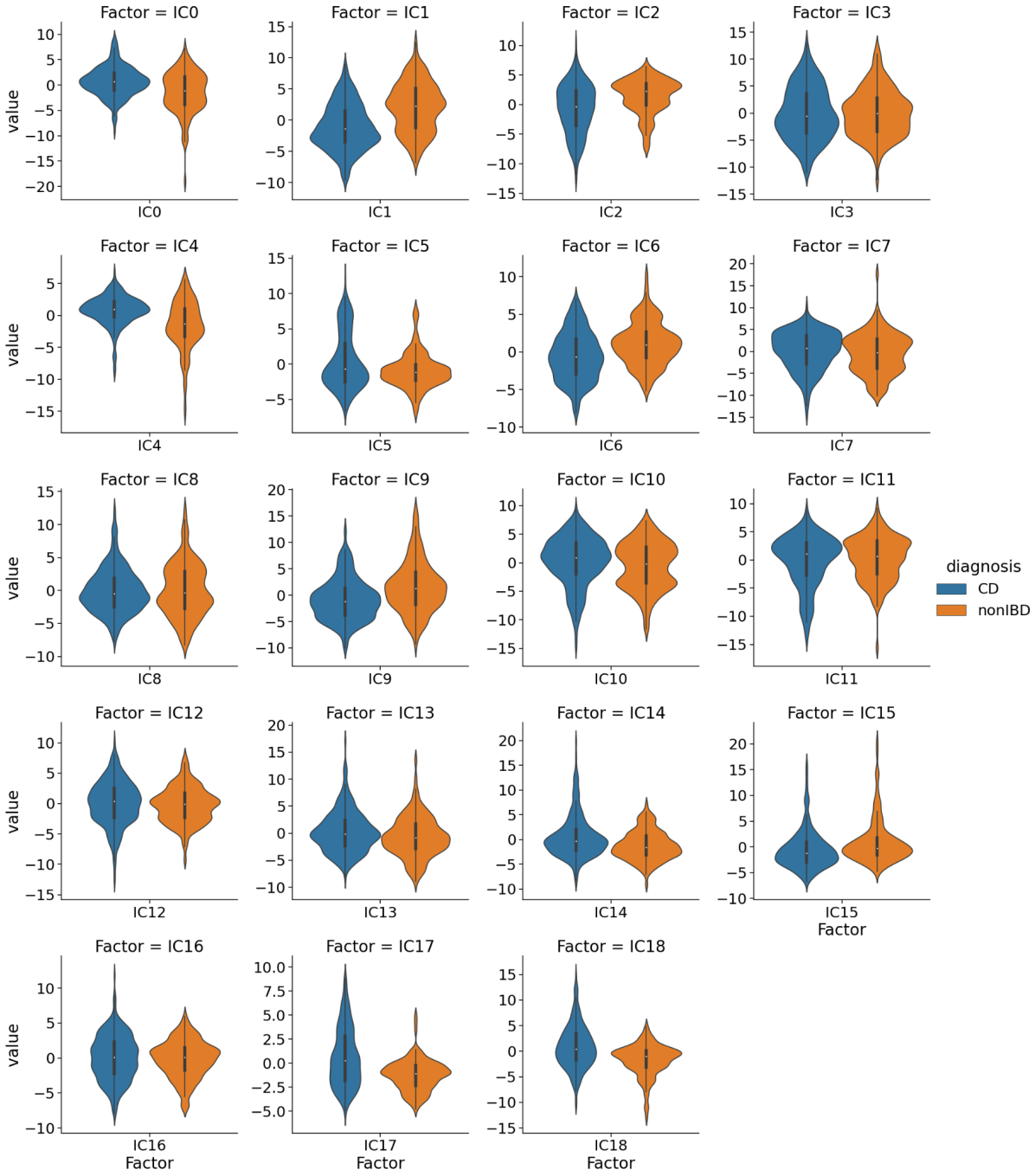
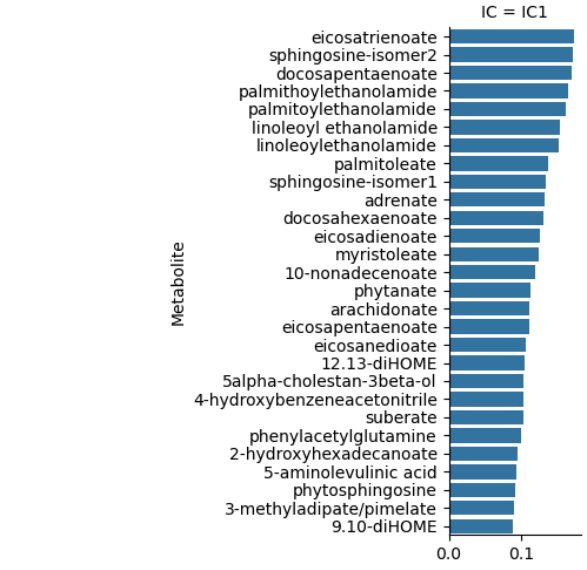
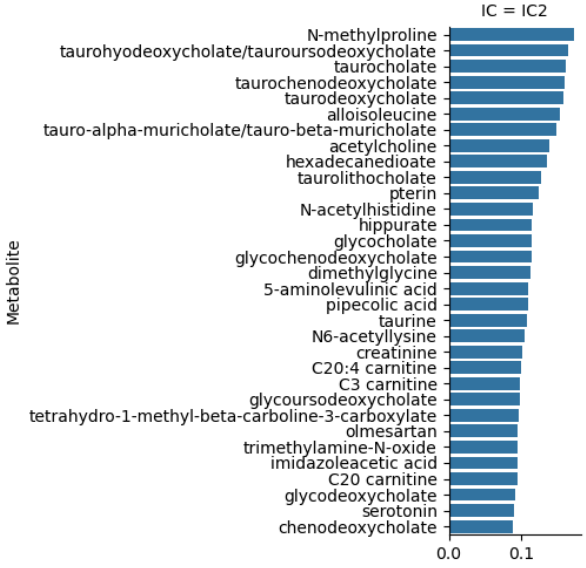
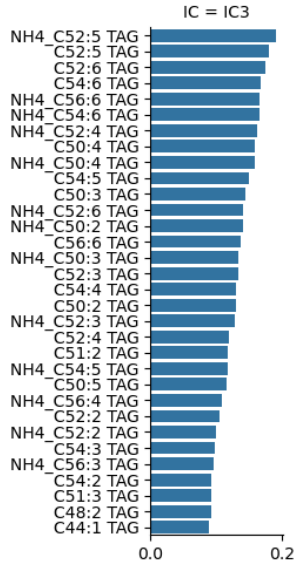
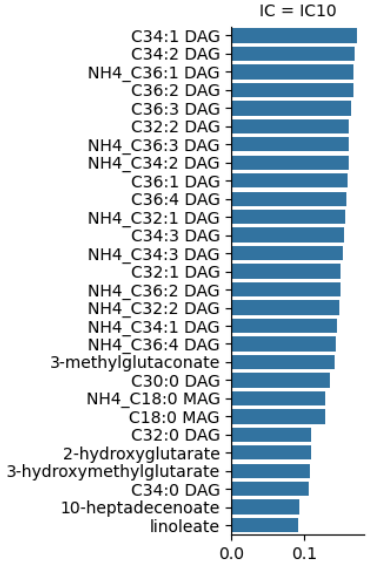


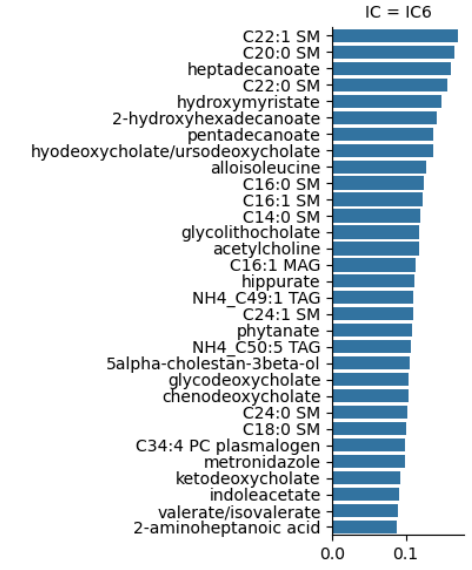
Figure 4.12. Top metabolites extracted from ICs capture a signal that can stratify samples between CD and healthy controls. The weights are thresholded by only selecting weights within the cutoff of 2 standard deviations from the mean. The absolute weights are taken to account for the arbitrary weight values.


IC = IC1
Metabolite
eicosatrienoate
sphingosine-isomer2
docosapentaenoate
palmithoylethanolamide
palmitoylethanolamide
linoleoyl ethanolamide
linoleoylethanolamide
palmitoleate
sphingosine-isomer1
adrenate
docosahexaenoate
eicosadienoate
myristoleate
10-nonadecenoate
phytanate
arachidonate
eicosapentaenoate
eicosanedioate
12.13-diHOME
5alpha-cholestan-3beta-ol
4-hydroxybenzeneacetonitrile
suberate
phenylacetylglutamine
2-hydroxyhexadecanoate
5-aminolevulinic acid
phytosphingosine
3-methyladipate/pimelate
9.10-diHOME
0.0
0.1


IC = IC2
Metabolite
N-methylproline
taurohyodeoxycholate/tauroursodeoxycholate
taurocholate
taurochenodeoxycholate
taurodeoxycholate
alloisoleucine
tauro-alpha-muricholate/tauro-beta-muricholate
acetylcholine
hexadecanedioate
taurolithocholate
pterin
N-acetylhistidine
hippurate
glycocholate
glycochenodeoxycholate
dimethylglycine
5-aminolevulinic acid
pipecolic acid
taurine
N6-acetyllysine
creatinine
C20:4 carnitine
C3 carnitine
glycoursodeoxycholate
tetrahydro-1-methyl-beta-carboline-3-carboxylate
olmesartan
trimethylamine-N-oxide
imidazoleacetic acid
C20 carnitine
glycodeoxycholate
serotonin
chenodeoxycholate
0.0
0.1


IC = IC3
NH4_C52:5 TAG
C52:5 TAG
C52:6 TAG
C54:6 TAG
NH4_C56:6 TAG
NH4_C54:6 TAG
NH4_C52:4 TAG
C50:4 TAG
NH4_C50:4 TAG
C54:5 TAG
C50:3 TAG
NH4_C52:6 TAG
NH4_C50:2 TAG
C56:6 TAG
NH4_C50:3 TAG
C52:3 TAG
C54:4 TAG
C50:2 TAG
NH4_C52:3 TAG
C52:4 TAG
C51:2 TAG
NH4_C54:5 TAG
C50:5 TAG
NH4_C56:4 TAG
C52:2 TAG
NH4_C52:2 TAG
C54:3 TAG
NH4_C56:3 TAG
C54:2 TAG
C51:3 TAG
C48:2 TAG
C44:1 TAG
0.0
0.2


IC = IC10
C34:1 DAG
C34:2 DAG
NH4_C36:1 DAG
C36:2 DAG
C36:3 DAG
C32:2 DAG
NH4_C36:3 DAG
NH4_C34:2 DAG
C36:1 DAG
C36:4 DAG
NH4_C32:1 DAG
C34:3 DAG
NH4_C34:3 DAG
C32:1 DAG
NH4_C36:2 DAG
NH4_C32:2 DAG
NH4_C34:1 DAG
NH4_C36:4 DAG
3-methylglutaconate
C30:0 DAG
NH4_C18:0 MAG
C18:0 MAG
C32:0 DAG
2-hydroxyglutarate
3-hydroxymethylglutarate
C34:0 DAG
10-heptadecenoate
linoleate
0.0
0.1


IC = IC6
C22:1 SM
C20:0 SM
heptadecanoate
C22:0 SM
hydroxymyristate
2-hydroxyhexadecanoate
pentadecanoate
hyodeoxycholate/ursodeoxycholate
alloisoleucine
C16:0 SM
C16:1 SM
C14:0 SM
glycolithocholate
acetylcholine
C16:1 MAG
hippurate
NH4_C49:1 TAG
C24:1 SM
phytanate
NH4_C50:5 TAG
5alpha-cholestan-3beta-ol
glycodeoxycholate
chenodeoxycholate
C24:0 SM
C18:0 SM
C34:4 PC plasmalogen
metronidazole
ketodeoxycholate
indoleacetate
valerate/isovalerate
2-aminoheptanoic acid
0.0
0.1

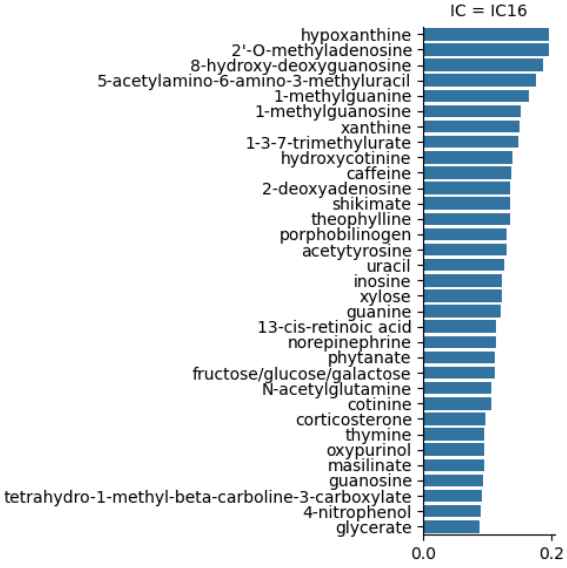
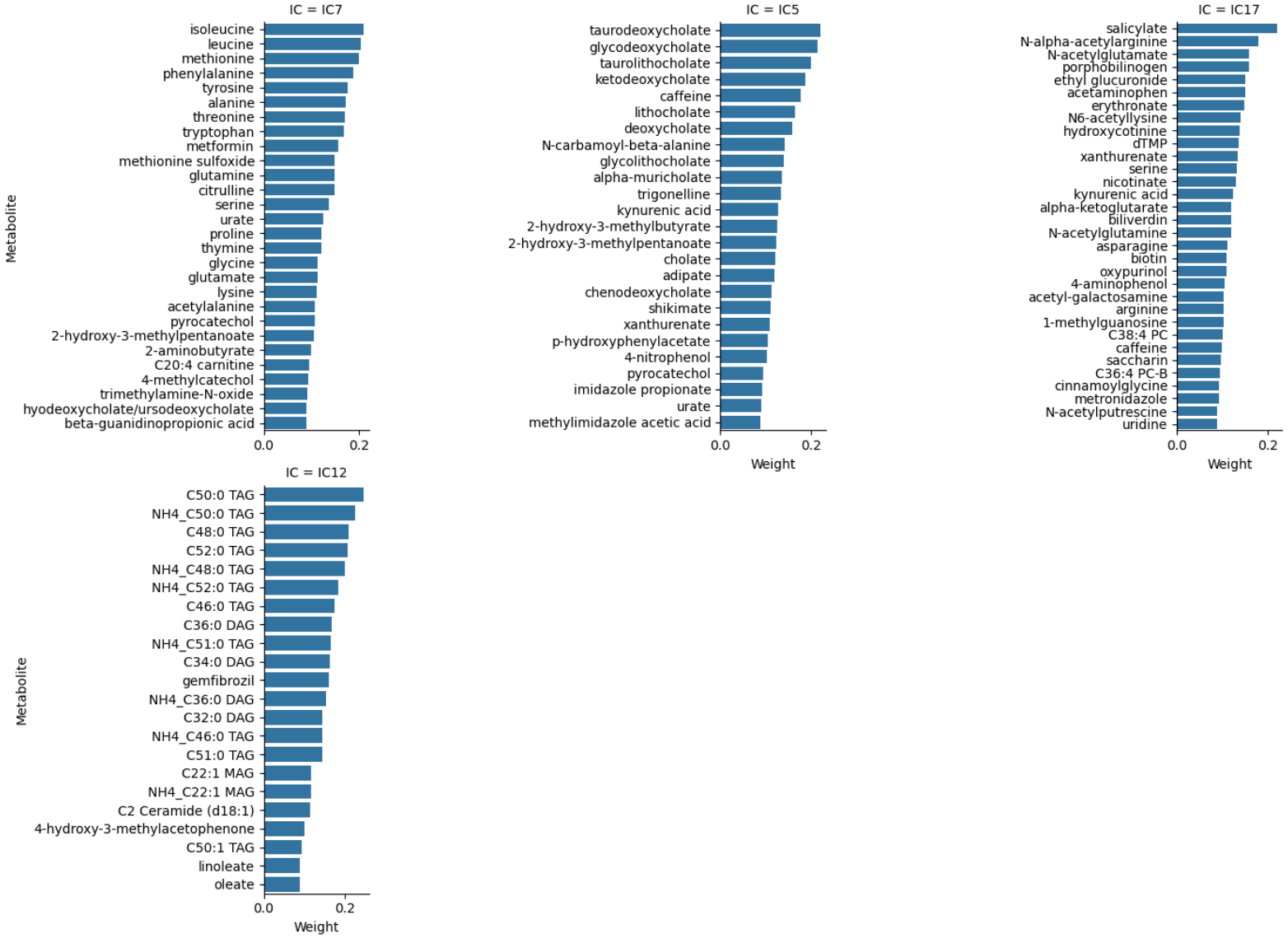

IC = IC16
hypoxanthine
2'-O-methyladenosine
8-hydroxy-deoxyguanosine
5-acetylamino-6-amino-3-methyluracil
1-methylguanine
1-methylguanosine
xanthine
1-3-7-trimethylurate
hydroxycotinine
caffeine
2-deoxyadenosine
shikimate
theophylline
porphobilinogen
acetytyrosine
uracil
inosine
xylose
guanine
13-cis-retinoic acid
norepinephrine
phytanate
fructose/glucose/galactose
N-acetylglutamine
cotinine
corticosterone
thymine
oxypurinol
masilinate
guanosine
tetrahydro-1-methyl-beta-carboline-3-carboxylate
4-nitrophenol
glycerate
0.0
0.2

Figure 4.13. Top metabolites extracted from ICs capture a signal that can stratify samples between CD and healthy controls (previous page). The weights are thresholded by only selecting weights within the cutoff of 2 standard deviations from the mean. The absolute weights are taken to account for the arbitrary weight values.

Figure 4.14. Top metabolites extracted from ICs capture a signal that can stratify samples between CD and healthy controls continued… (next page). The weights are thresholded by only selecting weights within the cutoff of 2 standard deviations from the mean. The absolute weights are taken to account for the arbitrary weight values.


IC = IC7
isoleucine
leucine
methionine
phenylalanine
tyrosine
alanine
threonine
tryptophan
metformin
methionine sulfoxide
glutamine
citrulline
serine
urate
proline
thymine
glycine
glutamate
lysine
acetylalanine
pyrocatechol
2-hydroxy-3-methylpentanoate
2-aminobutyrate
C20:4 carnitine
4-methylcatechol
trimethylamine-N-oxide
hyodeoxycholate/ursodeoxycholate
beta-guanidinopropionic acid
Metabolite
0.0
0.2
IC = IC5
taurodeoxycholate
glycodeoxycholate
taurolithocholate
ketodeoxycholate
caffeine
lithocholate
deoxycholate
N-carbamoyl-beta-alanine
glycolithocholate
alpha-muricholate
trigonelline
kynurenic acid
2-hydroxy-3-methylbutyrate
2-hydroxy-3-methylpentanoate
cholate
adipate
chenodeoxycholate
shikimate
xanthurenate
p-hydroxyphenylacetate
4-nitrophenol
pyrocatechol
imidazole propionate
urate
methylimidazole acetic acid
0.0
0.2
Weight
IC = IC17
salicylate
N-alpha-acetylarginine
N-acetylglutamate
porphobilinogen
ethyl glucuronide
acetaminophen
erythronate
N6-acetyllysine
hydroxycotinine
dTMP
xanthurenate
serine
nicotinate
kynurenic acid
alpha-ketoglutarate
biliverdin
N-acetylglutamine
asparagine
biotin
oxypurinol
4-aminophenol
acetyl-galactosamine
arginine
1-methylguanosine
C38:4 PC
caffeine
saccharin
C36:4 PC-B
cinnamoylglycine
metronidazole
N-acetylputrescine
uridine
0.0
0.2
Weight
IC = IC12
C50:0 TAG
NH4_C50:0 TAG
C48:0 TAG
C52:0 TAG
NH4_C48:0 TAG
NH4_C52:0 TAG
C46:0 TAG
C36:0 DAG
NH4_C51:0 TAG
C34:0 DAG
gemfibrozil
NH4_C36:0 DAG
C32:0 DAG
NH4_C46:0 TAG
C51:0 TAG
C22:1 MAG
NH4_C22:1 MAG
C2 Ceramide (d18:1)
4-hydroxy-3-methylacetophenone
C50:1 TAG
linoleate
oleate
Metabolite
0.0
0.2
Weight

4.4 Discussion

4.4.1 Metagenomic and Metabolomics prognostic indicator identification

Overall, this study showed how difficult it is to classify IBD using metagenomic data. Although the best performance was seen when predicting between CD and non-IBD controls, it should also be noted that this was with a small dataset (CD n=50, nonIBD n=20). This means the data is highly imbalanced, and when taking into account the temporal element, it results in an order of magnitude higher ratio of imbalance. Classifying between UC and non-IBD was overall very poor with most models displaying a wide confidence interval in both their F1-score and Bier scores (Figure 4.3-4.4). The top F1 score was between CD and non-IBD, achieving 0.761 and a ROC-AUC of 0.614 which was significantly worse than other studies achieved using longitudinal microbiome data (Clooney et al., 2021). However, it should be noted that for the study conducted by Clooney et al, the authors reported only doing standard KFold cross-validation, implying the same patient samples were in both the train and test datasets. If this is the case, this would lead to potential data leak in their model, hence their reported values of model performance exceeding what was reported in this chapter.

In comparison, the metabolomic profiles of the patients allowed for much greater predictive power with most models achieving an F1 score greater than 0.72 and the highest reaching 0.826. The RTF performed much better on metabolomic data than it did on the microbiome, which can be put down to the overall performance of PCA on this dataset. This is due to RTF using PCA as a fundamental part of its model. Almost all models performed well; however, further work and introspection are needed to assess the limitations of these models. This could be achieved through the use of SHAP, LIME or permutation feature importance to gain a greater understanding of why the model is predicting the outcome it is. Moreover, to make this study more robust it should have been applied to multiple different datasets. Using a validation cohort these models and methods could be more strictly assessed. Knowing these limitations all the results should not be taken at face value but it shows the potential power of RTF and combined matrix factorisation for classification.

4.4.2 Metabolomics blind source separation

IC4 and IC1 loadings present with a high degree of enrichment for bile acids and carnitine (Supplementary Table 4.1). IBD studies have shown an enrichment of several bile acids including the primary bile acids (PBA), such as chenodeoxycholic acid and cholic acid as well as their conjugated forms. In addition, less well-studied bile acids such as keto deoxycholic acid are also hypothesised to be enriched in IBD.

Other metabolites associated with IBD are secondary bile acids (Roda et al., 2019; Thomas et al., 2022; Heinken et al., 2019) (SBA) (deoxycholic acid and lithocholic acid), alpha-muricholic (Zhang et al., 2023) acid and 7-oxo-DCA (Yang et al., 2021b). More specifically, there has been evidence of a deduction of secondary bile acids in patients with IBD compared to healthy controls (Vich Vila et al., 2023). Due to dysbiosis in IBD, there is a disturbance in the transformation of PBAs to SBAs resulting in a relative increase in PBAs and a reduction in SBAs (Yang et al., 2021b; Bromke and Krzystek-Korpacka, 2021). For PBAs to be transformed into SBAs they first need to be deconjugated. This means removing amino acids such as taurine and glycine that allow the PBAs to be water-soluble and be secreted into bile. After this, they undergo several reactions such as desulphation, dehydrogenation and dehydroxylation by various bacteria that contain bile-acid-induced (BAI) operon enzymes. However, these various bacterial transformations are only recently being mapped, so the precise nature of how they directly relate and interact with the bacteria and then, ultimately the host is still unknown. These bacteria are perturbed to different degrees in IBD due to dysbiosis, and this is what results in various changes in bile acids (Lavelle and Sokol, 2020; Zhang et al., 2023; Pratt et al., 2021).

The blanket statement of in IBD the gut metabolome sees an increase in PBA and a decrease in SBA is a broad generalisation - for instance in some cohorts the dysbiosis occurs in such a way that some conjugated forms of SBAs are increased in IBD patients (e.g. 7-ketodeoxycholic acid). In the HMP data however we see a similar trend of PBA increased in disease and a decrease in SBA when compared to the control which was also seen by the authors of the data (Lloyd-Price et al., 2019).

These various bile acid metabolites have been linked to immune regulatory roles and can also affect the gut epithelium. Normally the bile acids help the absorption of lipids as the conjugated PBAs form micelles. Still, interestingly from a translational perspective, bile acids

are also important determinants of FMT success in *Clostridioides difficile* infection (Brown et al., 2018). This again demonstrates that they have immune modulatory effects in the gut. Moreover, in CD patients, studies have shown a correlation with reduced abundances of certain bacteria that contained bile salt hydrolase (BSH) and 7α-dehydroxylation enzymes (Wang et al., 2021; Thomas et al., 2022). This correlates the microbiome composition directly to PBA present in the gut. Studies have shown associations between several genera, such as *Bacteroides*, *Bifidobacteria*, *Clostridium*, *Lactobacillus* and *Eubacterium* (*Ridlon et al.*, 2014; *Staley et al.*, 2017; *Guzior and Quinn*, 2021).

The distinction between UC and CD is not very well defined in terms of bile acids. Moreover, this distinction is even more difficult to elucidate between the various subgroups of UC and CD (e.g. ileocolonic CD vs colonic CD) (Thomas et al., 2022; Verstockt et al., 2022). Theoretically, one would expect differences given the enterohepatic circulation of bile acids that occurs via the distal ileum. Moreover, some UC patients may also have subclinical primary sclerosing cholangitis (PSC), which can affect the bile acid pool. This demonstrates the need to better define the changes in bile acid occurring across the spectrum of IBD clinical phenotypes rather than just between IBD and healthy controls (Thomas et al., 2022).

4.4.3 Reviewing methodologies

One limitation of the work in this chapter is the number of datasets used. Depending on the dataset, different algorithms may perform significantly better or worse. Therefore the algorithm chosen should match not only the data it is being applied to but also the question that is being asked of it. For example, ICA has the advantage of separating multiple independent sources of signal, being efficient when applied to large data sets, and it preserves global structures in the data. However, due to very specific assumptions that are made beforehand, especially that none of the independent sources is normally distributed, ICA has the limitation that it can suffer from crowding in the presence of a large number of observations and also, without further modifications, can lack reproducibility. Furthermore, ICA is sensitive to zero inflation or minimal values resulting in heavy tails. In this case, the heavy tails resulted from the model separating individual patients' microbiomes rather than phenotype-specific signals. This was evident from the increase in performance after log normalisation methods.

Since this study, there have been multiple methods created for the application of dimensionality reduction methods to longitudinal omics data (Mor et al., 2022; Velten et al., 2022). One such method is Tensor Component Analysis with M-product between tensors (TCAM) (Mor et al., 2022). Tensor Component Analysis (TCA) structure allows for a natural integration of the 3-dimensional array used in longitudinal data analysis (i.e. the 3rd dimensional represents the time). This follows from work conducted by Martino et al, where they created a Compositional Tensor Factorization (CTF) to uncover driving differences in microbiome compositions between phenotypes (Quinn et al., 2019).

Another method is an extension of the Bayesian factor analysis tool MOFA (Argelaguet et al., 2018, 2019) called MEFISTO. MEFISTO (Velten et al., 2022) is a method for functionally integrating spatial and temporal omic data. The model builds on the multimodal sparse factor analysis framework and uses the Gaussian process to provide a functional view of the latent factors obtained by the model. It also has temporal and spatial alignment capabilities through dynamic time warping. Although not reported in the main text, I did apply MEFISTO (Velten et al., 2022) to the HMP dataset. The 3 approaches to frame the problem were at the patient level with a group kernel, patient level without a group kernel and phenotype level with a group kernel. Each of these models was built with and without DTW as well. However, these models performed poorly and the model was unable to leverage the data. This was due to the overall size of the data (e.g. small group sides at the patient level) and the very irregular sampling of the original dataset.

For small datasets, like within a lot of omic studies, dimensionality reduction might not always perform as well as feature selection. Feature selection methods such as mutual information-based feature Selection, minimum redundancy maximum relevance, normalised mutual information feature selection, discriminative feature selection, recursive feature elimination, K best feature selection, feature selection through genetic algorithms and other wrapper-based methods (i.e. feature selection built into the model itself) may achieve great performance on the test set. However, these results do not tend to generalise well to validation cohorts. This is evident from numerous studies that have investigated IBD using microbiome data. Although there are numerous limitations to these approaches, the more of these studies we conduct, the greater our understanding of the disease and the methods we obtain.

Since this work was conducted, there has been extensive work and focus on inferring and extracting information from longitudinal microbiome data (Zhang, Guo and Yi, 2020; Luna, Mansbach and Shaw, 2020; Lugo-Martinez et al., 2019; Sharma and Xu, 2021; Armoni and Borenstein, 2022; Joseph, Pasarkar and Pe'er, 2020; Mor et al., 2022; Laccourreye, Bielza and Larrañaga, 2022; Velten et al., 2022). Interestingly there is an overlap between these methods and the methods that have been developed in this chapter as well as the approach developed in Chapter 3.

4.4.4 Future work

This study has shown the advantage of matrix factorisation methods for extracting biologically meaningful insights from various microbiome related omics data. In particular, using ICA, these subtle biological signals can be isolated from the noisy environment and then combined together to represent a meaning factor. These factors can be generalised to studies with the goal of creating a biomarker panel. Though this work shows promise, for a complex disease such as IBD a single Omic layer is not enough to uncover the underlying information. Future work could combine the processing, transformation and interpretation methodologies explored here into a multi-omic model. Furthermore, using pathway analysis and functional analysis these identified ICs can use both predictive and explainable features (Wieder, Lai and Ebbels, 2022). This model would be Independent Vector Analysis (IVA) (Kim, Lee and Lee, 2006). IVA is similar to ICA but is designed for multi-modal blind source separation problems. As of writing, there are no biological models that leverage this implementation. This would allow for the metagenomics and metabolomics layers to be combined into one model with the hope of not only improving the model's predictive performance but also our understanding of the interplay between the microbiome, metabolites and ultimately the host as well. For example, the changes in the BA pool, and the relationship this has with microbial species can be further explored (Thomas et al., 2022).

Chapter 5: Predicting the effect of the gut microbiome on the host in inflammatory bowel disease

5.1 Introduction

Multiple studies have demonstrated the role the human gut microbiome plays in both healthy and unhealthy conditions (Integrative HMP (iHMP) Research Network Consortium, 2014; Malla et al., 2018; Valdes et al., 2018; Hou et al., 2022). The previous chapters of this thesis demonstrated how we can leverage the composition of the microbiome-related omic data to find non-invasive clinical biomarkers for disease stratification. These biomarkers may provide a powerful diagnostic tool, but it does not explain the role these microbes are playing within the stratified groups.

Discovering the composition of the microbiome is crucial because the imbalance between beneficial and harmful bacteria causes a dysbiotic state that can result in inflammation. The gut microbiome is responsible for preserving the gut lining's integrity and regulating the immune response. If the balance is disturbed, it can activate the immune system excessively and trigger inflammation through altered signalling pathways. Additionally, gut inflammation can cause dysbiosis, as the inflammatory response can damage the epithelial layer and alter the gut microbiome's environment. This can create a challenging environment for beneficial bacteria to survive and thrive while allowing harmful bacteria to dominate. Various inflammatory disorders, such as inflammatory bowel disease (IBD), autoimmune disorders, allergies, and metabolic conditions, have been linked to dysbiosis (Zeng, Inohara and Nuñez, 2017).

Few studies have determined effective models for engineering the human microbiome from an unhealthy state back to a healthy state. There have been some successful therapeutic applications, either from treating recurrent *Clostridium difficile* infections from Faecal Microbiota Transplantation (FMT) (Samarkos, Mastrogianni and Kampouropoulou, 2018). One potential reason for this is the lack of translation from biomarker identification to

effective therapy, is the limited tools to describe the complex system of interactions occurring in the microbiome.

There are several ways to investigate host-microbiome crosstalk. One way is through protein-protein interactions (PPIs). It has been shown that both commensal and pathogenic bacteria have highly conserved regions, known as microbe-associated molecular patterns (MAMPs), which have the ability to trigger host-signalling pathways through pattern recognition receptors present on epithelial and immune cells (Lebeer, Vanderleyden and De Keersmaecker, 2010; Zhou, Beltrán and Brito, 2022).

These PPIs can be modelled in a systematic way using networks. PPI networks are mathematical representations of the physical interactions that take place between proteins within a cell (Barabási and Oltvai, 2004; Bebek et al., 2012). These interactions are highly specific and only occur between well-defined binding regions on the proteins. Importantly, PPIs are responsible for specific biological processes and essential functions within the cell.

Although these interactions are well-documented and described within the host. They are not as well annotated between the microbe and the host, meaning at present there are limited tools and databases to model these interactions to a high degree of certainty.

Predicted interactions from large-scale language models, like AlphaFold (Jumper et al., 2021) and ESM (Lin et al., 2022) will begin to fill this gap as the interactions identified are validated and the prediction improves. At present, there are two approaches. 1) building a network which contains all the proteins of interest but has a low certainty and annotation level; or 2) a smaller network containing fewer proteins but with a higher degree of certainty and annotation.

5.1.1 Aims

In this chapter, I leverage microbiome data (metaproteomics) and host data (transcriptomics) to investigate the role of microbes associated with increased disease activity in Inflammatory Bowel Disease (IBD) patients. In turn, this would provide insights into the host-microbe interactions (HMI) but also a framework to provide biological interpretation to the findings of machine learning (ML) models. This chapter is a proof of concept on the extended version of MicrobioLink2, a tool developed within the Korcsmaros Group (led by Lejla Gul). Before my contribution to MicrobioLink2, Lejla Gul during their PhD developed the original code implementation, concept for modelling host-microbe

interactions and aided with the conducting the analysis in this chapter. My contribution to the development of the tool was at the data preprocessing stage, software engineering (i.e. creating reusable and robust implementations of the original code) and optimisation of the algorithms used at each stage.

This chapter's aims were as follows:

- Extend the MicrobioLink2 tool to be able to ingest and preprocess metaproteomics data
- Apply MicrobioLink2 to microbial proteins associated with IBD to investigate the role of HMI in unhealthy conditions
- Explore the functional role of both microbial and host proteins in unhealthy conditions
- Provide a proof of concept investigation into the effect selected microbial proteins to have on the host

5.2 Methods

5.2.1 Microbial proteins extraction

The metaprotomes were extracted from the study conducted by Mills et al (Mills et al., 2022). In this study, the data was generated from 40 UC patients using liquid chromatography (LC)–LC–MS2/MS3 proteomic data, identifying 36,391 proteins. The authors extracted and determined the protein levels using the following approach: The relative abundances were normalised first to the pooled standards for each protein and then to the median signal across the pooled standard. An average of these normalizations was used for the next step. To account for slight differences in amounts of protein labelled, these values were then normalised to the median of the entire dataset and reported as final normalised summed signal-to-noise ratios per protein per sample (Mills et al., 2022).

Proteins which had low expression levels across all patients were removed. The dataset was then filtered for specific bacterial species identified as most informative between the control condition and disease state (see Chapter 2, 3 and 4); in this case *Bacteroides vulgatus*. The proteins were then remapped to their UniProt (UniProt Consortium, 2021) identifiers and annotated with PFAM identifiers using a custom python script which made requests to UniProt (UniProt Consortium, 2021).

5.2.2 Processing human transcriptomics data

Bulk RNA-seq data from colonic tissue of healthy controls (n=123) and UC (n=169) patients from multiple combined studies was extracted from the IBD TAMMA resource (Massimino et al., 2021). The normalised count data were then subset based on the tissue location (colon) and the disease state (control or UC). Batch effects were already handled by the authors using ComBat (Stein et al., 2015), and dataset specific genes were regressed out. Differential expression analysis was conducted using DESeq2 (Love, Huber and Anders, 2014) and the average expression of all genes per condition was calculated independently of one another using a custom python script. This resulted in three matrices; differentially expressed genes (DEGs) begin control and UC, average expression of all genes in the control group and average expression of all genes in UC.

The average gene expression matrices were filtered to remove genes with low expression levels that can arise from technical or biological noise. The data were standardised using a z-score transformation (Cheadle et al., 2003). The z-score, also known as a standard score, is a statistic that indicates the number of standard deviations that a data point deviates from the mean of a distribution. The z-score, z, is calculated using the following equation,

$$z = \frac{x - \mu}{\sigma}.$$

(Equation 5.1)

where x is the raw score, μ is the mean of the population, and σ is the standard deviation of the population.

A z-value of 0 indicates the value is at the mean of the distribution, whereas a score of +n or -n implies that the value is n standard deviations away from the mean. In this case, the z-score is used to identify genes whose expression value differs the most across a distribution. Hart *et al* published a z-score-based normalisation method that determines which genes were expressed using a comparison between expressed genes and active promoters (Hart et al., 2013). After applying the z-score transformation to the average expression matrices, genes where the z-score was greater than -3 were kept. This cut-off of -3 is the default cut-off as suggested by the authors (Hart et al., 2013). This value includes those genes where the expression value is higher than three times the standard deviation below the mean.

5.2.3 Predicting the direct effect of microbial proteins on host

To study how bacterial proteins affect host proteins, host-microbe PPI networks were generated using MicrobioLink2. It should be noted that the underlying assumption made for this investigation is that a bacterial protein can bind to a human protein if a microbial protein domain targets a short linear motif (SLiM) - a specific amino acid motif - (domain-motif interaction (DMI) or domain-domain interaction (DDI)) on the host protein. These regions and their experimentally verified interactions are identified using the ELM database (Kumar et al., 2022).

The structure of bacterial proteins was analysed using the InterProScan tool (Jones et al., 2014) to determine potential domains which were represented as PFAM and IUPRED IDs. For

the study, I analysed the potential effect of bacterial proteins on host genes derived from UC conditions. Membrane-based host proteins were extracted from the transcriptomic dataset based on subcellular location annotations from the OmniPath database (Türei, Korcsmáros and Saez-Rodriguez, 2016). This step is required to filter the potential HMIs to those that can physically happen between host proteins and proteins secreted or displayed by non-invasive bacterial species. Finally, the sequences and domain structures of the selected host proteins were obtained from the UniProt database (UniProt Consortium, 2021). The microbial and host proteins were then connected by inferring their interactions using the MicorbioLink2 pipeline, resulting in a UC condition-specific host-microbe interactome.

5.2.4 Building up a downstream signalling network

To investigate the spread of signals derived from microbial-host interactions, network propagation algorithms were employed. These algorithms link the perturbation points, host proteins in contact with microbial proteins to the differentially expressed genes, via PPIs. In turn this yielded a comprehensive and ultimately mechanistic insight into signal dissemination.

This implementation utilised a network propagation algorithm called Tied Diffusion for Subnetwork Discovery (TieDie) (Paull et al., 2013). The TieDIE approach is a method that looks for connecting genes on a network using a diffusion strategy, based on a background interaction network. Which enabled an indirect evaluation of microbial effects on signalling pathways via their interaction with cell surface proteins. In turn, providing a framework to assess the effect of microbes on downstream signalling pathways.

In the current study, a network model is constructed to investigate the signalling processes altered in the context of UC. The final network delineated the order of signal propagation from bacterial proteins to human targets, downstream signalling pathways, transcription factors, and to differentially expressed genes. To manage the complexity of large interactomes, the analysis focuses on the top 150 upregulated genes from the transcriptomic dataset.

To identify and visualise the main signalling pathways and functions connecting the membrane-based proteins and transcription factors, the intracellular network was clustered with GLay community cluster analysis (Su et al., 2010) using the clusterMaker Cytoscape plug-in.

5.2.5 Functional analysis

Functional analysis was run through gene set enrichment analysis (GSEA). GSEA determines if a specific set of genes (or pathway) is statistically significant and, therefore overrepresented within the sample genes or between conditions. Here, the observed gene set includes the nodes that are potentially bound by the bacterial proteins, and the background gene set contains all the expressed genes that are represented in the transcriptomic dataset. To perform the enrichment analysis, ReactomePA (Yu and He, 2016), clusterProfiler (Wu et al., 2021) and ggplot2 R packages were used to further visualise the results.

For network-based functional analysis, the ClueGO Cytoscape plugin was used to visualise all the functions that the human target proteins play a role in. This tool also uses data from Reactome and therefore gives consistent results with the enrichment analysis outlined above. The parameters for the tool are the following: (1) medium network specificity between the global functions and detailed reactions (3-8 hierarchical levels from the ranked pathway database), the minimum requirement is that at least 4% of the mapped genes are represented in the total associated gene list; (2) Kappa-score = 0.5 - the score measures inter-rater agreement for categorical items. In ClueGO, Kappa-score defines the term-term interrelations and functional groups based on shared genes between terms; (3) a Two-sided hypergeometric test for enrichment calculation and Bonferroni step-down p-value correction.

5.3 Results

In this study, *Bacteroides vulgatus* was selected as a focal point due to compelling evidence presented by Mills et al. Their research illuminated a significant relationship between the proteases of *Bacteroides vulgatus* and the severity of UC. However, Mills et al did not extend to modelling or elucidating the interaction mechanisms between *Bacteroides vulgatus* and the host. Our work aimed to bridge this gap by exploring the potential interactions of B. vulgatus within the host environment. Moreover, *Bacteroides vulgatus* has been implicated by multiple models used in chapter 3, but the role *Bacteroides vulgatus* plays in IBD is widely unknown (Liu et al., 2022; Mills et al., 2022). This approach was intended to frame a realistic use case of the Microbiolink2 pipeline.

5.3.1. Identification of domain-domain and domain-motif interactions

I found 812 bacterial proteins identified in the microbiome data of the UC cohort (outlined in Methods 5.2.1) all of them derived from or associated with *Bacteroides vulgatus*. These proteins were then mapped based on their sequences to Uniprot and their PFAM identifiers were extracted. Of these, 66 bacterial proteins (Supplementary Table 5.1) were connected to 290 human proteins through DDIs, resulting in 899 PPIs. Meanwhile, the DMI analysis revealed six bacterial proteins that have domains connecting to a motif on host protein sequences. Because the DDI-mediated PPIs are undirected and less specific compared to the DMI-based PPIs, I worked with the latter results in the following. These six bacterial proteins have been outlined in Table 5.1. The two proteins identified as A6KXF4 and A6L2K1 were both from a specific strain of *Bacteroides vulgatus*, strain ATCC 8482 while A0A076IWM7, W4UP76 and A0A108T7M9 came from other Bacteroides species and E6MLK6 derived from *Prevotella salivae.*

Table 5.1. *Bacteroides vulgatus* proteins were identified and predicted to bind to host membrane proteins.

Uniprot ID	Description	Organism
A6KXF4	Serine/threonine-protein kinase, AfsK-like	*Bacteroides vulgatus* strain ATCC 8482
A6L2K1	Putative integral membrane protein, with calcineurin-like phosphoesterase domain	*Bacteroides vulgatus* strain ATCC 8482
A0A076IWM7	RNA-binding protein	*Bacteroides dorei*
E6MLK6	Phosphorylase family	*Prevotella salivae* DSM 15606 strain
W4UP76	Apolipoprotein N-acyltransferase	*Bacteroides reticulotermitis* JCM 10512 strain
A0A108T7M9	Putative serine protease, AprX-like	*Bacteroides stercoris*

5.3.2 Reconstructing the bacteria-human interactome

The MicrobioLink2 pipeline identified 590 HMIs between 6 *Bacteroides vulgatus* proteins and 455 human proteins through DMIs (Supplementary Figure 5.1). I found 5 bacterial domains out of the 562 that are able to connect to the target motifs on the human protein sequence.

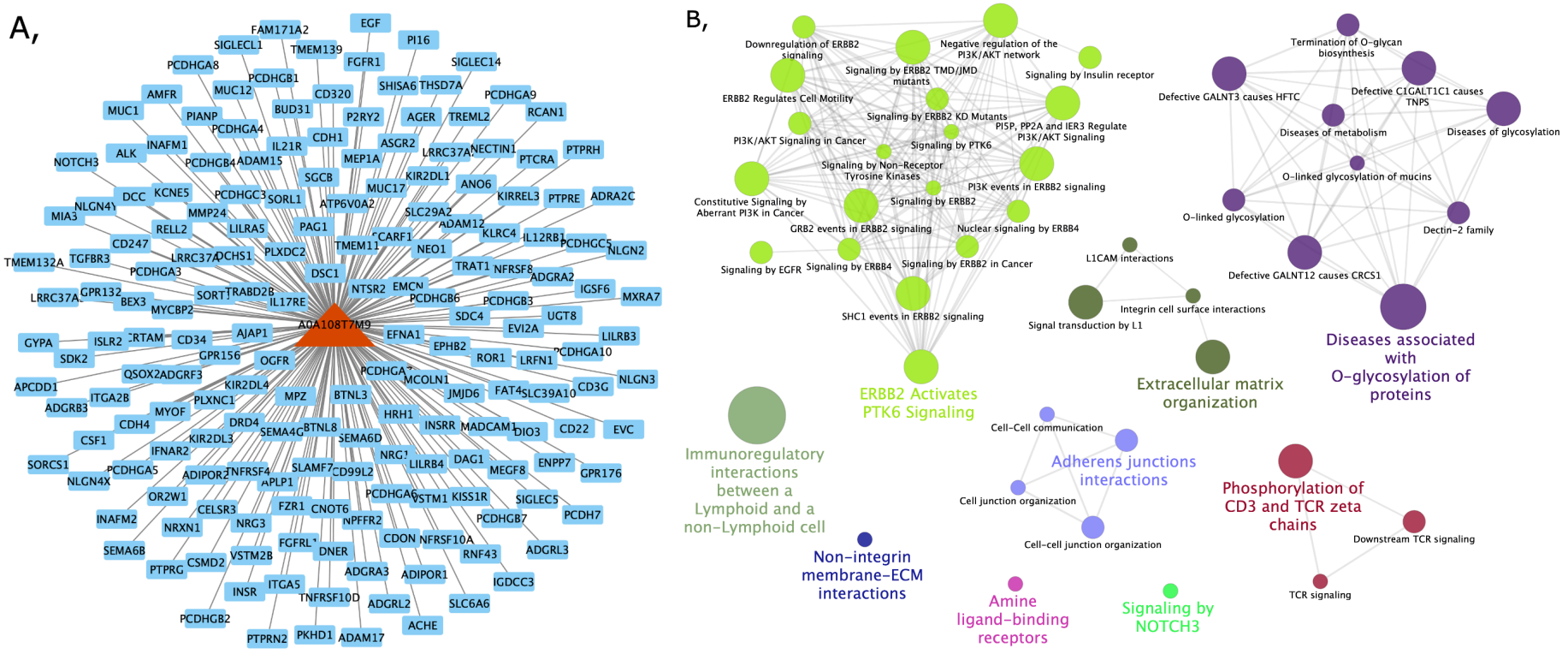


Figure 5.1. Predicted host-microbe interactions between a putative serine protease A0A108T7M9 (red triangle) and host membrane based proteins (blue rectangles). (A) direct PPIs interactions (B) results of functional analysis to determine which biological pathways are enriched in host proteins that are directly interacting with the microbial protein. The colour of the nodes represents the group of reactions that belong to the broader term (highlighted by bold font type). The size of the nodes correlates with the p-value corrected with Bonferroni step down approach. The edge between nodes shows the relationship between reactions. The figure was created in Cytoscape using the ClueGO package.

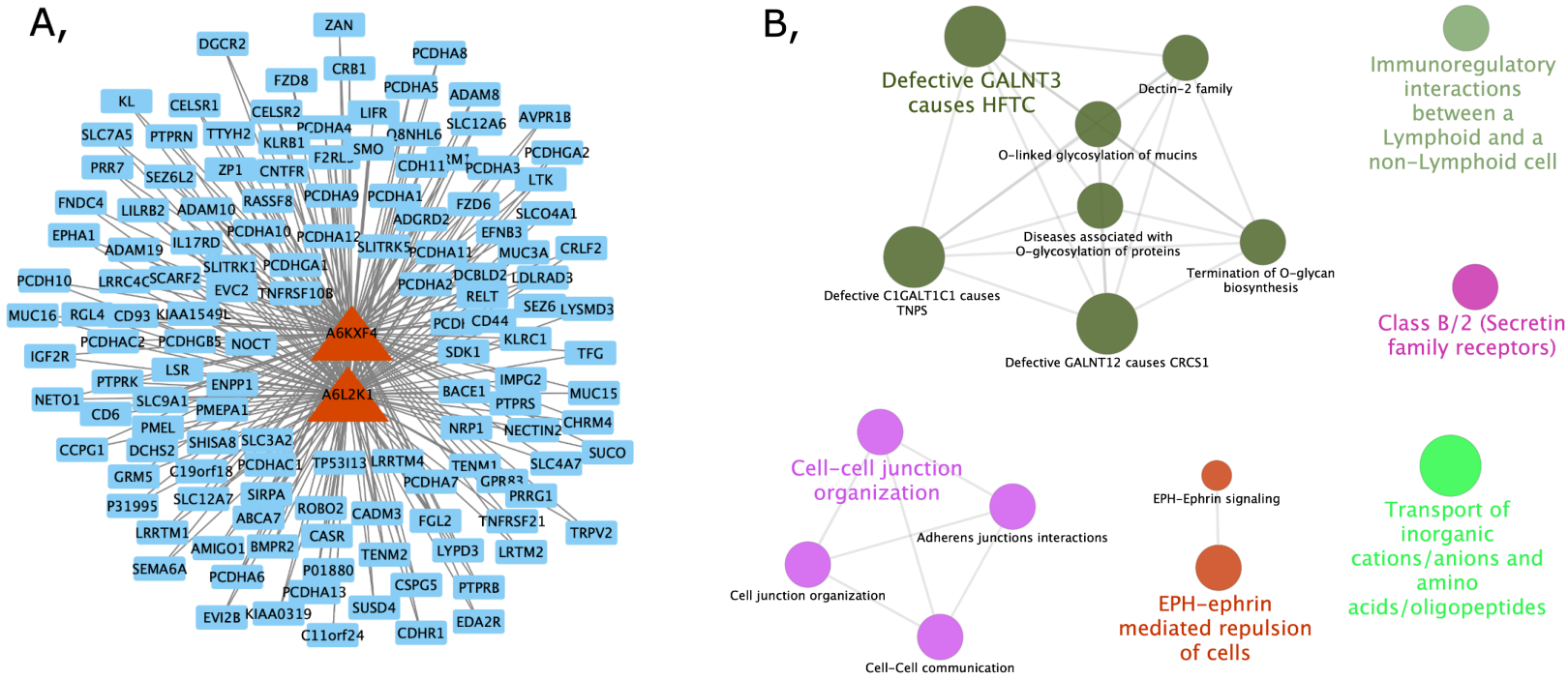


Figure 5.2. Predicted host-microbiome interactions between A6KXF4 and A6L2K1 (red triangles) with the host membrane based proteins (blue rectangles). (A) direct PPIs interactions. (B) results of functional analysis to determine which biological pathways are enriched in host proteins that are directly interacting with the microbial protein. The colour of the nodes represents the group of reactions that belong to the broader term (highlighted by bold font type). The size of the nodes correlates with the p-value corrected with Bonferroni step down approach. The edge between nodes shows the relationship between reactions. The figure was created in Cytoscape using the ClueGO package.

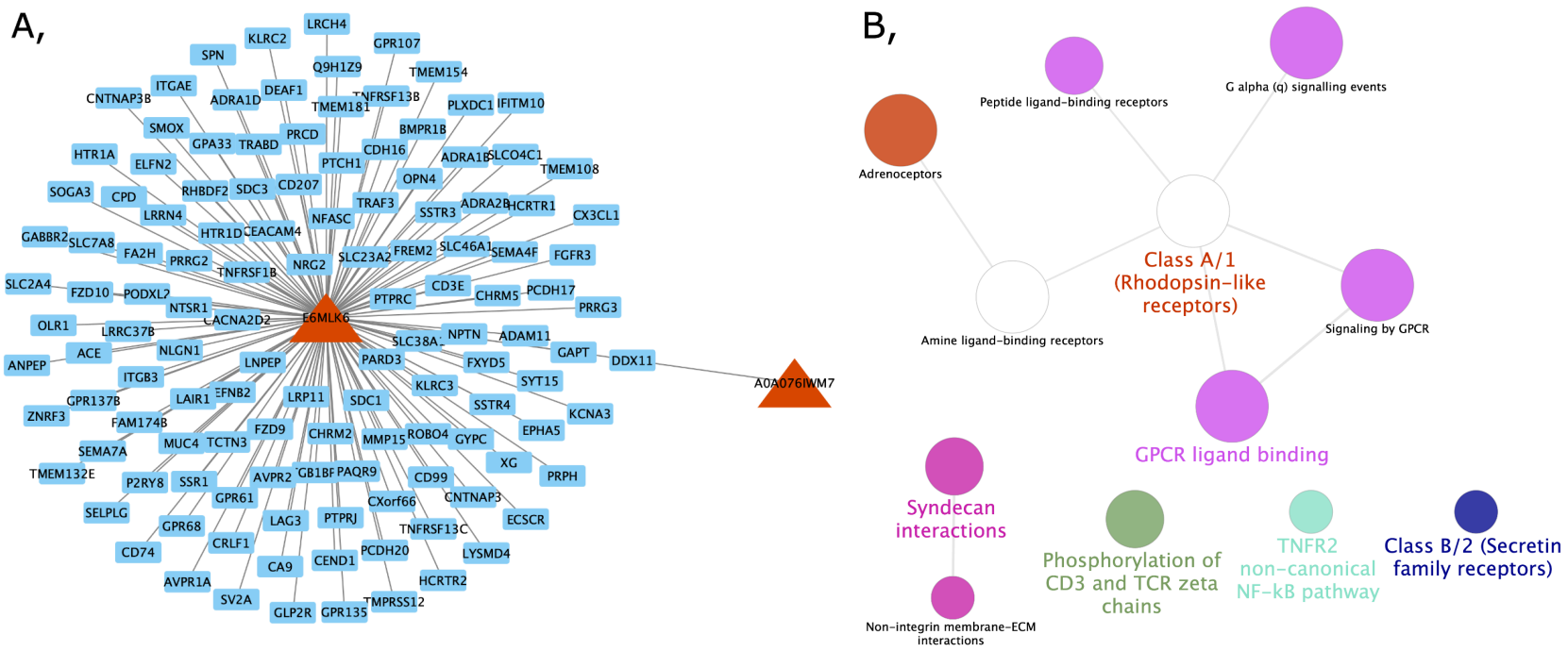


Figure 5.3. Predicted host-microbiome interactions between E6MLK6 and A0A076IWM7 (red triangles) with the host membrane based proteins (blue rectangles). (A) direct PPIs interactions. (B) results of functional analysis to determine which biological pathways are enriched in host proteins that are directly interacting with the microbial protein. The colour of the nodes represents the group of reactions that belong to the broader term (highlighted by bold font type). The size of the nodes correlates with the p-value corrected with Bonferroni step down approach. The edge between nodes shows the relationship between reactions. The figure was created in Cytoscape using the ClueGO package.

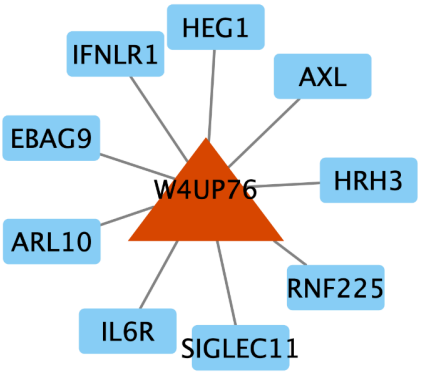


Figure 5.4. Predicted host-microbiome interactions between W4UP76 (red triangle) with the host membrane based proteins (blue rectangles). The small number of membrane-based host proteins were not enough to perform a functional enrichment analysis. The figure was created in the Cytoscape network visualisation tool.

5.3.2 Functions of human target proteins

The ClueGO functional analysis highlighted the main pathways and reactions for each cluster of human target genes (Figure 5.1,, 5.2., and 5.3.). The first cluster included the 186 targets of the A0A108TZM bacterial protein. These molecules are involved in the PTK6 and Notch signalling, extracellular matrix organisation, post-translational modification of proteins, intercellular interactions, etc (details in Figure 5.1.). The second cluster involved 134 human proteins targeted by A6L2K1 and A6KXF4 bacterial proteins. Similarly to cluster 1, proteins related to intercellular interactions are affected but this group of proteins involves Secretin family receptors and members of the EPH-Ephrin signalling (details in Figure 5.2). Cluster 3 described 126 targets potentially bound by A0A076IWM7 and E6MLK6 proteins enriched with rhodopsin-like receptors, Secretin family receptors, members of the GPCR signalling and intercellular interaction related molecules (details in Figure 5.3). Finally cluster 4 consisted of 9 human targets where ClueGO could not identify enriched functions (details in Figure 5.4) but targets included proteins like an Interferon lambda receptor and Interleukin 6 receptor.

All host proteins that have direct interactions with the microbial proteins from every cluster were then aggregated together. clusterProfiler (Wu et al., 2021) and ReactomePA as the pathway database was used for the over-representation analysis and highlighted the GPCR signalling as the most enriched pathway but the analysis revealed cell-cell interaction-related processes as well (Figure 5.5).

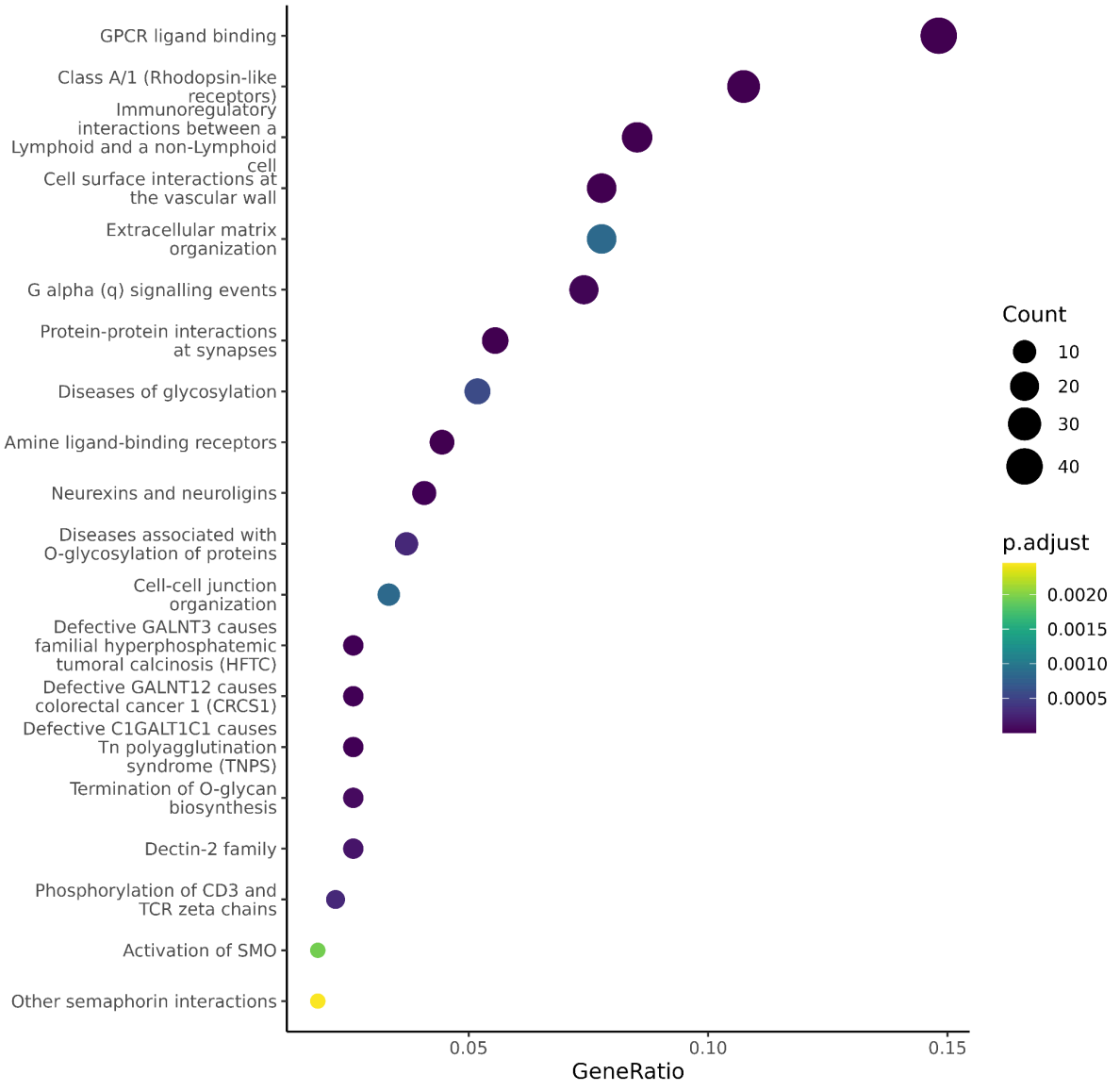


Figure 5.5. Functional enrichment analysis of the human target proteins. Over-representation analysis of the host proteins which have direct interactions with microbial proteins across clusters found in the HMI network.

5.3.3 Effect of bacterial proteins on downstream signalling

A multi-layered network was constructed to model the potential effect of microbial proteins on downstream signalling. The network consisted of three types of molecular connections, including host-microbe, human protein-protein, and transcription factor-target gene interactions. The TieDie algorithm was used to integrate various inputs, including the expression of 455 host membrane proteins affected by bacteria inUC, 51 transcription factors regulating the expression of the top 150 upregulated DEGs in UC condition, and an intracellular signalling network. This resulted in an intermediate contextualised PPI network with 18,248 directed and signed interactions among 5,390 proteins (derived from the transcriptomics dataset) in the UC samples.

The output of the algorithm included the inferred signalling network and the heat of each node in the network. The heat represents the influence or activity generated by a particular node or set of nodes in the network. The greater the value of the heat the greater the influence of the node and therefore the behaviour or information is propagated throughout the network. The inferred network included five types of nodes: bacterial proteins (5), human membrane proteins (136), intermediate signalling proteins (324), transcription factors (39), and DEGs (24). Each of these layers has been annotated in Figure 5.6.

The ClueGO analysis identified enriched functions for each cluster in the intracellular PPI network separately (see Figure 5.6). The combination of pathway information from Reactome and heat values derived from TieDie revealed the diverse signalling pathways involved in the effect of bacterial proteins. 14 functional clusters were identified, including those involved in MAPK, VEGF, TLR4, TGF-beta signalling, apoptosis, and DNA repair. Figure 5.6 provides more detailed information on these clusters.

The analysis showed that the highest heat values were observed in clusters 2, 3, 6, and 9, which were associated with inflammation-related processes (Supplementary Figure 5.1). These findings suggest that the identified signalling pathways and functional clusters may play a critical role in UC pathogenesis mediated by bacteria.

Finally, I examined the over-represented functions among the reached DEGs in the TieDie network. Not surprisingly, the cytokine-mediated signalling was significantly enriched (p-value < 0.05) compared to the top 150 upregulated DEGs in UC samples compared to healthy condition (Figure 5.7).

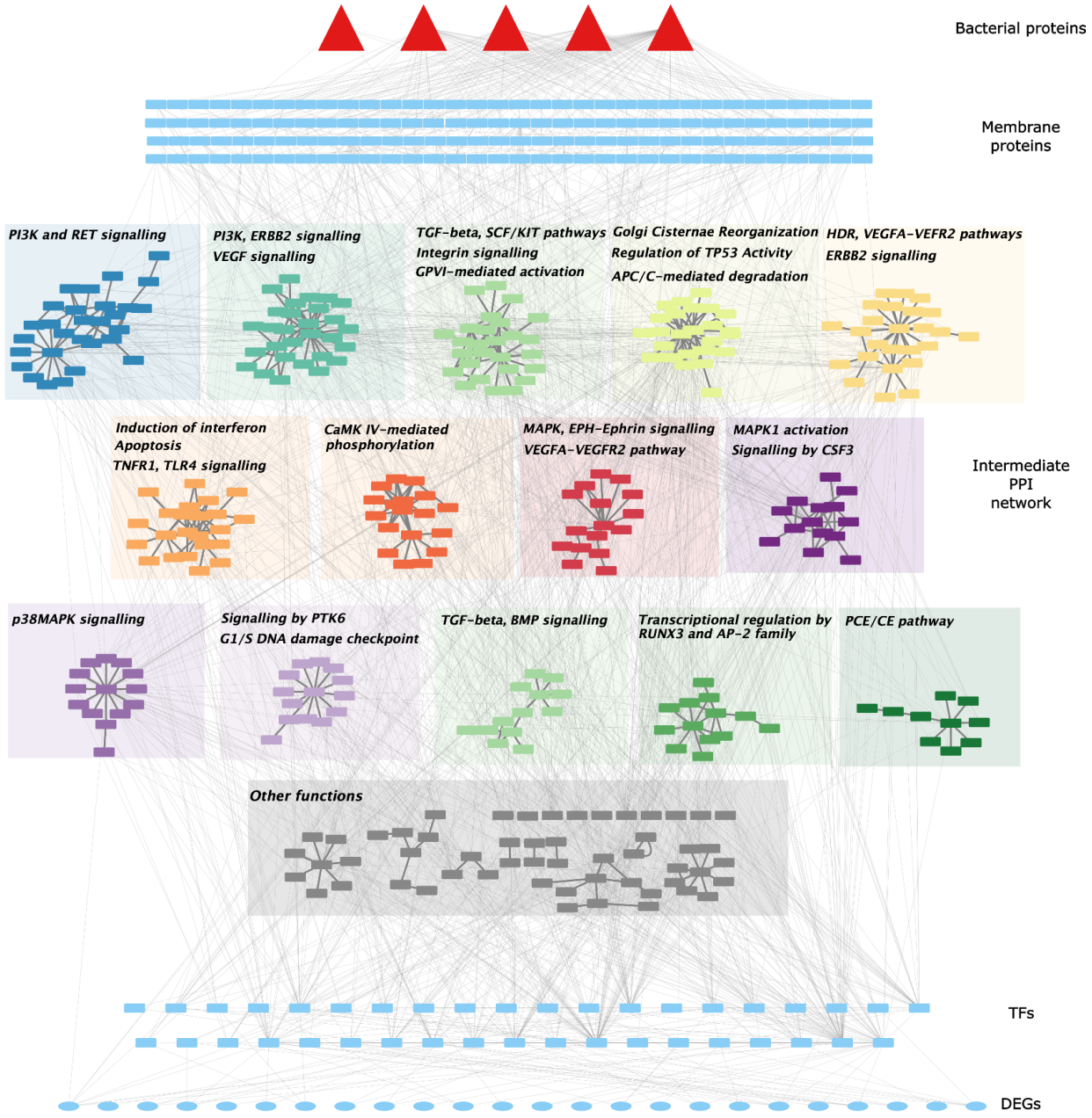


Figure 5.6. Inferred multi-layer host-microbe protein-protein interaction (PPI) network from the source *Bacteroides vulgatus* microbial proteins (red triangles) to the host differentially expressed genes through downstream signalling proteins in ulcerative colitis. The full resulting PPI network from Microbiolink2 pipeline is annotated for the functional clusters in the intermediate PPI network. This figure was created using Cytoscape (Shannon et al., 2003) network visualisation and analysis software environment.

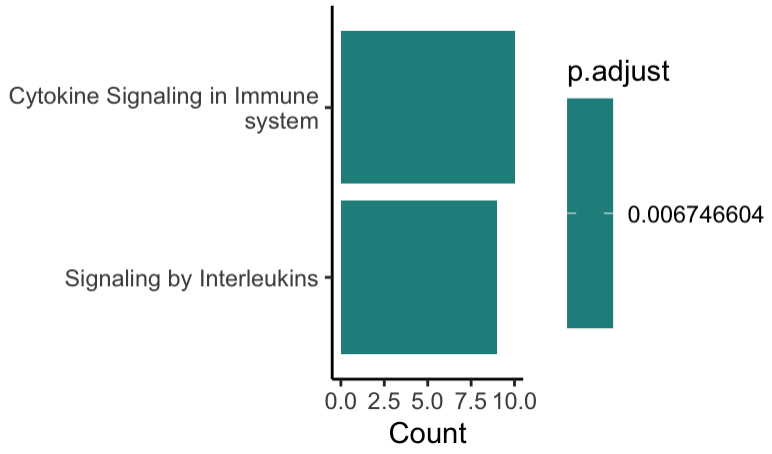


Figure 5.7. Enriched functions among the DEGs in TieDie compared to the top 150 upregulated genes in UC. The functional enrichment demonstrates a large host immune response in UC through both cytokine and interleukin signalling.

5.3.4 Effect of *Bacteroides vulgatus* on GPCR and MAPK pathways

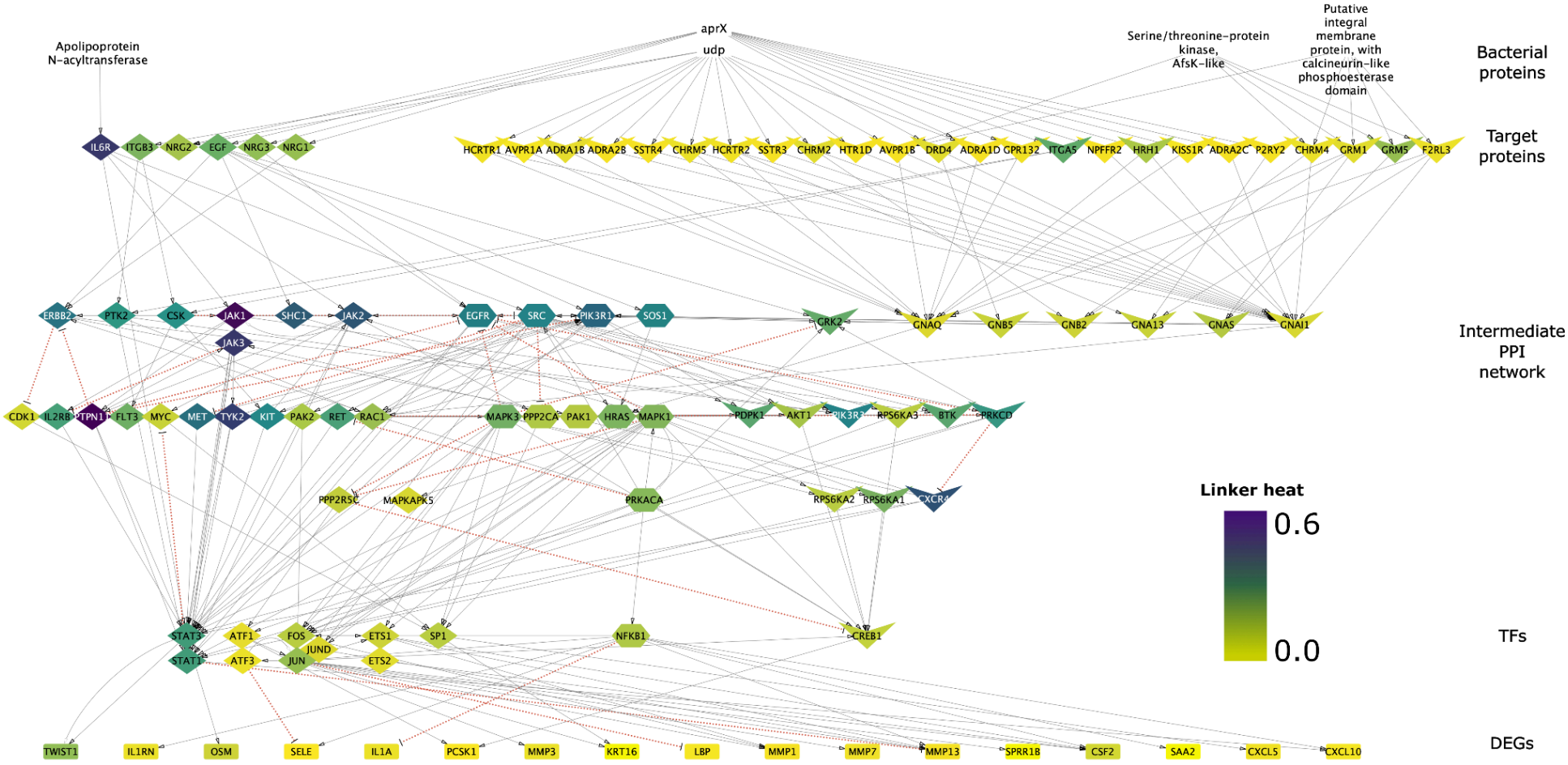
Functional analysis was performed on a diverse set of proteins, bacteria targets and downstream signalling networks, identifying the potential role of GPCR and MAPK signalling to mediate the effect of *Bacteroides vulgatus*. To elucidate this complex TieDie networks shown in Figure 5.6, the pathways and the cross-talk between bacteria and the host proteins were reconstructed (Figure 5.8) using collected pathway members from the ReactomeDB and the transcription factors from the literature (Liu et al., 2018; Guo et al., 2020; Coulthard et al., 2009; Cuadrado and Nebreda, 2010; Huang, Shi and Chi, 2009).

MAPK and GPCR pathway members were selected from the TieDie network and only DEGs where the expression is influenced by a transcription factor (TF) in the MAPK and GPCR pathways were kept. This enabled the focused analysis and modelling of the effect HMIs have on differentially expressed genes in UC through the MAPK and GPCR signalling pathways.

Finally, to observe the propagation of the signal through the network, heat values from the TieDie analysis were used to annotate the network. The result highlighted that the heat

values are higher in the MAPK pathway suggesting the significant role of the pathway in downstream signalling (Figure 5.8.).

Figure 5.8. **(Next page) Subset network modelling the host-microbe interactions and regulatory interactions between bacterial proteins and GPCR/MAPK pathway in ulcerative colitis.** Nodes shapes represent the pathways involved: diamond-shaped nodes are MAPK pathways members; V-shaped nodes are GPCR pathway members; hexagon-shaped nodes are common members between the two pathways and therefore highlight cross-talk between MAPK and GPCR pathways; rectangle-shaped nodes represent differentially expressed genes in ulcerative colitis. Node colour represents the linker heat of the signal propagation throughout the network. The red dotted line shows inhibitory interaction and the black line shows stimulatory interaction. This figure was created using Cytoscape (Shannon et al., 2003) network visualisation and analysis software environment.


Apolipoprotein N-acyltransferase
aprX
udp
Serine/threonine-protein kinase, AfsK-like
Putative integral membrane protein, with calcineurin-like phosphoesterase domain
Bacterial proteins
IL6R ITGB3 NRG2 EGF NRG3 NRG1
HCRTR1 AVPR1A ADRA1B ADRA2B SSTR4 CHRM5 HCRTR2 SSTR3 CHRM2 HTR1D AVPR1B DRD4 ADRA1D GPR132 ITGA5 NPFFR2 HRH1 KISS1R ADRA2C P2RY2 CHRM4 GRM1 GRM5 F2RL3
Target proteins
ERBB2 PTK2 CSK JAK1 SHC1 JAK2 JAK3 EGFR SRC PIK3R1 SOS1 CRK2 GNAQ GNB5 GNB2 GNA13 GNAS GNAI1
Intermediate PPI network
CDK1 IL2RB TPN1 FLT3 MYC MET TYK2 KIT PAK2 RET RAC1 MAPK3 PPP2CA PAK1 HRAS MAPK1 PDPK1 AKT1 PIK3R2 RPS6KA3 BTK PRKCD
PPP2R5C MAPKAPK5 PRKACA RPS6KA2 RPS6KA1 CXCR4
Linker heat
0.6
0.0
STAT3 STAT1 ATF1 ATF3 FOS JUND JUN ETS1 ETS2 SP1 NFKB1 CREB1
TFs
TWIST1 IL1RN OSM SELE IL1A PCSK1 MMP3 KRT16 LBP MMP1 MMP7 MMP13 SPRR1B CSF2 SAA2 CXCL5 CXCL10
DEGs

5.4. Discussion

The microbiome plays an important role in homeostatic processes in the host therefore, the altered community composition leads to differences in host signalling (Wu and Wu, 2012). Currently, there is a lack of studies describing the connection between altered microbial communities and host signalling responses by analysing and integrating multi-omic datasets. In this chapter, a proof of concept was presented to predict interactions between host and microbes by focusing on selected bacterial proteins derived or associated with strains of *Bacteroides vulgatus*, and its potential signalling mechanisms in driving UC. Although *Bacteroides vulgatus* was previously considered as a commensal bacteria that exhibits probiotic properties in mouse models, more recent research has shown how this bacteria can also play a role in not only IBD pathogenesis but also increased disease activity for individuals with UC (Mills et al., 2022). The presented study highlights the potential involvement of *Bacteroides vulgatus* in the development of UC.

I analysed public metagenomic and transcriptomic datasets and combined them with network resources to establish a host-microbe interactome and the consequence of these HMIs on the downstream signalling network in UC conditions. The identified 6 bacterial proteins are potentially able to bind to and modify host membrane proteins mostly through enzymatic domains, including kinase, protease, phosphoesterase, phosphorylase and acyltransferase functions. The classic example of host-microbe interactions comes from pathogenic bacteria secreting proteins that selectively bind to proteins to regulate the host cell's biological activity; these proteins are known as effector proteins (Weigele et al., 2017). However, bacterial proteins can also interact through other mechanisms such as secreted human proteins, bacterial proteins secreted into extracellular spaces, membrane vesicles that are endocytosed or fuse with the human cell membranes, bacterial cellular lysate, translocation due to dysfunction and direction contact with M cells, dendritic cells or epithelial cells. The 6 bacterial proteins identified in this chapter are all membrane proteins, which means they have the ability to interact with the host through a complex system of signal transduction. By targeting proteins on the plasma membrane, the bacteria leverage a core part of eukaryotic signalling networks. However, the precise signalling mechanics for some bacterial proteins are largely unknown, which shows the potential for a tool such as

MicrobioLink, which enables biologists to create hypotheses about how the bacterial proteins could be targeting host proteins (Weigele et al., 2017; Zhou, Beltrán and Brito, 2022).

The main processes that these proteins target included the GPCR signalling and the cell-cell interaction. Both functions play critical roles in host-microbe interactions: G protein-coupled receptors (GPCRs) are cell surface receptors that transmit signals from outside the cell and are involved in numerous physiological processes. Commensal bacteria, such as *Bacteroides vulgatus*, often mimic the ligands for these receptors, therefore perturbing the signalling in hosts (Cohen et al., 2017). In UC, GPCR signalling contributes to the recruitment and activation of immune cells, causing chronic inflammation and tissue damage in the colon (Zeng et al., 2020). The gut microbiome can affect cell-cell interactions, therefore modifying the integrity of the intestinal epithelial barrier that leads to increased permeability and allows bacterial products to stimulate the immune system (Gieryńska et al., 2022).

The TieDie network propagation algorithm (Paull et al., 2013) revealed the effect of the bacteria-perturbed membrane proteins on differential gene expression in UC. The integrated, multi-layered network highlighted the main clusters where the signal is going through, including several already published functions, such as the PI3K and MAPK signalling, but also highlighted new potential candidates (e.g. ERBB2-signalling).

Phosphatidylinositol-3 kinase (PI3K) signalling contributes to the activation and migration of immune cells and to the disruption of the intestinal epithelial barrier, two factors that play a key role in UC pathogenesis. Specifically, the PI3K signalling pathway can promote the production of pro-inflammatory cytokines and chemokines, which can recruit and activate immune cells in the colon. In addition, the PI3K pathway can influence the integrity of the intestinal epithelial barrier by regulating the expression of tight junction proteins, which help to maintain the physical barrier between the gut lumen and the underlying immune system (Huang et al., 2011). Evidence shows that gut microbiota composition depends on the PI3K signalling, which has been shown to regulate the production of antimicrobial peptides by intestinal epithelial cells. Also, the gut microbiome can modulate PI3K signalling, with certain gut bacteria producing metabolites that activate or inhibit PI3K signalling in host cells (Mohseni et al., 2021).

Similarly to the PI3K pathway, MAPK signalling is a potential candidate that mediates the effect of the altered gut microbiome on inflammatory processes by enhancing

proinflammatory cytokine production. MAPKs are enzymes that regulate cellular processes such as cell growth, differentiation, and survival, as well as immune response and inflammation (Yang et al., 2022).

Studies have also demonstrated that dysregulated MAPK signalling influences the gut microbiome structure and function, which may contribute to the development of IBD (Guardamagna et al., 2022).

While the functional analysis highlighted the GPCR pathway as the most significant function targeted by bacterial proteins, the downstream analysis revealed the crucial role of the MAPK pathway in mediating the effect of the interspecies interactions. The literature describes a cross-talk between the pathways: upon ligand binding to a GPCR, the receptor undergoes a conformational change that allows it to interact with a G protein. The activated G protein dissociates from the receptor and activates downstream effectors, including the MAPK pathway. The specific G protein that is activated depends on the type of GPCR, and different G proteins can activate different MAPK pathways (Hur and Kim, 2002). The reconstructed GPCR-MAPK network supported the same model, as the bacterial proteins mostly connected to the GPCR signalling members and then the signal connected to the MAPK pathway through the shared pathway members. The analysis of the heat showed that the MAPK signalling components have higher values, suggesting that these proteins have more influence on the network. The novelty of the established workflow presented in this chapter comes from the exploratory power it provides. Typically, in microbiome analysis, KEGG or Enzymatic pathways are used to determine the functional potential of a community of bacteria (Kanehisa et al., 2023). These pathways are derived from a consensus of all literature research (Creixell et al., 2015). Although PPI networks are often oversimplifications of complex biological processes, they process the ability to reveal potential information that cannot be identified or is hidden within a well-defined pathway (Barabási, Gulbahce and Loscalzo, 2011; Gosak et al., 2018; Creixell et al., 2015; Yang et al., 2019).

There are several limitations to using MicrobioLink2 for predicting HMIs. One challenge is that predicting interactions between bacterial and human proteins is difficult due to limited knowledge of the motifs bound by bacterial domains. The ELM (Kumar et al., 2022) and 3did (Mosca et al., 2014) databases also limit results to domains found in eukaryotes, which can miss bacteria-specific structures. To overcome this, machine learning-based approaches can be used to predict bacterial domain structures and potential target motifs. Another

limitation is that it is unclear whether bacterial proteins activate or inhibit host proteins. To address this, manually curated information about domain-binding motifs could be integrated into HMI predictions. Additionally, assuming that every expressed transcript in transcriptomics leads to a functional protein is not accurate, as post-transcriptional and post-translational modifications can affect the RNA structure and protein activity. Analysing proteomics and transcriptomics from the same samples could improve the accuracy of the model. Finally, as these connections are predicted interactions until the resulting pathways are validated, it is difficult to know if the biological system will behave as described.

There are other existing HMI tools, such as interSPPI (Yang et al., 2019), which use an ensemble of different machine learning models to score and predict the likelihood of interspecies interactions. InterSPPI provides a much higher level of coverage than MicrobioLink2 and, therefore, has the potential to explore a larger range of microbial proteins. Nevertheless, as MicrobioLink2 uses experimentally validated domain-motif (SLiM) interactions from ELM upstream of the network, there is a higher level of certainty that the predicted interactions will hold true.

Despite the limitations described here, the applied HMI pipeline combines gap-filling approaches, such as structural PPI prediction and network analysis, which highlight the importance of condition-specific gene expression. I not only identified the potentially diverse role of *Bacteroides vulgatus* in UC conditions but also revealed the background of biological processes on the molecular interaction level.

As the prediction of PPIs has improved in the last few years, and more and more machine learning approaches have come to light, I plan to focus more on deep learning methods that use neural networks to model the sequence, structure, or both of the interacting proteins. AlphaFold2 is a protein structure prediction algorithm developed by the European Molecular Biology Laboratory and the University of Washington. It uses deep learning techniques to predict the 3D structure of a protein from its amino acid sequence. While the original aim of AlphaFold2 is to predict 3D protein structures, bacterial domains can also be inferred with the same algorithm. Google's nearest rival in this space is Meta, which released its tool for protein structure prediction called ESM (Rives et al., 2021). However, they also extended this to bacterial proteins, releasing the ESM Metagenomic Atlas in 2022 (https://esmatlas.com/).

It's important to keep in mind that the accuracy of the prediction usually depends on the specific input, and the quality of the prediction may vary for different bacterial domains. In

this study a strict filtering threshold was placed on the database used such that only verified DMIs and DDIs were carried forwards. Although this limited the scope of the study, it did increase the certainty of the predictions. It is, however, recommended that these predictions be validated using experimental methods. An example of this would be the study conducted by Balkenhol *et al.* In this study, they validated the host-pathogen interactions using *Aspergillus fumigatus* in mice as a case study. After identifying candidate PPIs, they experimentally validated using a ligand binding assay (Balkenhol et al., 2022). However, this is a non-trivial task as the majority of the bacteria in the human gut are not currently culturable, thus limiting the experimental validation to a select group of bacteria (Balint and Brito, 2023). To infer networks on a smaller scale shows how effective MicrobioLink2 is overall and compared to other interspecies interaction prediction approaches. In this chapter, I presented a use case that describes the need for this level of granularity when predicting host-microbe interaction networks.

Chapter 6: Integrated Discussion

The role the gut microbiome plays in IBD pathogenesis and disease severity remains a key challenge for researchers and clinicians worldwide. With an ever-increasing incidence rate, new tools and methods for predicting patient-level diagnosis, prognosis and drug response are needed. And importantly, we must ensure that these tests should be as accessible and non-invasive as possible. This is where the human gut microbiome has a vital advantage as it is completely non-invasive to take faecal samples from IBD patients. That being said, understanding the complexity of the diseases and that they are not just one component, but instead, the combination of genetics, environmental factors, and the microbiome, requires more than just the analysis of a single omics.

In Chapter 2, I showed the current standard of microbiome analysis by applying it to the largest publicly available longitudinal microbiome study in IBD (Lloyd-Price et al., 2019). This analysis identified areas of weakness in the current methods to explore longitudinal microbiome data. Most of the popular tools are not appropriate for compositional analysis, i.e. those which have been derived from transcriptomic analysis like DESEQ2 (Love, Huber and Anders, 2014) and EdgeR (Robinson, McCarthy and Smyth, 2010; McCarthy, Chen and Smyth, 2012; Chen, Lun and Smyth, 2016)), or have not been designed to account for longitudinal samples, such as LEfSe (Segata et al., 2011). Furthermore, even with increasingly large studies taking place, there are still ongoing issues with the ordination and clustering of the microbiome data. The application to longitudinal data would only exacerbate the poor ordination and clustering of the data.

In an attempt to address these issues, in Chapter 3, I developed an approach to try and account for the patient-specific baseline as well as to try and identify a global IBD microbiome signature when compared to healthy controls. The method identified bacterial species that were more likely to experience a shift over the course of an individual time course in UC, CD, and healthy controls.

The outcomes of Chapter 3 highlighted the non-linear, highly-dimensional, noisy and complex nature of microbiome data. In Chapter 4, I extended the approaches to handle longitudinal data, namely using patient-specific baseline transformation (FCBT and SBT), to

different matrix factorisation algorithms and machine learning classification algorithms. Moreover, as matrix factorisation methods are not count data specific like the Pooled Species Precision (SPM) developed in Chapter 3, the models were also applied to metabolomics data. This is important as metabolomics are considered to be the closest omics to phenotype (Johnson and Gonzalez, 2012; Johnson, Ivanisevic and Siuzdak, 2016; Patti, Yanes and Siuzdak, 2012). Using metabolomics to identify biomarkers has a few major advantages over metagenomics as extensive metabolomic pathway research has already been conducted, metabolites directly interact with both the host and microbiome and, like with metagenomics, it can also be extracted from faecal samples. However, there are two main challenges with extracting biomarkers from metabolomic data: (1) accurately identifying the best biomarkers molecules and (2) which molecules among the numerous other dysregulated metabolites are the best phenotype modulator remains an open question (Guijas et al., 2018).

To address the limitations in biomarker identification, there is clearly more work to be done to further utilise ML and DL models. One of the biggest steps forward in the AI models is the recent introduction of foundational models. A foundation model, for example, a large language model (LLM), can be trained on broad sets of data that can be adapted to a broad spectrum of downstream tasks with little to no fine-tuning of parameters. These models require an extremely large amount of data, which the model uses to apply the information it has learned to the new question it has been asked. It achieves this through self-supervised and transfer learning and can perform both generative tasks (i.e. predicting new protein structures) or classification tasks (e.g. predicting cell-type annotations) depending on the architecture used. The longitudinal nature of the microbiome and metabolome can be understood through the information captured across a “gut microbiome atlas”, which could result in a better understanding of the dynamic systems at play. At the same time, the model can also retain knowledge from the host’s immune system, for example. But like all foundational models, it needs to be treated with care, and a huge amount of work will be needed to validate the models’ findings properly.

One limitation of AI models is their lack of interpretability. Interpretability is the ability of the model to explain the results they are predicting. This is one of the biggest barriers to the adoption of more complex machine learning algorithms. To address this, a key area of research with current AI methods is explainable AI. This research area focuses on the

development of tools and frameworks for interpreting the results of the ML models. This is particularly important for any biological insights found by ML models. There are several ways to approach this mathematically; however, in the application of healthcare, the explainitiablity of a model needs to also be combined with a prior understanding of the biological system. In Chapter 5, I demonstrate and provide a proof of concept for a framework to perform downstream analysis and interpretation of how a selected subset of microbial proteins interact with the host through host-microbiome interactions. In complex diseases, it is known that the microbiome does not work in isolation, and it is the interplay between the host and microbes that can result in different biological responses. Thus the MicrobioLink2 pipeline provides a way to provide additional information to a single omics. The framework is model agnostic and, therefore, can be used to help explain and explore any selected microbial proteins and their effect on the host. Moreover, future work could extend the framework of MicrobioLink2 so that it can also be used on the metabolome. Once again, further utilising network diffusion and the wealth of existing metabolomic data as a reference database to model patient-specific host-metabolite interactions. Even though I did not apply this approach to metabolomic data in this thesis, this, in turn, could address the second challenge of identifying which dysregulated metabolites are the best phenotype modulators. The pipeline used in this chapter will be released as a publication and made open-access. To ensure support for the tool, I will work with other members of the group to ensure a smooth handover of the tool is made.

In addition to the biological insights described in this thesis, another outcome is the developed methods. Each of these chapters builds up a framework and workflow for analysing longitudinal microbiome data. Because of this, there was a heavy emphasis on productisation and software engineering throughout this thesis. Each chapter has its own module, which allows for further data analysis. To ensure that the methods implemented are correct, there are also unit tests throughout the code base. This has the additional advantage of ensuring reproducibility; for example, after a new package is updated, the unit test will flag if this affects the codebase or not. The entire code base is wrapped in Python and is deployable either through local installation, Docker or Singularity image, depending on the platform; see Figure 6.1 for more details.

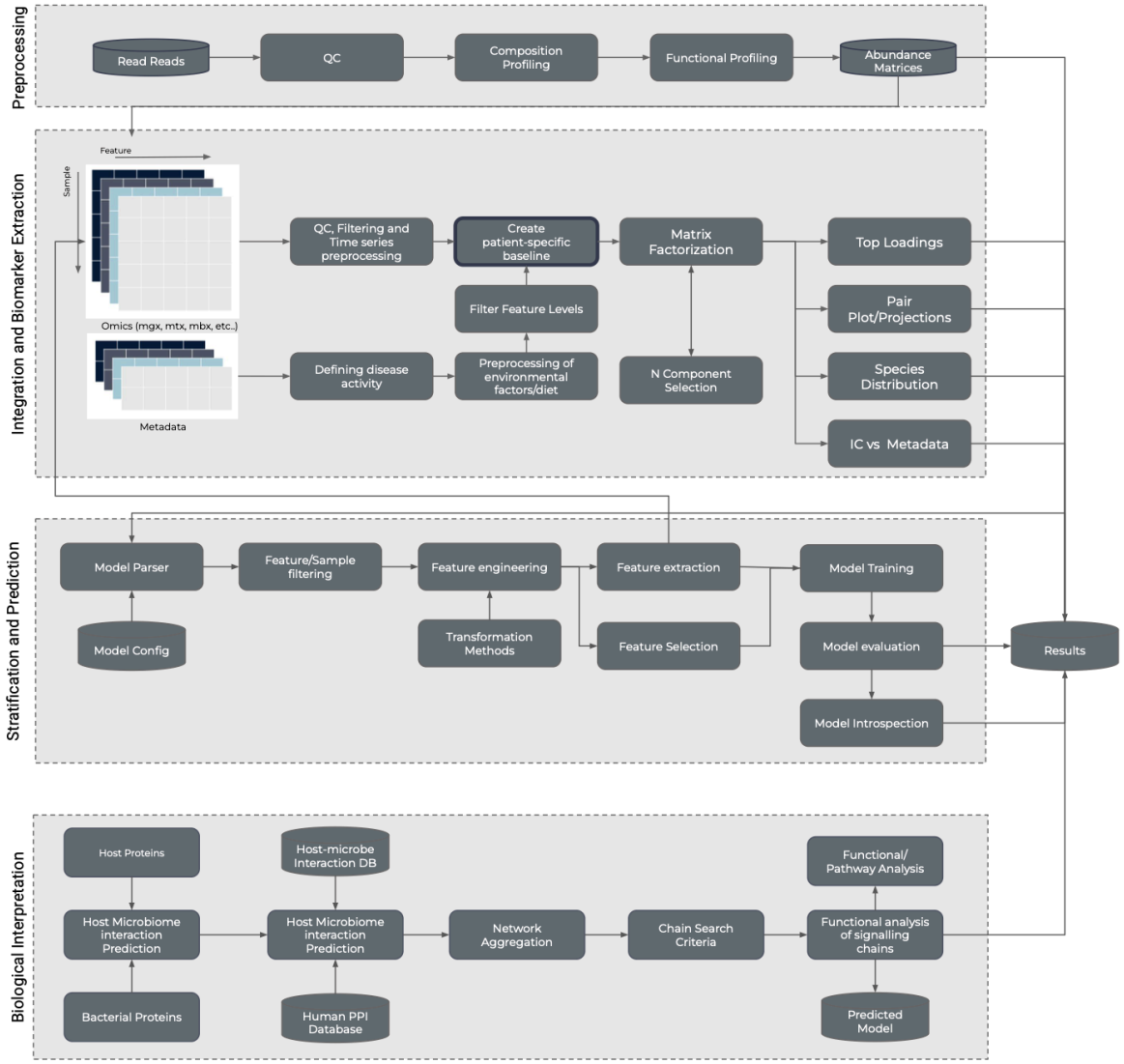


Figure 6.1. Oviewer of the framework created in this PhD research project. The framework can be broken into 4 key parts. From top to bottom. (1) Preprocessing is a module which can be replaced easily with any preprocessing scripts/functions that are needed. Therefore can be quickly adapted to omic data. (2) Integration and biomarker extraction encompass the work seen in chapter 4. (3) Stratification and prediction encompass the work in Chapter 3, and 4. (4) For biological interpretation and functional analysis, which can be seen in Chapter 5.

The MOTION study is a longitudinal microbiome and multi-omic study following three different risk categories of dementia patients over multiple years with the aim of identifying prognostic indicators. The original plan of my PhD project was to apply the developed models to data generated in parallel through the BBSRC-funded MOTION study. The restrictions put in place due to COVID-19 resulted in none of the planned data being generated in time to analyse it for my PhD, meaning the decision was made to switch to publicly available data. If the MOTION study had gone ahead, the key difference would have been the sampling rate. Although this is difficult to control in longitudinal studies, the monthly rate proposed for the MOTION study would've meant there would have been consistent sampling, and therefore, the data could have been used as a time-series.

In conclusion, this PhD research has provided and explored new ways to investigate and extract potential prognostic indicators from the human gut microbiome over longitudinal omic data. This was achieved through both unsupervised or supervised methods, depending on the amount of metadata or the question at hand. To aid in the overall interpretability of the model developed, network and systems biology approaches were combined together to explain how the extracted microbe(s) could interact with the host. This ultimately led to further mechanistic insights and understanding of the interplay between the host and the microbes during healthy and unhealthy conditions, as demonstrated with a specific example.

References

Aden, K., Rehman, A., Waschina, S., Pan, W.-H., Walker, A., Lucio, M., Nunez, A.M., Bharti, R., Zimmerman, J., Bethge, J., Schulte, B., Schulte, D., Franke, A., Nikolaus, S., Schroeder, J.O., Vandeputte, D., Raes, J., Szymczak, S., Waetzig, G.H., Zeuner, R. and Rosenstiel, P., 2019. Metabolic Functions of Gut Microbes Associate With Efficacy of Tumor Necrosis Factor Antagonists in Patients With Inflammatory Bowel Diseases. *Gastroenterology*, 157(5), pp.1279-1292.e11.

Aitchison, J., 1982. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2), pp.139–160.

Akalin, A., 2020. *Computational Genomics with R*. Chapman and Hall/CRC.

Akiva, E., Friedlander, G., Itzhaki, Z. and Margalit, H., 2012. A dynamic view of domain-motif interactions. *PLoS Computational Biology*, 8(1), p.e1002341.

Alex, F., ALEX, G., Bertr, RE.GRAMFORTINRIA.F., BERTR, T. and THIRION, n.d. Scikit-learn: Machine Learning in Python.

Alkasir, R., Li, J., Li, X., Jin, M. and Zhu, B., 2017. Human gut microbiota: the links with dementia development. *Protein & cell*, 8(2), pp.90–102.

Alves, L. de F., Westmann, C.A., Lovate, G.L., de Siqueira, G.M.V., Borelli, T.C. and Guazzaroni, M.-E., 2018. Metagenomic approaches for understanding new concepts in microbial science. *International journal of genomics*, 2018, p.2312987.

Anon 2012. *Numerical Ecology*. Developments in environmental modelling. Elsevier.

Anon 2014. Gently Clarifying the Application of Horn's Parallel Analysis to Principal Component Analysis Versus Factor Analysis. *Community Health Faculty Publications and Presentations*.

Anscombe, F.J. and Glynn, W.J., 1983. Distribution of the kurtosis statistic b_2 for normal samples. *Biometrika*, 70(1), pp.227–234.

Arango-Argoty, G., Garner, E., Pruden, A., Heath, L.S., Vikesland, P. and Zhang, L., 2018. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, 6(1), p.23.

Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J.C. and Stegle, O., 2019. MOFA+: a probabilistic framework for comprehensive integration of structured single-cell data. *BioRxiv*.

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W. and Stegle, O., 2018. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6), p.e8124.

Armoni, R. and Borenstein, E., 2022. Temporal alignment of longitudinal microbiome data. *Frontiers in Microbiology*, 13, p.909313.

Armstrong, G., Martino, C., Rahman, G., Gonzalez, A., Vázquez-Baeza, Y., Mishne, G. and Knight, R., 2021. Uniform manifold approximation and projection (UMAP) reveals composite patterns and resolves visualization artifacts in microbiome data. *mSystems*, 6(5), p.e0069121.

Armstrong, G., Rahman, G., Martino, C., McDonald, D., Gonzalez, A., Mishne, G. and Knight, R., 2022. Applications and comparison of dimensionality reduction methods for microbiome data.

Frontiers in Bioinformatics, 2, p.821861.

Bagnall, A., Flynn, M., Large, J., Line, J., Bostrom, A. and Cawley, G., 2018. Is rotation forest the best classifier for problems with continuous features? *arXiv*.

Bajer, L., Kverka, M., Kostovcik, M., Macinga, P., Dvorak, J., Stehlikova, Z., Brezina, J., Wohl, P., Spicak, J. and Drastich, P., 2017. Distinct gut microbiota profiles in patients with primary sclerosing cholangitis and ulcerative colitis. *World Journal of Gastroenterology*, 23(25), pp.4548–4558.

Bakir-Gungor, B., Hacılar, H., Jabeer, A., Nalbantoglu, O.U., Aran, O. and Yousef, M., 2022. Inflammatory bowel disease biomarkers of human gut microbiota selected via different feature selection methods. *PeerJ*, 10, p.e13205.

Balint, D. and Brito, I.L., 2023. Human-gut bacterial protein-protein interactions: understudied but impactful to human health. *Trends in Microbiology*.

Balkenhol, J., Bencurova, E., Gupta, S.K., Schmidt, H., Heinekamp, T., Brakhage, A., Pottikkadavath, A. and Dandekar, T., 2022. Prediction and validation of host-pathogen interactions by a versatile inference approach using Aspergillus fumigatus as a case study. *Computational and structural biotechnology journal*, 20, pp.4225–4237.

Bangsgaard Bendtsen, K.M., Krych, L., Sørensen, D.B., Pang, W., Nielsen, D.S., Josefsen, K., Hansen, L.H., Sørensen, S.J. and Hansen, A.K., 2012. Gut microbiota composition is correlated to grid floor induced stress and behavior in the BALB/c mouse. *Plos One*, 7(10), p.e46231.

Barabási, A.-L., Gulbahce, N. and Loscalzo, J., 2011. Network medicine: a network-based approach to human disease. *Nature Reviews. Genetics*, 12(1), pp.56–68.

Barabási, A.-L. and Oltvai, Z.N., 2004. Network biology: understanding the cell's functional organization. *Nature Reviews. Genetics*, 5(2), pp.101–113.

Barnich, N. and Darfeuille-Michaud, A., 2007. Adherent-invasive Escherichia coli and Crohn's disease. *Current Opinion in Gastroenterology*, 23(1), pp.16–20.

Bauermeister, A., Mannochio-Russo, H., Costa-Lotufo, L.V., Jarmusch, A.K. and Dorrestein, P.C., 2022. Mass spectrometry-based metabolomics in microbiome investigations. *Nature Reviews. Microbiology*, 20(3), pp.143–160.

Bebek, G., Koyutürk, M., Price, N.D. and Chance, M.R., 2012. Network biology methods integrating biological data for translational science. *Briefings in Bioinformatics*, 13(4), pp.446–459.

Beghini, F., McIver, L.J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A.M., Valles-Colomer, M., Weingart, G., Zhang, Y., Zolfo, M., Huttenhower, C., Franzosa, E.A. and Segata, N., 2021. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife*, 10.

Bellman, R., 1966. Dynamic programming. *Science*, 153(3731), pp.34–37.

Bent, S.J., Pierson, J.D., Forney, L.J., Danovaro, R., Luna, G.M., Dell'anno, A. and Pietrangeli, B., 2007. Measuring species richness based on microbial community fingerprints: the emperor has no clothes. *Applied and Environmental Microbiology*, 73(7), pp.2399–401; author reply 2399.

Biagioni, D.J., Astling, D.P., Graf, P. and Davis, M.F., 2011. Orthogonal projection to latent structures solution properties for chemometrics and systems biology data. *Journal of chemometrics*, 25(9), pp.514–525.

Bishop, C.M., 2006. *Pattern recognition and machine learning*. Springer New York.

Biton, A., Bernard-Pierrot, I., Lou, Y., Krucker, C., Chapeaublanc, E., Rubio-Pérez, C., López-Bigas, N., Kamoun, A., Neuzillet, Y., Gestraud, P., Grieco, L., Rebouissou, S., de Reyniès, A., Benhamou, S., Lebret, T., Southgate, J., Barillot, E., Allory, Y., Zinovyev, A. and Radvanyi, F., 2014. Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. *Cell reports*, 9(4), pp.1235–1245.

Bjerrum, J.T., Wang, Y., Hao, F., Coskun, M., Ludwig, C., Günther, U. and Nielsen, O.H., 2015. Metabonomics of human fecal extracts characterize ulcerative colitis, Crohn's disease and healthy individuals. *Metabolomics : Official journal of the Metabolomic Society*, 11, pp.122–133.

Bjerrum, J.T., Wang, Y.L., Seidelin, J.B. and Nielsen, O.H., 2021. IBD metabonomics predicts phenotype, disease course, and treatment response. *EBioMedicine*, 71, p.103551.

Björkqvist, O., Repsilber, D., Seifert, M., Brislawn, C., Jansson, J., Engstrand, L., Rangel, I. and Halfvarson, J., 2019. Alterations in the relative abundance of Faecalibacterium prausnitzii correlate with changes in fecal calprotectin in patients with ileal Crohn's disease: a longitudinal study. *Scandinavian Journal of Gastroenterology*, 54(5), pp.577–585.

Bokulich, N.A., Chung, J., Battaglia, T., Henderson, N., Jay, M., Li, H., D Lieber, A., Wu, F., Perez-Perez, G.I., Chen, Y., Schweizer, W., Zheng, X., Contreras, M., Dominguez-Bello, M.G. and Blaser, M.J., 2016. Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Science Translational Medicine*, 8(343), p.343ra82.

Bray, J.R. and Curtis, J.T., 1957. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological monographs*, 27(4), pp.325–349.

Brito, A.F. and Pinney, J.W., 2017. Protein-Protein Interactions in Virus-Host Systems. *Frontiers in Microbiology*, 8, p.1557.

Bromke, M.A. and Krzystek-Korpacka, M., 2021. Bile acid signaling in inflammatory bowel disease. *International Journal of Molecular Sciences*, 22(16).

Brown, J.R.-M., Flemer, B., Joyce, S.A., Zulquernain, A., Sheehan, D., Shanahan, F. and O'Toole, P.W., 2018. Changes in microbiota composition, bile and fatty acid metabolism, in successful faecal microbiota transplantation for Clostridioides difficile infection. *BMC Gastroenterology*, 18(1), p.131.

Bull, M.J. and Plummer, N.T., 2014. Part 1: The Human Gut Microbiome in Health and Disease. *Integrative medicine (Encinitas, Calif.)*, 13(6), pp.17–22.

Cammarota, G., Ianiro, G., Ahern, A., Carbone, C., Temko, A., Claesson, M.J., Gasbarrini, A. and Tortora, G., 2020. Gut microbiome, big data and machine learning to promote precision medicine for cancer. *Nature Reviews. Gastroenterology & Hepatology*, 17(10), pp.635–648.

Cantini, L., Kairov, U., de Reyniès, A., Barillot, E., Radvanyi, F. and Zinovyev, A., 2019. Assessing reproducibility of matrix factorization methods in independent transcriptomes. *Bioinformatics*, 35(21), pp.4307–4313.

Cao, Y., Shen, J. and Ran, Z.H., 2014. Association between Faecalibacterium prausnitzii Reduction and Inflammatory Bowel Disease: A Meta-Analysis and Systematic Review of the Literature. *Gastroenterology research and practice*, 2014, p.872725.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J. and Knight, R., 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5),

pp.335–336.

Casella, G. and Berger, R.L., 2001. *Statistical Inference*. 2nd ed. Australia: Cengage Learning.p.660.

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C.A., Holland, T.A., Keseler, I.M., Kothari, A., Kubo, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D.S., Weerasinghe, D., Zhang, P. and Karp, P.D., 2014. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research*, 42(Database issue), pp.D459-71.

Cheadle, C., Vawter, M.P., Freed, W.J. and Becker, K.G., 2003. Analysis of Microarray Data Using Z Score Transformation. *The Journal of Molecular Diagnostics*, 5(2), pp.73–81.

Cheng, L., Qi, C., Yang, H., Lu, M., Cai, Y., Fu, T., Ren, J., Jin, Q. and Zhang, X., 2022. gutMGene: a comprehensive database for target genes of gut microbes and microbial metabolites. *Nucleic Acids Research*, 50(D1), pp.D795–D800.

Chen, T. and Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. 22nd ACM SIGKDD International Conference. New York, New York, USA: ACM Press.pp.785–794.

Chen, Y., Lun, A.T.L. and Smyth, G.K., 2016. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. [version 2; peer review: 5 approved]. *F1000Research*, 5, p.1438.

Chiappetta, P., Roubaud, M.C. and Torrésani, B., 2004. Blind source separation and the analysis of microarray data. *Journal of Computational Biology*, 11(6), pp.1090–1109.

Chiu, C.Y. and Miller, S.A., 2019. Clinical metagenomics. *Nature Reviews. Genetics*, 20(6), pp.341–355.

Chuang, H.-Y., Hofree, M. and Ideker, T., 2010. A decade of systems biology. *Annual Review of Cell and Developmental Biology*, 26, pp.721–744.

Clemente, J.C., Ursell, L.K., Parfrey, L.W. and Knight, R., 2012. The impact of the gut microbiota on human health: an integrative view. *Cell*, 148(6), pp.1258–1270.

Clooney, A.G., Eckenberger, J., Laserna-Mendieta, E., Sexton, K.A., Bernstein, M.T., Vagianos, K., Sargent, M., Ryan, F.J., Moran, C., Sheehan, D., Sleator, R.D., Targownik, L.E., Bernstein, C.N., Shanahan, F. and Claesson, M.J., 2021. Ranking microbiome variance in inflammatory bowel disease: a large longitudinal intercontinental study. *Gut*, 70(3), pp.499–510.

Cohen, L.J., Esterhazy, D., Kim, S.-H., Lemetre, C., Aguilar, R.R., Gordon, E.A., Pickard, A.J., Cross, J.R., Emiliano, A.B., Han, S.M., Chu, J., Vila-Farres, X., Kaplitt, J., Rogoz, A., Calle, P.Y., Hunter, C., Bitok, J.K. and Brady, S.F., 2017. Commensal bacteria make GPCR ligands that mimic human signalling molecules. *Nature*, 549(7670), pp.48–53.

Conte, M.P., Schippa, S., Zamboni, I., Penta, M., Chiarini, F., Seganti, L., Osborn, J., Falconieri, P., Borrelli, O. and Cucchiara, S., 2006. Gut-associated bacterial microbiota in paediatric patients with inflammatory bowel disease. *Gut*, 55(12), pp.1760–1767.

Costello, S.P., Soo, W., Bryant, R.V., Jairath, V., Hart, A.L. and Andrews, J.M., 2017. Systematic review with meta-analysis: faecal microbiota transplantation for the induction of remission for active ulcerative colitis. *Alimentary Pharmacology & Therapeutics*, 46(3), pp.213–224.

Coulthard, L.R., White, D.E., Jones, D.L., McDermott, M.F. and Burchill, S.A., 2009. p38(MAPK): stress responses from molecular mechanisms to therapeutics. *Trends in Molecular Medicine*,

15(8), pp.369–379.

Crawford, A.V., Green, S.B., Levy, R., Lo, W.-J., Scott, L., Svetina, D. and Thompson, M.S., 2010. Evaluation of parallel analysis methods for determining the number of factors. *Educational and psychological measurement*, 70(6), pp.885–901.

Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., Mustonen, V., Gonzalez-Perez, A., Pearson, J., Sander, C., Raphael, B.J., Marks, D.S., Ouellette, B.F.F., Valencia, A., Bader, G.D., Boutros, P.C., Stuart, J.M., Linding, R., Lopez-Bigas, N., Stein, L.D. and Mutation Consequences and Pathway Analysis Working Group of the International Cancer Genome Consortium, 2015. Pathway and network analysis of cancer genomes. *Nature Methods*, 12(7), pp.615–621.

Cuadrado, A. and Nebreda, A.R., 2010. Mechanisms and functions of p38 MAPK signalling. *The Biochemical Journal*, 429(3), pp.403–417.

Deek, R.A. and Li, H., 2020. A Zero-Inflated Latent Dirichlet Allocation Model for Microbiome Studies. *Frontiers in genetics*, 11, p.602594.

De Preter, V., Machiels, K., Joossens, M., Arijs, I., Matthys, C., Vermeire, S., Rutgeerts, P. and Verbeke, K., 2015. Faecal metabolite profiling identifies medium-chain fatty acids as discriminating compounds in IBD. *Gut*, 64(3), pp.447–458.

Del Toro, N., Shrivastava, A., Ragueneau, E., Meldal, B., Combe, C., Barrera, E., Perfetto, L., How, K., Ratan, P., Shirodkar, G., Lu, O., Mészáros, B., Watkins, X., Pundir, S., Licata, L., Iannuccelli, M., Pellegrini, M., Martin, M.J., Panni, S., Duesbury, M. and Hermjakob, H., 2022. The IntAct database: efficient access to fine-grained molecular interaction data. *Nucleic Acids Research*, 50(D1), pp.D648–D653.

Dempster, A., Petitjean, F. and Webb, G.I., 2020. ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. *Data mining and knowledge discovery*, 34(5), pp.1454–1495.

Dieterle, F., Ross, A., Schlotterbeck, G. and Senn, H., 2006. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Analytical Chemistry*, 78(13), pp.4281–4290.

Dovrolis, N., Kolios, G., Spyrou, G.M. and Maroulakou, I., 2019. Computational profiling of the gut-brain axis: microflora dysbiosis insights to neurological disorders. *Briefings in Bioinformatics*, 20(3), pp.825–841.

Duncan, K., Carey-Ewend, K. and Vaishnava, S., 2021. Spatial analysis of gut microbiome reveals a distinct ecological niche associated with the mucus layer. *Gut microbes*, 13(1), p.1874815.

Eckburg, P.B., Bik, E.M., Bernstein, C.N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S.R., Nelson, K.E. and Relman, D.A., 2005. Diversity of the human intestinal microbial flora. *Science*, 308(5728), pp.1635–1638.

Engreitz, J.M., Daigle, B.J., Marshall, J.J. and Altman, R.B., 2010. Independent component analysis: mining microarray data for fundamental human gene expression modules. *Journal of Biomedical Informatics*, 43(6), pp.932–944.

Erickson, A.R., Cantarel, B.L., Lamendella, R., Darzi, Y., Mongodin, E.F., Pan, C., Shah, M., Halfvarson, J., Tysk, C., Henrissat, B., Raes, J., Verberkmoes, N.C., Fraser, C.M., Hettich, R.L. and Jansson, J.K., 2012. Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *Plos One*, 7(11), p.e49138.

Faith, M., 2015. Centered Log-Ratio (clr) Transformation and Robust Principal Component Analysis of Long-Term NDVI Data Reveal Vegetation Activity Linked to Climate Processes. *Climate*, 3(1), pp.135–149.

Filyk, H.A. and Osborne, L.C., 2016. The multibiome: the intestinal ecosystem's influence on immune homeostasis, health, and disease. *EBioMedicine*, 13, pp.46–54.

Franzosa, E.A., McIver, L.J., Rahnavard, G., Thompson, L.R., Schirmer, M., Weingart, G., Lipson, K.S., Knight, R., Caporaso, J.G., Segata, N. and Huttenhower, C., 2018. Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods*, 15(11), pp.962–968.

Franzosa, E.A., Sirota-Madi, A., Avila-Pacheco, J., Fornelos, N., Haiser, H.J., Reinker, S., Vatanen, T., Hall, A.B., Mallick, H., McIver, L.J., Sauk, J.S., Wilson, R.G., Stevens, B.W., Scott, J.M., Pierce, K., Deik, A.A., Bullock, K., Imhann, F., Porter, J.A., Zhernakova, A. and Xavier, R.J., 2019. Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nature Microbiology*, 4(2), pp.293–305.

Gallagher, K., Catesson, A., Griffin, J.L., Holmes, E. and Williams, H.R.T., 2021. Metabolomic analysis in inflammatory bowel disease: A systematic review. *Journal of Crohn's & colitis*, 15(5), pp.813–826.

Garrett, W.S., Gallini, C.A., Yatsunenko, T., Michaud, M., DuBois, A., Delaney, M.L., Punit, S., Karlsson, M., Bry, L., Glickman, J.N., Gordon, J.I., Onderdonk, A.B. and Glimcher, L.H., 2010. Enterobacteriaceae act in concert with the gut microbiota to induce spontaneous and maternally transmitted colitis. *Cell Host & Microbe*, 8(3), pp.292–300.

Gibbons, R.J., Socransky, S.S., Dearaujo, W.C. and Vanhoute, J., 1964. Studies of the predominant cultivable microbiota of dental plaque. *Archives of Oral Biology*, 9, pp.365–370.

Gibbons, S.M., Kearney, S.M., Smillie, C.S. and Alm, E.J., 2017. Two dynamic regimes in the human gut microbiome. *PLoS Computational Biology*, 13(2), p.e1005364.

Gieryńska, M., Szulc-Dąbrowska, L., Struzik, J., Mielcarska, M.B. and Gregorczyk-Zboroch, K.P., 2022. Integrity of the Intestinal Barrier: The Involvement of Epithelial Cells and Microbiota-A Mutual Relationship. *Animals : an open access journal from MDPI*, 12(2).

Glassner, K.L., Abraham, B.P. and Quigley, E.M.M., 2020. The microbiome and inflammatory bowel disease. *The Journal of Allergy and Clinical Immunology*, 145(1), pp.16–27.

Gloor, G.B. and Reid, G., 2016. Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. *Canadian Journal of Microbiology*, 62(8), pp.692–703.

Glorfeld, L.W., 1995. An improvement on horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and psychological measurement*, 55(3), pp.377–393.

Gosak, M., Markovič, R., Dolenšek, J., Slak Rupnik, M., Marhl, M., Stožer, A. and Perc, M., 2018. Network science of biological systems at different scales: A review. *Physics of life reviews*, 24, pp.118–135.

Greenacre, M., Martínez-Álvaro, M. and Blasco, A., 2021. Compositional Data Analysis of Microbiome and Any-Omics Datasets: A Validation of the Additive Logratio Transformation. *Frontiers in Microbiology*, 12, p.727398.

Gregorius, H. and Kosman, E., 2017. On the notion of dispersion: from dispersion to diversity. *Methods in Ecology and Evolution*, 8(3), pp.278–287.

Guardamagna, M., Berciano-Guerrero, M.-A., Villaescusa-González, B., Perez-Ruiz, E., Oliver, J., Lavado-Valenzuela, R., Rueda-Dominguez, A., Barragán, I. and Queipo-Ortuño, M.I., 2022. Gut microbiota and therapy in metastatic melanoma: focus on MAPK pathway inhibition. *International Journal of Molecular Sciences*, 23(19).

Guijas, C., Montenegro-Burke, J.R., Warth, B., Spilker, M.E. and Siuzdak, G., 2018. Metabolomics activity screening for identifying metabolites that modulate phenotype. *Nature Biotechnology*, 36(4), pp.316–320.

Guo, Y.-J., Pan, W.-W., Liu, S.-B., Shen, Z.-F., Xu, Y. and Hu, L.-L., 2020. ERK/MAPK signalling pathway and tumorigenesis. *Experimental and therapeutic medicine*, 19(3), pp.1997–2007.

Guzior, D.V. and Quinn, R.A., 2021. Review: microbial transformations of human bile acids. *Microbiome*, 9(1), p.140.

Haran, J.P., Bhattarai, S.K., Foley, S.E., Dutta, P., Ward, D.V., Bucci, V. and McCormick, B.A., 2019. Alzheimer's Disease Microbiome Is Associated with Dysregulation of the Anti-Inflammatory P-Glycoprotein Pathway. *mBio*, 10(3).

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., Del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P. and Oliphant, T.E., 2020. Array programming with NumPy. *Nature*, 585(7825), pp.357–362.

Hart, T., Komori, H.K., LaMere, S., Podshivalova, K. and Salomon, D.R., 2013. Finding the active genes in deep RNA-seq gene expression studies. *BMC Genomics*, 14, p.778.

Harvey, R.F. and Bradshaw, J.M., 1980. A simple index of Crohn's-disease activity. *The Lancet*, 1(8167), p.514.

Hassouneh, S.A.-D., Loftus, M. and Yooseph, S., 2021. Linking inflammatory bowel disease symptoms to changes in the gut microbiome structure and function. *Frontiers in Microbiology*, 12, p.673632.

Heinken, A., Ravcheev, D.A., Baldini, F., Heirendt, L., Fleming, R.M.T. and Thiele, I., 2019. Systematic assessment of secondary bile acid metabolism in gut microbes reveals distinct metabolic capabilities in inflammatory bowel disease. *Microbiome*, 7(1), p.75.

Heinken, A. and Thiele, I., 2015. Systems biology of host-microbe metabolomics. *Wiley interdisciplinary reviews. Systems biology and medicine*, 7(4), pp.195–219.

Herault, J. and Jutten, C., 1986. Space or time adaptive signal processing by neural network models. In: *AIP Conference Proceedings*. AIP Conference Proceedings Volume 151. AIP.pp.206–211.

He, Q., Gao, Y., Jie, Z., Yu, X., Laursen, J.M., Xiao, L., Li, Y., Li, L., Zhang, F., Feng, Q., Li, X., Yu, J., Liu, C., Lan, P., Yan, T., Liu, X., Xu, X., Yang, H., Wang, J., Madsen, L. and Jia, H., 2017. Two distinct metacommunities characterize the gut microbiota in Crohn's disease patients. *GigaScience*, 6(7), pp.1–11.

Hildebrand, F., 2021. Ultra-resolution Metagenomics: When Enough Is Not Enough. *mSystems*, p.e0088121.

Hildebrand, F., Moitinho-Silva, L., Blasche, S., Jahn, M.T., Gossmann, T.I., Huerta-Cepas, J., Hercog, R., Luetge, M., Bahram, M., Pryszlak, A., Alves, R.J., Waszak, S.M., Zhu, A., Ye, L., Costea, P.I., Aalvink, S., Belzer, C., Forslund, S.K., Sunagawa, S., Hentschel, U. and Bork, P., 2019. Antibiotics-induced monodominance of a novel gut bacterial order. *Gut*, 68(10), pp.1781–1790.

Hill, J.M., Clement, C., Pogue, A.I., Bhattacharjee, S., Zhao, Y. and Lukiw, W.J., 2014. Pathogenic microbes, the microbiome, and Alzheimer's disease (AD). *Frontiers in aging neuroscience*, 6, p.127.

Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. *Neural Computation*, 9(8), pp.1735–1780.

Hoffman, M.D. and Gelman, A., 2014. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*.

Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., Oda, Y., Kakazu, Y., Kusano, M., Tohge, T., Matsuda, F., Sawada, Y., Hirai, M.Y., Nakanishi, H., Ikeda, K., Akimoto, N. and Nishioka, T., 2010. MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, 45(7), pp.703–714.

Horn, J.L., 1965. A RATIONALE AND TEST FOR THE NUMBER OF FACTORS IN FACTOR ANALYSIS. *Psychometrika*, 30, pp.179–185.

Hou, K., Wu, Z.-X., Chen, X.-Y., Wang, J.-Q., Zhang, D., Xiao, C., Zhu, D., Koya, J.B., Wei, L., Li, J. and Chen, Z.-S., 2022. Microbiota in health and diseases. *Signal transduction and targeted therapy*, 7(1), p.135.

Huang, G., Shi, L.Z. and Chi, H., 2009. Regulation of JNK and p38 MAPK in the immune system: signal integration, propagation and termination. *Cytokine*, 48(3), pp.161–169.

Huang, X.L., Xu, J., Zhang, X.H., Qiu, B.Y., Peng, L., Zhang, M. and Gan, H.T., 2011. PI3K/Akt signaling pathway is involved in the pathogenesis of ulcerative colitis. *Inflammation Research*, 60(8), pp.727–734.

Human Microbiome Project Consortium, 2012. A framework for human microbiome research. *Nature*, 486(7402), pp.215–221.

Hur, E.M. and Kim, K.T., 2002. G protein-coupled receptor signalling and cross-talk: achieving rapidity and specificity. *Cellular Signalling*, 14(5), pp.397–405.

Hyvärinen, A. and Oja, E., 2000. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4–5), pp.411–430.

Idrees, S., Pérez-Bercoff, Å. and Edwards, R.J., 2018. SLiM-Enrich: computational assessment of protein-protein interaction data as a source of domain-motif interactions. *PeerJ*, 6, p.e5858.

Imker, H.J., 2018. 25 years of molecular biology databases: A study of proliferation, impact, and maintenance. *Frontiers in Research Metrics and Analytics*, 3.

Integrative HMP (iHMP) Research Network Consortium, 2014. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host & Microbe*, 16(3), pp.276–289.

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L. and Muller, P.-A., 2019. Deep learning for time series classification: a review. *Data mining and knowledge discovery*, pp.1–47.

Issa Isaac, N., Philippe, D., Nicholas, A., Raoult, D. and Eric, C., 2019. Metaproteomics of the human gut microbiota: Challenges and contributions to other OMICS. *Clinical mass spectrometry (Del Mar, Calif.)*, 14 Pt A, pp.18–30.

Itzhaki, Z., Akiva, E., Altuvia, Y. and Margalit, H., 2006. Evolutionary conservation of domain-domain interactions. *Genome Biology*, 7(12), p.R125.

Jebara, T., 2004. *Machine Learning*. Boston, MA: Springer US.

Johnson, C.H. and Gonzalez, F.J., 2012. Challenges and opportunities of metabolomics. *Journal of Cellular Physiology*, 227(8), pp.2975–2981.

Johnson, C.H., Ivanisevic, J. and Siuzdak, G., 2016. Metabolomics: beyond biomarkers and towards mechanisms. *Nature Reviews. Molecular Cell Biology*, 17(7), pp.451–459.

Joseph, T.A., Pasarkar, A.P. and Pe'er, I., 2020. Efficient and accurate inference of microbial trajectories from longitudinal count data. *BioRxiv*.

Juez-Gil, M., Arnaiz-González, Á., Rodríguez, J.J., López-Nozal, C. and García-Osorio, C., 2021. Rotation forest for big data. *Information Fusion*, 74, pp.39–49.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T. and Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), pp.583–589.

Kalathur, R.K.R., Pinto, J.P., Hernández-Prieto, M.A., Machado, R.S.R., Almeida, D., Chaurasia, G. and Futschik, M.E., 2014. UniHI 7: an enhanced database for retrieval and interactive analysis of human molecular interaction networks. *Nucleic Acids Research*, 42(Database issue), pp.D408-14.

Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. and Ishiguro-Watanabe, M., 2023. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research*, 51(D1), pp.D587–D592.

Kent, J.T., 1983. Information gain and a general measure of correlation. *Biometrika*, 70(1), pp.163–173.

Kim, T., Lee, I. and Lee, T.-W., 2006. Independent vector analysis: definition and algorithms. In: *2006 Fortieth Asilomar Conference on Signals, Systems and Computers*. 2006 Fortieth Asilomar Conference on Signals, Systems and Computers. IEEE.pp.1393–1396.

Klimovskaia, A., Lopez-Paz, D., Bottou, L. and Nickel, M., 2020. Poincaré maps for analyzing complex hierarchies in single-cell data. *Nature Communications*, 11(1), p.2966.

Kobayashi, T., Siegmund, B., Le Berre, C., Wei, S.C., Ferrante, M., Shen, B., Bernstein, C.N., Danese, S., Peyrin-Biroulet, L. and Hibi, T., 2020. Ulcerative colitis. *Nature reviews. Disease primers*, 6(1), p.74.

Kodikara, S., Ellul, S. and Lê Cao, K.-A., 2022. Statistical challenges in longitudinal microbiome data analysis. *Briefings in Bioinformatics*, 23(4).

Kolmeder, C.A. and de Vos, W.M., 2014. Metaproteomics of our microbiome - developing insight in function and activity in man and model systems. *Journal of Proteomics*, 97, pp.3–16.

Kotlowski, R., Bernstein, C.N., Sepehri, S. and Krause, D.O., 2007. High prevalence of Escherichia coli belonging to the B2+D phylogenetic group in inflammatory bowel disease. *Gut*, 56(5), pp.669–675.

Krumsiek, J., Suhre, K., Illig, T., Adamski, J. and Theis, F.J., 2012. Bayesian independent component analysis recovers pathway signatures from blood metabolomics data. *Journal of Proteome Research*, 11(8), pp.4120–4131.

Kubinski, R., Djamen-Kepaou, J.-Y., Zhanabaev, T., Hernandez-Garcia, A., Bauer, S., Hildebrand, F., Korcsmaros, T., Karam, S., Jantchou, P., Kafi, K. and Martin, R.D., 2022. Benchmark of Data Processing Methods and Machine Learning Models for Gut Microbiome-Based Diagnosis of Inflammatory Bowel Disease. *Frontiers in genetics*, 13, p.784397.

Kumar, M., Michael, S., Alvarado-Valverde, J., Mészáros, B., Sámano-Sánchez, H., Zeke, A., Dobson, L., Lazar, T., Örd, M., Nagpal, A., Farahi, N., Käser, M., Kraleti, R., Davey, N.E., Pancsa, R., Chemes, L.B. and Gibson, T.J., 2022. The Eukaryotic Linear Motif resource: 2022 release. *Nucleic Acids Research*, 50(D1), pp.D497–D508.

Laccourreye, P., Bielza, C. and Larrañaga, P., 2022. Explainable Machine Learning for Longitudinal Multi-Omic Microbiome. *Mathematics*, 10(12), p.1994.

Lamb, C.A., Kennedy, N.A., Raine, T., Hendy, P.A., Smith, P.J., Limdi, J.K., Hayee, B., Lomer, M.C.E., Parkes, G.C., Selinger, C., Barrett, K.J., Davies, R.J., Bennett, C., Gittens, S., Dunlop, M.G., Faiz, O., Fraser, A., Garrick, V., Johnston, P.D., Parkes, M. and Hawthorne, A.B., 2019. British Society of Gastroenterology consensus guidelines on the management of inflammatory bowel disease in adults. *Gut*, 68(Suppl 3), pp.s1–s106.

Langille, M.G.I., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J.A., Clemente, J.C., Burkepile, D.E., Vega Thurber, R.L., Knight, R., Beiko, R.G. and Huttenhower, C., 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, 31(9), pp.814–821.

Lavelle, A. and Sokol, H., 2020. Gut microbiota-derived metabolites as key actors in inflammatory bowel disease. *Nature Reviews. Gastroenterology & Hepatology*, 17(4), pp.223–237.

Layeghifard, M., Li, H., Wang, P.W., Donaldson, S.L., Coburn, B., Clark, S.T., Caballero, J.D., Zhang, Y., Tullis, D.E., Yau, Y.C.W., Waters, V., Hwang, D.M. and Guttman, D.S., 2019. Microbiome networks and change-point analysis reveal key community changes associated with cystic fibrosis pulmonary exacerbations. *npj Biofilms and Microbiomes*, 5(1), p.4.

Lebeer, S., Vanderleyden, J. and De Keersmaecker, S.C.J., 2010. Host interactions of probiotic bacterial surface molecules: comparison with commensals and pathogens. *Nature Reviews. Microbiology*, 8(3), pp.171–184.

Lee, M. and Chang, E.B., 2021. Inflammatory Bowel Diseases (IBD) and the Microbiome-Searching the Crime Scene for Clues. *Gastroenterology*, 160(2), pp.524–537.

Lee, P.Y., Chin, S.-F., Neoh, H.-M. and Jamal, R., 2017. Metaproteomic analysis of human gut microbiota: where are we heading? *Journal of Biomedical Science*, 24(1), p.36.

Lehmann, T., Schallert, K., Vilchez-Vargas, R., Benndorf, D., Püttker, S., Sydor, S., Schulz, C., Bechmann, L., Canbay, A., Heidrich, B., Reichl, U., Link, A. and Heyer, R., 2019. Metaproteomics of fecal samples of Crohn's disease and Ulcerative Colitis. *Journal of Proteomics*, 201, pp.93–103.

Ley, R.E., Turnbaugh, P.J., Klein, S. and Gordon, J.I., 2006. Microbial ecology: human gut microbes associated with obesity. *Nature*, 444(7122), pp.1022–1023.

Lima, S.F., Gogokhia, L., Viladomiu, M., Chou, L., Putzel, G., Jin, W.-B., Pires, S., Guo, C.-J., Gerardin, Y., Crawford, C.V., Jacob, V., Scherl, E., Brown, S.-E., Hambor, J. and Longman, R.S., 2022. Transferable Immunoglobulin A-Coated Odoribacter splanchnicus in Responders to Fecal Microbiota Transplantation for Ulcerative Colitis Limits Colonic Inflammation. *Gastroenterology*, 162(1), pp.166–178.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S. and Rives, A., 2022. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*.

Liu, F., Yang, X., Geng, M. and Huang, M., 2018. Targeting ERK, an Achilles' Heel of the MAPK pathway, in cancer therapy. *Acta pharmaceutica Sinica. B*, 8(4), pp.552–562.

Liu, Y., Smirnov, K., Lucio, M., Gougeon, R.D., Alexandre, H. and Schmitt-Kopplin, P., 2016. MetICA: independent component analysis for high-resolution mass-spectrometry based non-targeted metabolomics. *BMC Bioinformatics*, 17, p.114.

Li, J., Shen, X. and Li, Y., 2021. Modeling the temporal dynamics of gut microbiota from a local community perspective. *Ecological Modelling*, 460, p.109733.

Litvak, Y., Byndloss, M.X. and Bäumler, A.J., 2018. Colonocyte metabolism shapes the gut microbiota. *Science*, 362(6418).

Liu, L., Xu, M., Lan, R., Hu, D., Li, X., Qiao, L., Zhang, S., Lin, X., Yang, J., Ren, Z. and Xu, J., 2022. Bacteroides vulgatus attenuates experimental mice colitis through modulating gut microbiota and immune responses. *Frontiers in Immunology*, 13, p.1036196.

Liu, Y.-Y., 2023. Controlling the human microbiome. *Cell Systems*, 14(2), pp.135–159.

Lloyd-Price, J., Arze, C., Ananthakrishnan, A.N., Schirmer, M., Avila-Pacheco, J., Poon, T.W., Andrews, E., Ajami, N.J., Bonham, K.S., Brislawn, C.J., Casero, D., Courtney, H., Gonzalez, A., Graeber, T.G., Hall, A.B., Lake, K., Landers, C.J., Mallick, H., Plichta, D.R., Prasad, M. and Huttenhower, C., 2019. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 569(7758), pp.655–662.

Lopez-Siles, M., Duncan, S.H., Garcia-Gil, L.J. and Martinez-Medina, M., 2017. Faecalibacterium prausnitzii: from microbiology to diagnostics and prognostics. *The ISME Journal*, 11(4), pp.841–852.

Love, M.I., Huber, W. and Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), p.550.

Luca, F., Kupfer, S.S., Knights, D., Khoruts, A. and Blekhman, R., 2018. Functional Genomics of Host-Microbiome Interactions in Humans. *Trends in Genetics*, 34(1), pp.30–40.

Lugo-Martinez, J., Ruiz-Perez, D., Narasimhan, G. and Bar-Joseph, Z., 2019. Dynamic interaction network inference from longitudinal microbiome data. *Microbiome*, 7(1), p.54.

Luna, P.N., Mansbach, J.M. and Shaw, C.A., 2020. A joint modeling approach for longitudinal microbiome data improves ability to detect microbiome associations with disease. *PLoS Computational Biology*, 16(12), p.e1008473.

van der Maaten, L. and Hinton, G., 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*.

Machiels, K., Joossens, M., Sabino, J., De Preter, V., Arijs, I., Eeckhaut, V., Ballet, V., Claes, K., Van Immerseel, F., Verbeke, K., Ferrante, M., Verhaegen, J., Rutgeerts, P. and Vermeire, S., 2014. A decrease of the butyrate-producing species Roseburia hominis and Faecalibacterium prausnitzii defines dysbiosis in patients with ulcerative colitis. *Gut*, 63(8), pp.1275–1283.

Malla, M.A., Dubey, A., Kumar, A., Yadav, S., Hashem, A. and Abd Allah, E.F., 2018. Exploring the Human Microbiome: The Potential Future Role of Next-Generation Sequencing in Disease Diagnosis and Treatment. *Frontiers in Immunology*, 9, p.2868.

Mandal, S., Van Treuren, W., White, R.A., Eggesbø, M., Knight, R. and Peddada, S.D., 2015. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial ecology in health and disease*, 26, p.27663.

Manichanh, C., Borruel, N., Casellas, F. and Guarner, F., 2012. The gut microbiota in IBD. *Nature Reviews. Gastroenterology & Hepatology*, 9(10), pp.599–608.

Manor, O., Dai, C.L., Kornilov, S.A., Smith, B., Price, N.D., Lovejoy, J.C., Gibbons, S.M. and Magis, A.T., 2020. Health and disease markers correlate with gut microbiome composition across thousands of people. *Nature Communications*, 11(1), p.5206.

Marchesi, J.R., Holmes, E., Khan, F., Kochhar, S., Scanlan, P., Shanahan, F., Wilson, I.D. and Wang, Y., 2007. Rapid and noninvasive metabonomic characterization of inflammatory bowel disease. *Journal of Proteome Research*, 6(2), pp.546–551.

Mark Welch, J.L., Hasegawa, Y., McNulty, N.P., Gordon, J.I. and Borisy, G.G., 2017. Spatial organization of a model 15-member human gut microbiota established in gnotobiotic mice. *Proceedings of the National Academy of Sciences of the United States of America*, 114(43), pp.E9105–E9114.

Martino, C., Morton, J.T., Marotz, C.A., Thompson, L.R., Tripathi, A., Knight, R. and Zengler, K., 2019. A novel sparse compositional technique reveals microbial perturbations. *mSystems*, 4(1).

Martin, B.D., Witten, D. and Willis, A.D., 2020. Modeling microbial abundances and dysbiosis with beta-binomial regression. *The annals of applied statistics*, 14(1), pp.94–115.

Ma, S. and Dai, Y., 2011. Principal component analysis based methods in bioinformatics studies. *Briefings in Bioinformatics*, 12(6), pp.714–722.

Martino, C., Shenhav, L., Marotz, C.A., Armstrong, G., McDonald, D., Vázquez-Baeza, Y., Morton, J.T., Jiang, L., Dominguez-Bello, M.G., Swafford, A.D., Halperin, E. and Knight, R., 2021. Context-aware dimensionality reduction deconvolutes gut microbial community dynamics. *Nature Biotechnology*, 39(2), pp.165–168.

Massimino, L., Lamparelli, L.A., Houshyar, Y., D'Alessio, S., Peyrin-Biroulet, L., Vetrano, S., Danese, S. and Ungaro, F., 2021. The Inflammatory Bowel Disease Transcriptome and Metatranscriptome Meta-Analysis (IBD TaMMA) framework. *Nature Computational Science*, 1(8), pp.511–515.

McCarthy, D.J., Chen, Y. and Smyth, G.K., 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10), pp.4288–4297.

McInnes, L., Healy, J., Saul, N. and Großberger, L., 2018. UMAP: uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29), p.861.

McMurdie, P.J. and Holmes, S., 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *Plos One*, 8(4), p.e61217.

Mendes-Soares, H., Mundy, M., Soares, L.M. and Chia, N., 2016. MMinte: an application for predicting metabolic interactions among the microbial species in a community. *BMC Bioinformatics*, 17(1), p.343.

Mills, R.H., Dulai, P.S., Vázquez-Baeza, Y., Sauceda, C., Daniel, N., Gerner, R.R., Batachari, L.E., Malfavon, M., Zhu, Q., Weldon, K., Humphrey, G., Carrillo-Terrazas, M., Goldasich, L.D., Bryant, M., Raffatellu, M., Quinn, R.A., Gewirtz, A.T., Chassaing, B., Chu, H., Sandborn, W.J. and Gonzalez, D.J., 2022. Multi-omics analyses of the ulcerative colitis gut microbiome link Bacteroides vulgatus proteases with disease severity. *Nature Microbiology*, 7(2), pp.262–276.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., Finn, R.D. and Bateman, A., 2021. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1), pp.D412–D419.

Mohseni, A.H., Casolaro, V., Bermúdez-Humarán, L.G., Keyvani, H. and Taghinezhad-S, S., 2021.

Modulation of the PI3K/Akt/mTOR signaling pathway by probiotics as a fruitful target for orchestrating the immune response. *Gut microbes*, 13(1), pp.1–17.

Moldakarimov, S. and Sejnowski, T.J., 2017. Neural computation theories of learning ☆. In: *Learning and memory: A comprehensive reference*. Elsevier.pp.579–589.

Mor, U., Cohen, Y., Valdés-Mas, R., Kviatcovsky, D., Elinav, E. and Avron, H., 2022. Dimensionality reduction of longitudinal 'omics data using modern tensor factorizations. *PLoS Computational Biology*, 18(7), p.e1010212.

Mosca, R., Céol, A., Stein, A., Olivella, R. and Aloy, P., 2014. 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Research*, 42(Database issue), pp.D374-9.

Moschen, A.R., Gerner, R.R., Wang, J., Klepsch, V., Adolph, T.E., Reider, S.J., Hackl, H., Pfister, A., Schilling, J., Moser, P.L., Kempster, S.L., Swidsinski, A., Orth Höller, D., Weiss, G., Baines, J.F., Kaser, A. and Tilg, H., 2016. Lipocalin 2 Protects from Inflammation and Tumorigenesis Associated with Gut Microbiota Alterations. *Cell Host & Microbe*, 19(4), pp.455–469.

Nayfach, S., Shi, Z.J., Seshadri, R., Pollard, K.S. and Kyrpides, N.C., 2019. New insights from uncultivated genomes of the global human gut microbiome. *Nature*, 568(7753), pp.505–510.

Nazarov, P.V., Wienecke-Baldacchino, A.K., Zinovyev, A., Czerwinska, U., Muller, A., Nashan, D., Dittmar, G., Azuaje, F. and Kreis, S., 2018. Independent component analysis provides clinically relevant insights into the biology of melanoma patients. *BioRxiv*.

Nguyen, Q.P., Karagas, M.R., Madan, J.C., Dade, E., Palys, T.J., Morrison, H.G., Pathmasiri, W.W., McRitche, S., Sumner, S.J., Frost, H.R. and Hoen, A.G., 2021. Associations between the gut microbiome and metabolome in early life. *BMC Microbiology*, 21(1), p.238.

Nicholson, J.K., Holmes, E., Kinross, J., Burcelin, R., Gibson, G., Jia, W. and Pettersson, S., 2012. Host-gut microbiota metabolic interactions. *Science*, 336(6086), pp.1262–1267.

Nikolaus, S., Schulte, B., Al-Massad, N., Thieme, F., Schulte, D.M., Bethge, J., Rehman, A., Tran, F., Aden, K., Häsler, R., Moll, N., Schütze, G., Schwarz, M.J., Waetzig, G.H., Rosenstiel, P., Krawczak, M., Szymczak, S. and Schreiber, S., 2017. Increased tryptophan metabolism is associated with activity of inflammatory bowel diseases. *Gastroenterology*, 153(6), pp.1504-1516.e2.

Oliphant, K. and Allen-Vercoe, E., 2019. Macronutrient metabolism by the human gut microbiome: major fermentation by-products and their impact on host health. *Microbiome*, 7(1), p.91.

Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.-J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F., Dolma, S., Coulombe-Huntington, J., Chatr-Aryamontri, A., Dolinski, K. and Tyers, M., 2021. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science*, 30(1), pp.187–200.

Palmela, C., Chevarin, C., Xu, Z., Torres, J., Sevrin, G., Hirten, R., Barnich, N., Ng, S.C. and Colombel, J.-F., 2018. Adherent-invasive Escherichia coli in inflammatory bowel disease. *Gut*, 67(3), pp.574–587.

Parada Venegas, D., De la Fuente, M.K., Landskron, G., González, M.J., Quera, R., Dijkstra, G., Harmsen, H.J.M., Faber, K.N. and Hermoso, M.A., 2019. Short Chain Fatty Acids (SCFAs)-Mediated Gut Epithelial and Immune Regulation and Its Relevance for Inflammatory Bowel Diseases. *Frontiers in Immunology*, 10, p.277.

Parker, B.J., Wearsch, P.A., Veloo, A.C.M. and Rodriguez-Palacios, A., 2020. The genus alistipes: gut bacteria with emerging implications to inflammation, cancer, and mental health. *Frontiers in Immunology*, 11, p.906.

Parker, R.B. and Snyder, M.L., 1961. Interactions of the oral microbiota I. A system for the defined study of mixed cultures. *Experimental biology and medicine*, 108(3), pp.749–752.

Patti, G.J., Yanes, O. and Siuzdak, G., 2012. Innovation: Metabolomics: the apogee of the omics trilogy. *Nature Reviews. Molecular Cell Biology*, 13(4), pp.263–269.

Paull, E.O., Carlin, D.E., Niepel, M., Sorger, P.K., Haussler, D. and Stuart, J.M., 2013. Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics*, 29(21), pp.2757–2764.

Paulson, J.N., Stine, O.C., Bravo, H.C. and Pop, M., 2013. Differential abundance analysis for microbial marker-gene surveys. *Nature Methods*, 10(12), pp.1200–1202.

Pavlidis, P., Gulati, S., Dubois, P., Chung-Faye, G., Sherwood, R., Bjarnason, I. and Hayee, B., 2016. Early change in faecal calprotectin predicts primary non-response to anti-TNFα therapy in Crohn's disease. *Scandinavian Journal of Gastroenterology*, 51(12), pp.1447–1452.

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, 2(11), pp.559–572.

Petersen, A.M., Nielsen, E.M., Litrup, E., Brynskov, J., Mirsepasi, H. and Krogfelt, K.A., 2009. A phylogenetic group of Escherichia coli associated with active left-sided inflammatory bowel disease. *BMC Microbiology*, 9, p.171.

Petriz, B.A. and Franco, O.L., 2017. Metaproteomics as a complementary approach to gut microbiota in health and disease. *Frontiers in chemistry*, 5, p.4.

Petrosino, J.F., 2018. The microbiome in precision medicine: the way forward. *Genome Medicine*, 10(1), p.12.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., NAEISPACK authors, Heisterkamp, S., Van Willigen, B., Ranke, J. and R Core Team, 2023. *Linear and Nonlinear Mixed Effects Models, R package nlme version 3.1-162*. RCAN.

Pinheiro, J.C. and Bates, D.M., 2000. *Mixed-Effects Models in S and S-PLUS*. New York: Springer-Verlag.

Pratt, M., Forbes, J.D., Knox, N.C., Bernstein, C.N. and Van Domselaar, G., 2021. Microbiome-Mediated Immune Signaling in Inflammatory Bowel Disease and Colorectal Cancer: Support From Meta-omics Data. *Frontiers in cell and developmental biology*, 9, p.716604.

Quévrain, E., Maubert, M.A., Michon, C., Chain, F., Marquant, R., Tailhades, J., Miquel, S., Carlier, L., Bermúdez-Humarán, L.G., Pigneur, B., Lequin, O., Kharrat, P., Thomas, G., Rainteau, D., Aubry, C., Breyner, N., Afonso, C., Lavielle, S., Grill, J.P., Chassaing, G. and Seksik, P., 2016. Identification of an anti-inflammatory protein from Faecalibacterium prausnitzii, a commensal bacterium deficient in Crohn's disease. *Gut*, 65(3), pp.415–425.

Quinn, T.P., Erb, I., Gloor, G., Notredame, C., Richardson, M.F. and Crowley, T.M., 2019. A field guide for the compositional analysis of any-omics data. *GigaScience*, 8(9).

Rabizadeh, S., Rhee, K.-J., Wu, S., Huso, D., Gan, C.M., Golub, J.E., Wu, X., Zhang, M. and Sears, C.L., 2007. Enterotoxigenic bacteroides fragilis: a potential instigator of colitis. *Inflammatory Bowel Diseases*, 13(12), pp.1475–1483.

Rashid, T. and Ebringer, A., 2011. Gut-mediated and HLA-B27-associated arthritis: an emphasis on ankylosing spondylitis and Crohn's disease with a proposal for the use of new treatment. *Discovery medicine*, 12(64), pp.187–194.

Rashid, T., Ebringer, A. and Wilson, C., 2013. The role of Klebsiella in Crohn's disease with a potential for the use of antimicrobial measures. *International journal of rheumatology*, 2013, p.610393.

Rehman, A., Lepage, P., Nolte, A., Hellmig, S., Schreiber, S. and Ott, S.J., 2010. Transcriptional activity of the dominant gut mucosal microbiota in chronic inflammatory bowel disease patients. *Journal of Medical Microbiology*, 59(Pt 9), pp.1114–1122.

Ricotta, C., 2021. From the euclidean distance to compositional dissimilarity: What is gained and what is lost. *Acta Oecologica*, 111, p.103732.

Ridlon, J.M., Kang, D.J., Hylemon, P.B. and Bajaj, J.S., 2014. Bile acids and the gut microbiome. *Current Opinion in Gastroenterology*, 30(3), pp.332–338.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J. and Fergus, R., 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 118(15).

Robinson, M.D., McCarthy, D.J. and Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), pp.139–140.

Roda, G., Chien Ng, S., Kotze, P.G., Argollo, M., Panaccione, R., Spinelli, A., Kaser, A., Peyrin-Biroulet, L. and Danese, S., 2020. Crohn's disease. *Nature reviews. Disease primers*, 6(1), p.22.

Roda, G., Porru, E., Katsanos, K., Skamnelos, A., Kyriakidi, K., Fiorino, G., Christodoulou, D., Danese, S. and Roda, A., 2019. Serum Bile Acids Profiling in Inflammatory Bowel Disease Patients Treated with Anti-TNFs. *Cells*, 8(8).

Samarkos, M., Mastrogianni, E. and Kampouropoulou, O., 2018. The role of gut microbiota in Clostridium difficile infection. *European Journal of Internal Medicine*, 50, pp.28–32.

Sankaran, K. and Holmes, S.P., 2019. Latent variable modeling for the microbiome. *Biostatistics*, 20(4), pp.599–614.

Sartor, R.B., 2006. Mechanisms of disease: pathogenesis of Crohn's disease and ulcerative colitis. *Nature Clinical Practice. Gastroenterology & Hepatology*, 3(7), pp.390–407.

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M.G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J. and Yau, C., 2021. Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1), p.1.

Schüssler-Fiorenza Rose, S.M., Contrepois, K., Moneghetti, K.J., Zhou, W., Mishra, T., Mataraso, S., Dagan-Rosenfeld, O., Ganz, A.B., Dunn, J., Hornburg, D., Rego, S., Perelman, D., Ahadi, S., Sailani, M.R., Zhou, Y., Leopold, S.R., Chen, J., Ashland, M., Christle, J.W., Avina, M. and Snyder, M.P., 2019. A longitudinal big data approach for precision health. *Nature Medicine*, 25(5), pp.792–804.

Segal, A.W., 2018. The role of neutrophils in the pathogenesis of Crohn's disease. *European Journal of Clinical Investigation*, 48 Suppl 2, p.e12983.

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S. and Huttenhower, C., 2011. Metagenomic biomarker discovery and explanation. *Genome Biology*, 12(6), p.R60.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11), pp.2498–2504.

Sharma, D. and Xu, W., 2021. phyLoSTM: a novel deep learning model on disease prediction from longitudinal microbiome data. *Bioinformatics*.

Shen, Z.-H., Zhu, C.-X., Quan, Y.-S., Yang, Z.-Y., Wu, S., Luo, W.-W., Tan, B. and Wang, X.-Y., 2018. Relationship between intestinal microbiota and ulcerative colitis: Mechanisms and clinical application of probiotics and fecal microbiota transplantation. *World Journal of Gastroenterology*, 24(1), pp.5–14.

Sheth, R.U., Li, M., Jiang, W., Sims, P.A., Leong, K.W. and Wang, H.H., 2019. Spatial metagenomic characterization of microbial biogeography in the gut. *Nature Biotechnology*, 37(8), pp.877–883.

Singh, V., Yeoh, B.S., Xiao, X., Kumar, M., Bachman, M., Borregaard, N., Joe, B. and Vijay-Kumar, M., 2015. Interplay between enterobactin, myeloperoxidase and lipocalin 2 regulates E. coli survival in the inflamed gut. *Nature Communications*, 6, p.7113.

Smith, C.A., Maille, G.O., Want, E.J., Qin, C., Trauger, S.A., Brandon, T.R., Custodio, D.E., Abagyan, R. and Siuzdak, G., 2005. METLIN. *Therapeutic Drug Monitoring*, 27(6), pp.747–751.

Smith, L.A. and Gaya, D.R., 2012. Utility of faecal calprotectin analysis in adult inflammatory bowel disease. *World Journal of Gastroenterology*, 18(46), pp.6782–6789.

Sokol, H., Landman, C., Seksik, P., Berard, L., Montil, M., Nion-Larmurier, I., Bourrier, A., Le Gall, G., Lalande, V., De Rougemont, A., Kirchgesner, J., Daguenel, A., Cachanado, M., Rousseau, A., Drouet, É., Rosenzwajg, M., Hagege, H., Dray, X., Klatzman, D., Marteau, P. and Simon, T., 2020. Fecal microbiota transplantation to maintain remission in Crohn's disease: a pilot randomized controlled study. *Microbiome*, 8(1), p.12.

Sokol, H., Pigneur, B., Watterlot, L., Lakhdari, O., Bermúdez-Humarán, L.G., Gratadoux, J.-J., Blugeon, S., Bridonneau, C., Furet, J.-P., Corthier, G., Grangette, C., Vasquez, N., Pochart, P., Trugnan, G., Thomas, G., Blottière, H.M., Doré, J., Marteau, P., Seksik, P. and Langella, P., 2008. Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proceedings of the National Academy of Sciences of the United States of America*, 105(43), pp.16731–16736.

Sompairac, N., Nazarov, P.V., Czerwinska, U., Cantini, L., Biton, A., Molkenov, A., Zhumadilov, Z., Barillot, E., Radvanyi, F., Gorban, A., Kairov, U. and Zinovyev, A., 2019. Independent component analysis for unraveling the complexity of cancer omics datasets. *International Journal of Molecular Sciences*, 20(18).

de Souza, H.S.P. and Fiocchi, C., 2016. Immunopathogenesis of IBD: current state of the art. *Nature Reviews. Gastroenterology & Hepatology*, 13(1), pp.13–27.

Stafford, I.S., Gosink, M.M., Mossotto, E., Ennis, S. and Hauben, M., 2022. A Systematic Review of Artificial Intelligence and Machine Learning Applications to Inflammatory Bowel Disease, with Practical Guidelines for Interpretation. *Inflammatory Bowel Diseases*, 28(10), pp.1573–1583.

Staley, C., Weingarden, A.R., Khoruts, A. and Sadowsky, M.J., 2017. Interaction of gut microbiota with bile acid metabolism and its influence on disease states. *Applied Microbiology and Biotechnology*, 101(1), pp.47–64.

Stein, C.K., Qu, P., Epstein, J., Buros, A., Rosenthal, A., Crowley, J., Morgan, G. and Barlogie, B., 2015. Removing batch effects from purified plasma cell gene expression microarrays with

modified ComBat. *BMC Bioinformatics*, 16, p.63.

Stenlund, H., Gorzsás, A., Persson, P., Sundberg, B. and Trygg, J., 2008. Orthogonal projections to latent structures discriminant analysis modeling on in situ FT-IR spectral imaging of liver tissue for identifying sources of variability. *Analytical Chemistry*, 80(18), pp.6898–6906.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. and Zeileis, A., 2008. Conditional variable importance for random forests. *BMC Bioinformatics*, 9, p.307.

Sudhakar, P., Jacomin, A.-C., Hautefort, I., Samavedam, S., Fatemian, K., Ari, E., Gul, L., Demeter, A., Jones, E., Korcsmaros, T. and Nezis, I.P., 2019. Targeted interplay between bacterial pathogens and host autophagy. *Autophagy*, 15(9), pp.1620–1633.

Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., Wu, C.H. and UniProt Consortium, 2015. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6), pp.926–932.

Su, G., Kuchinsky, A., Morris, J.H., States, D.J. and Meng, F., 2010. GLay: community structure analysis of biological networks. *Bioinformatics*, 26(24), pp.3135–3137.

Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N.T., Legeay, M., Fang, T., Bork, P., Jensen, L.J. and von Mering, C., 2021. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, 49(D1), pp.D605–D612.

Tavassoly, I., Goldfarb, J. and Iyengar, R., 2018. Systems biology primer: the basic methods and approaches. *Essays in biochemistry*, 62(4), pp.487–500.

Teschendorff, A.E., Journée, M., Absil, P.A., Sepulchre, R. and Caldas, C., 2007. Elucidating the altered transcriptional programs in breast cancer using independent component analysis. *PLoS Computational Biology*, 3(8), p.e161.

Tessler, M., Neumann, J.S., Afshinnekoo, E., Pineda, M., Hersch, R., Velho, L.F.M., Segovia, B.T., Lansac-Toha, F.A., Lemke, M., DeSalle, R., Mason, C.E. and Brugler, M.R., 2017. Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. *Scientific Reports*, 7(1), p.6589.

Thomas, J.P., Modos, D., Rushbrook, S.M., Powell, N. and Korcsmaros, T., 2022. The emerging role of bile acids in the pathogenesis of inflammatory bowel disease. *Frontiers in Immunology*, 13, p.829525.

Titterington, D.M., 1997. Introduction to Gelfand and Smith (1990) Sampling-Based Approaches to Calculating Marginal Densities. In: S. Kotz and N.L. Johnson, eds. *Breakthroughs in Statistics*, Springer series in statistics. New York, NY: Springer New York.pp.519–550.

Truong, D.T., Franzosa, E.A., Tickle, T.L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C. and Segata, N., 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, 12(10), pp.902–903.

Trygg, J. and Wold, S., 2002. Orthogonal projections to latent structures (O-PLS). *Journal of chemometrics*, 16(3), pp.119–128.

Türei, D., Korcsmáros, T. and Saez-Rodriguez, J., 2016. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nature Methods*, 13(12), pp.966–967.

Türei, D., Valdeolivas, A., Gul, L., Palacio-Escat, N., Klein, M., Ivanova, O., Ölbei, M., Gábor, A., Theis, F., Módos, D., Korcsmáros, T. and Saez-Rodriguez, J., 2021. Integrated intra- and

intercellular signaling knowledge for multicellular omics analysis. *Molecular Systems Biology*, 17(3).

Ungaro, R., Mehandru, S., Allen, P.B., Peyrin-Biroulet, L. and Colombel, J.-F., 2017. Ulcerative colitis. *The Lancet*, 389(10080), pp.1756–1770.

UniProt Consortium, 2021. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1), pp.D480–D489.

Valdes, A.M., Walter, J., Segal, E. and Spector, T.D., 2018. Role of the gut microbiota in nutrition and health. *BMJ (Clinical Research Ed.)*, 361, p.k2179.

Van Den Bossche, T., Arntzen, M.Ø., Becher, D., Benndorf, D., Eijsink, V.G.H., Henry, C., Jagtap, P.D., Jehmlich, N., Juste, C., Kunath, B.J., Mesuere, B., Muth, T., Pope, P.B., Seifert, J., Tanca, A., Uzzau, S., Wilmes, P., Hettich, R.L. and Armengaud, J., 2021. The Metaproteomics Initiative: a coordinated approach for propelling the functional characterization of microbiomes. *Microbiome*, 9(1), p.243.

Vázquez-Baeza, Y., Gonzalez, A., Xu, Z.Z., Washburne, A., Herfarth, H.H., Sartor, R.B. and Knight, R., 2018. Guiding longitudinal sampling in IBD cohorts. *Gut*, 67(9), pp.1743–1745.

Velliangiri, S., Alagumuthukrishnan, S. and Thankumar joseph, S.I., 2019. A review of dimensionality reduction techniques for efficient computation. *Procedia Computer Science*, 165, pp.104–111.

Velten, B., Braunger, J.M., Argelaguet, R., Arnol, D., Wirbel, J., Bredikhin, D., Zeller, G. and Stegle, O., 2022. Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO. *Nature Methods*, 19(2), pp.179–186.

Vernocchi, P., Del Chierico, F. and Putignani, L., 2016. Gut microbiota profiling: metabolomics based approach to unravel compounds affecting human health. *Frontiers in Microbiology*, 7, p.1144.

Verstockt, B., Vetrano, S., Salas, A., Nayeri, S., Duijvestein, M., Vande Casteele, N. and Alimentiv Translational Research Consortium (ATRC), 2022. Sphingosine 1-phosphate modulation and immune cell trafficking in inflammatory bowel disease. *Nature Reviews. Gastroenterology & Hepatology*, 19(6), pp.351–366.

Vich Vila, A., Hu, S., Andreu-Sánchez, S., Collij, V., Jansen, B.H., Augustijn, H.E., Bolte, L.A., Ruigrok, R.A.A.A., Abu-Ali, G., Giallourakis, C., Schneider, J., Parkinson, J., Al-Garawi, A., Zhernakova, A., Gacesa, R., Fu, J. and Weersma, R.K., 2023. Faecal metabolome and its determinants in inflammatory bowel disease. *Gut*, 72(8), pp.1472–1485.

Vinaixa, M., Schymanski, E.L., Neumann, S., Navarro, M., Salek, R.M. and Yanes, O., 2016. Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects. *TrAC Trends in Analytical Chemistry*, 78, pp.23–35.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J. and SciPy 1.0 Contributors, 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3), pp.261–272.

Walmsley, R.S., Ayres, R.C., Pounder, R.E. and Allan, R.N., 1998. A simple clinical colitis activity index. *Gut*, 43(1), pp.29–32.

Walters, K.E. and Martiny, J.B.H., 2020. Alpha-, beta-, and gamma-diversity of bacteria varies across habitats. *Plos One*, 15(9), p.e0233872.

Wang, Y., Gao, X., Zhang, X., Xiao, F., Hu, H., Li, X., Dong, F., Sun, M., Xiao, Y., Ge, T., Li, D., Yu, G., Liu, Z. and Zhang, T., 2021. Microbial and metabolic features associated with outcome of infliximab therapy in pediatric Crohn's disease. *Gut microbes*, 13(1), pp.1–18.

Weigele, B.A., Orchard, R.C., Jimenez, A., Cox, G.W. and Alto, N.M., 2017. A systematic exploration of the interactions between bacterial effector proteins and host cell membranes. *Nature Communications*, 8(1), p.532.

Weinstock, G.M., 2012. Genomic approaches to studying the human microbiota. *Nature*, 489(7415), pp.250–256.

Weiss, S., Xu, Z.Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J.R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E.R. and Knight, R., 2017. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(1), p.27.

Weng, Y.J., Gan, H.Y., Li, X., Huang, Y., Li, Z.C., Deng, H.M., Chen, S.Z., Zhou, Y., Wang, L.S., Han, Y.P., Tan, Y.F., Song, Y.J., Du, Z.M., Liu, Y.Y., Wang, Y., Qin, N., Bai, Y., Yang, R.F., Bi, Y.J. and Zhi, F.C., 2019. Correlation of diet, microbiota and metabolite networks in inflammatory bowel disease. *Journal of digestive diseases*, 20(9), pp.447–459.

Wieder, C., Lai, R.P.J. and Ebbels, T.M.D., 2022. Single sample pathway analysis in metabolomics: performance evaluation and application. *BMC Bioinformatics*, 23(1), p.481.

Wishart, D.S., Tzur, D., Knox, C., Eisner, R., Guo, A.C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M.-A., Forsythe, I., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., Amegbey, G. and Querengesser, L., 2007. HMDB: the human metabolome database. *Nucleic Acids Research*, 35(Database issue), pp.D521-6.

Wu, H.-J. and Wu, E., 2012. The role of gut microbiota in immune homeostasis and autoimmunity. *Gut microbes*, 3(1), pp.4–14.

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X. and Yu, G., 2021. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Cambridge (Mass.))*, 2(3), p.100141.

Xu, X., Liang, T., Zhu, J., Zheng, D. and Sun, T., 2018. Review of classical dimensionality reduction and sample selection methods for large-scale data processing. *Neurocomputing*, 328, pp.5–15.

Yang, C., Chowdhury, D., Zhang, Z., Cheung, W.K., Lu, A., Bian, Z. and Zhang, L., 2021a. A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Computational and structural biotechnology journal*, 19, pp.6301–6314.

Yang, J., Pei, G., Sun, X., Xiao, Y., Miao, C., Zhou, L., Wang, B., Yang, L., Yu, M., Zhang, Z.-S., Keller, E.T., Yao, Z. and Wang, Q., 2022. RhoB affects colitis through modulating cell signaling and intestinal microbiome. *Microbiome*, 10(1), p.149.

Yang, M., Gu, Y., Li, L., Liu, T., Song, X., Sun, Y., Cao, X., Wang, B., Jiang, K. and Cao, H., 2021b. Bile Acid-Gut Microbiota Axis in Inflammatory Bowel Disease: From Bench to Bedside. *Nutrients*, 13(9).

Yang, S., Li, H., He, H., Zhou, Y. and Zhang, Z., 2019. Critical assessment and performance improvement of plant-pathogen protein-protein interaction prediction methods. *Briefings in Bioinformatics*, 20(1), pp.274–287.

Yang, Y., Chen, N. and Chen, T., 2017. Inference of Environmental Factor-Microbe and Microbe-Microbe Associations from Metagenomic Data Using a Hierarchical Bayesian Statistical Model. *Cell Systems*, 4(1), pp.129-137.e5.

Yu, G. and He, Q.-Y., 2016. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Molecular Biosystems*, 12(2), pp.477–479.

Zamani, S., Hesam Shariati, S., Zali, M.R., Asadzadeh Aghdaei, H., Sarabi Asiabar, A., Bokaie, S., Nomanpour, B., Sechi, L.A. and Feizabadi, M.M., 2017. Detection of enterotoxigenic Bacteroides fragilis in patients with ulcerative colitis. *Gut Pathogens*, 9, p.53.

Zeng, M.Y., Inohara, N. and Nuñez, G., 2017. Mechanisms of inflammation-driven bacterial dysbiosis in the gut. *Mucosal Immunology*, 10(1), pp.18–26.

Zeng, Z., Mukherjee, A., Varghese, A.P., Yang, X.-L., Chen, S. and Zhang, H., 2020. Roles of G protein-coupled receptors in inflammatory bowel disease. *World Journal of Gastroenterology*, 26(12), pp.1242–1261.

Zhang, E., Yan, Y., Lei, Y., Qu, Y., Fan, Z., Zhang, T., Xu, Y., Du, Q., Brugger, D., Chen, Y. and Zhang, K., 2023. Bacteroides uniformis regulates TH17 cell differentiation and alleviates chronic colitis by producing alpha-muricholic acid. *Research square*.

Zhang, X., Deeke, S.A., Ning, Z., Starr, A.E., Butcher, J., Li, J., Mayne, J., Cheng, K., Liao, B., Li, L., Singleton, R., Mack, D., Stintzi, A. and Figeys, D., 2018a. Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease. *Nature Communications*, 9(1), p.2873.

Zhang, X., Guo, B. and Yi, N., 2020. Zero-Inflated gaussian mixed models for analyzing longitudinal microbiome data. *Plos One*, 15(11), p.e0242073.

Zhang, X., Li, L., Mayne, J., Ning, Z., Stintzi, A. and Figeys, D., 2018b. Assessing the impact of protein extraction methods for human gut metaproteomics. *Journal of Proteomics*, 180, pp.120–127.

Zheng, L., Wen, X.-L. and Duan, S.-L., 2022. Role of metabolites derived from gut microbiota in inflammatory bowel disease. *World journal of clinical cases*, 10(9), pp.2660–2677.

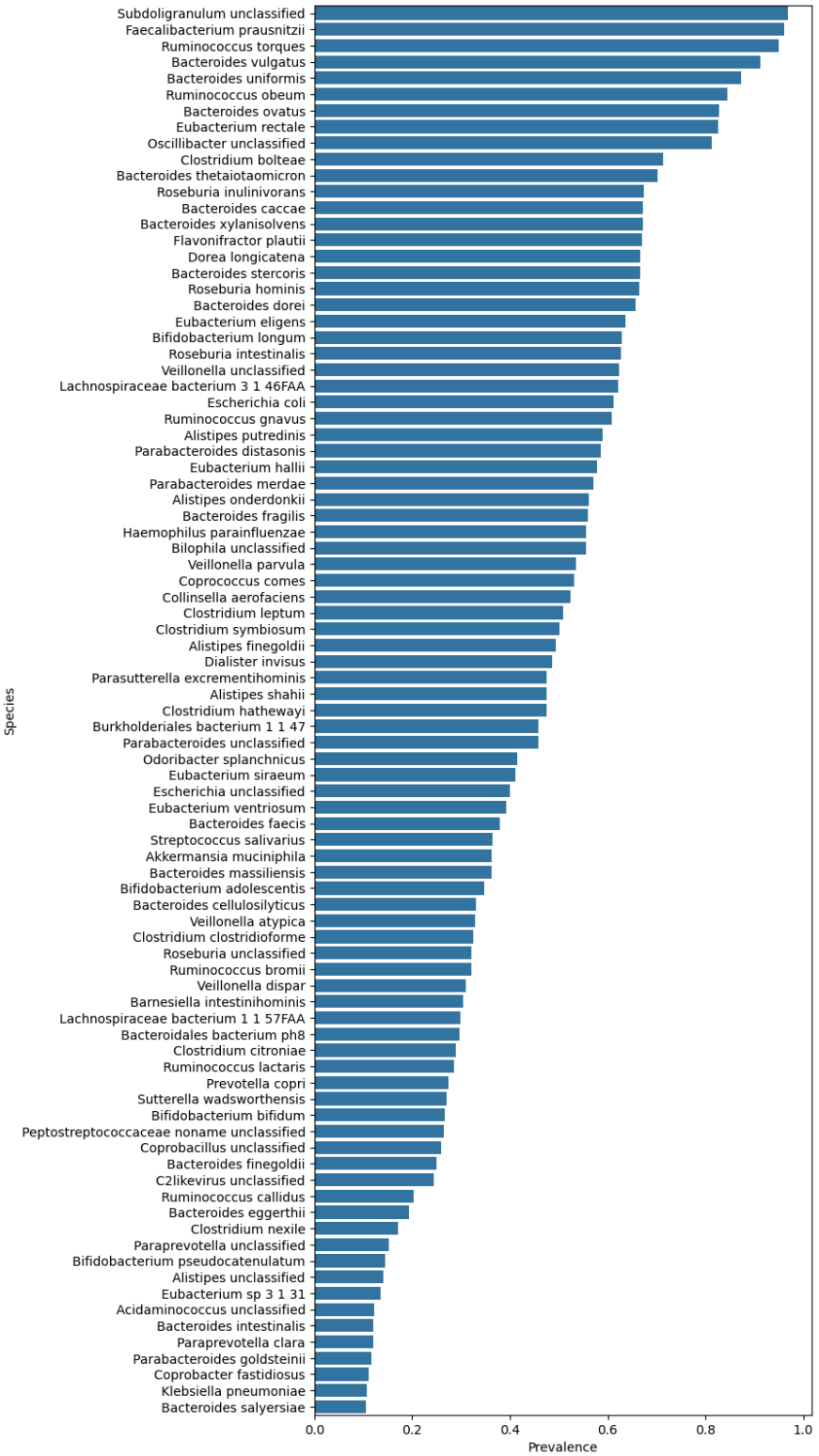
Zhou, F., He, K., Li, Q., Chapkin, R.S. and Ni, Y., 2022. Bayesian biclustering for microbial metagenomic sequencing data via multinomial matrix factorization. *Biostatistics*, 23(3), pp.891–909.

Zhou, H., Beltrán, J.F. and Brito, I.L., 2022. Host-microbiome protein-protein interactions capture disease-relevant pathways. *Genome Biology*, 23(1), p.72.

Zhou, Y. and Zhi, F., 2016. Lower Level of Bacteroides in the Gut Microbiota Is Associated with Inflammatory Bowel Disease: A Meta-Analysis. *BioMed research international*, 2016, p.5828959.

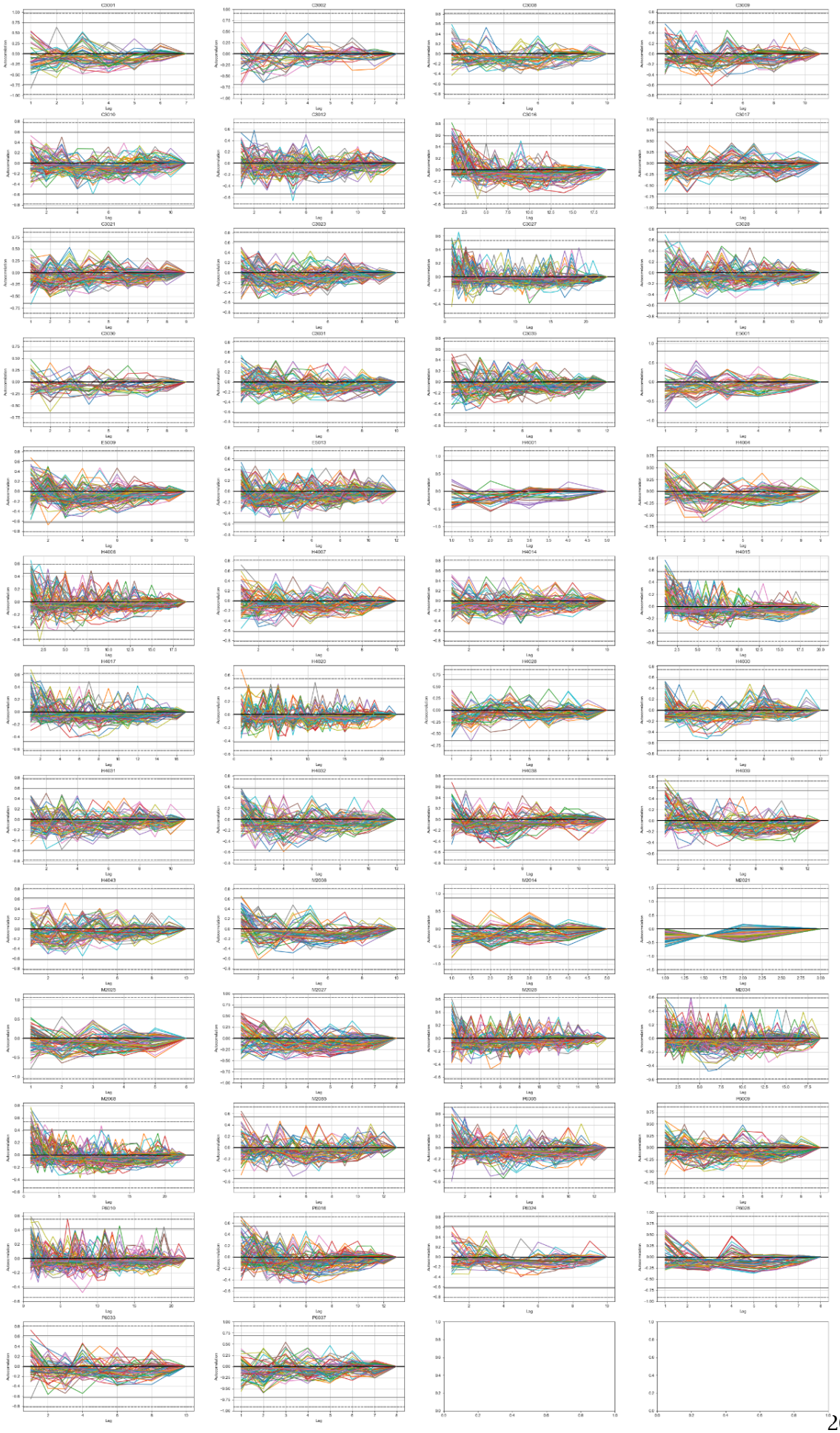
Zuo, K., Li, J., Li, K., Hu, C., Gao, Y., Chen, M., Hu, R., Liu, Y., Chi, H., Wang, H., Qin, Y., Liu, X., Li, S., Cai, J., Zhong, J. and Yang, X., 2019. Disordered gut microbiota and alterations in metabolic patterns are associated with atrial fibrillation. *GigaScience*, 8(6).

Appendix 1: Supplementary Chapter 2



Supplementary Figure 2.1. Prevalence of microbial species across the dataset that appear in more the 10% of all samples.

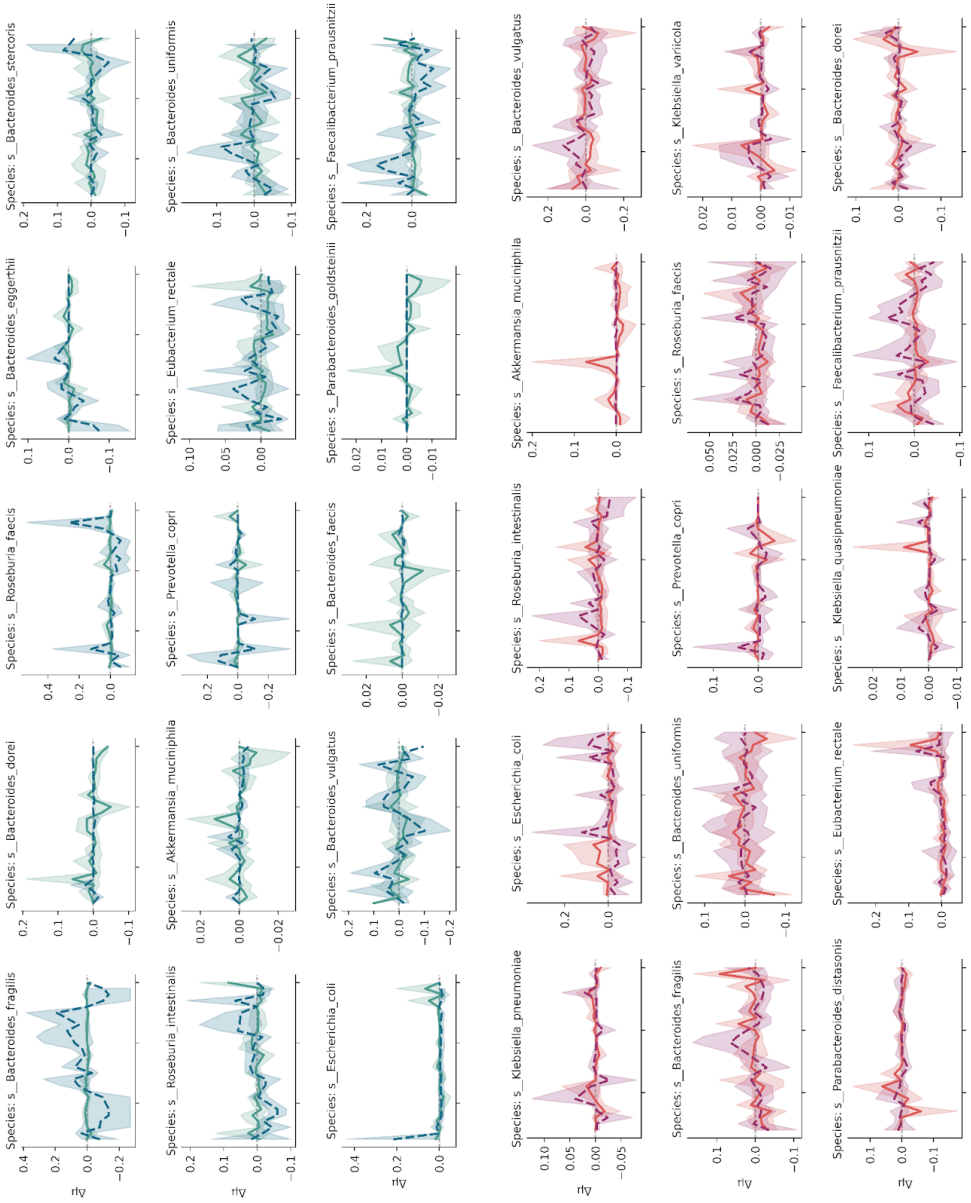

C3003
C3004
C3005
C3006
C3011
C3013
C3015
C3029
C3032
C3034
C3037
E9004
H4010
H4019
H4027
H4035
H4043
H4042
H4044
M2026
M2064
M2069
M2071
M2083
M2163
P6012
P6013
P6025
P6035
P6038


C3001
C3002
C3008
C3009
C3010
C3012
C3016
C3017
C3021
C3023
C3027
C3028
C3030
C3031
C3035
E5001
E5039
E5013
H4001
H4004
H4008
H4007
H4014
H4015
H4017
H4022
H4028
H4032
H4031
H4032
H4038
H4039
H4043
M2008
M2014
M2021
M2025
M2027
M2029
M2034
M2068
M2039
P6005
P6009
P6010
P6018
P6024
P6028
P6033
P6037

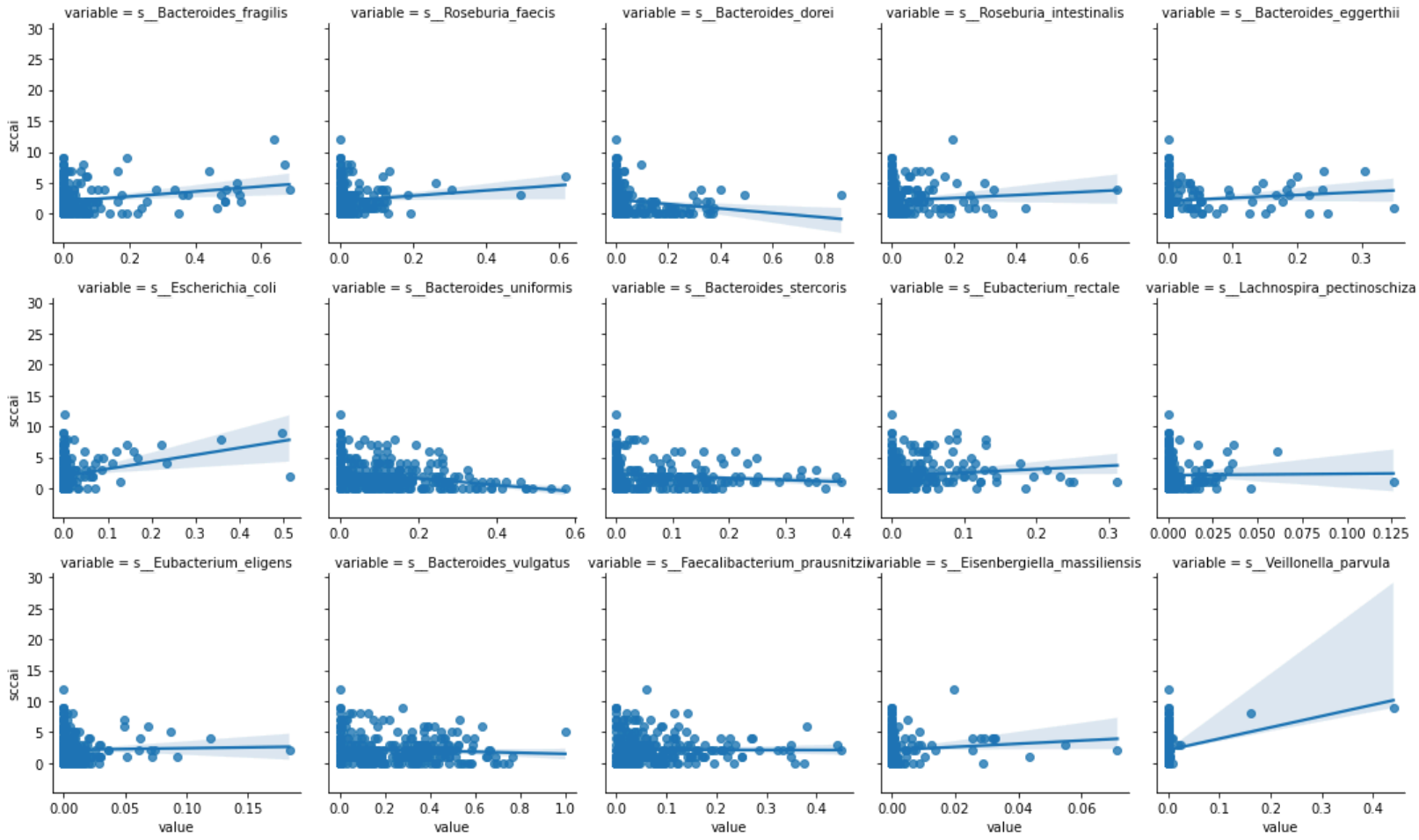

C3022
H4008
H4009
H4013
H4018
H4018
H4022
H4023
H4024
H4045
M2039
M2041
M2042
M2047
M2048
M2066
M2061
M2072
M2075
M2077
M2079
M2084
M2097
P6014
P6017
P6018

Supplementary Figure 2.2. (Previous pages 3 pages) Show the autocorrelation of each species in each patient in UC, CD and healthy controls respectively.

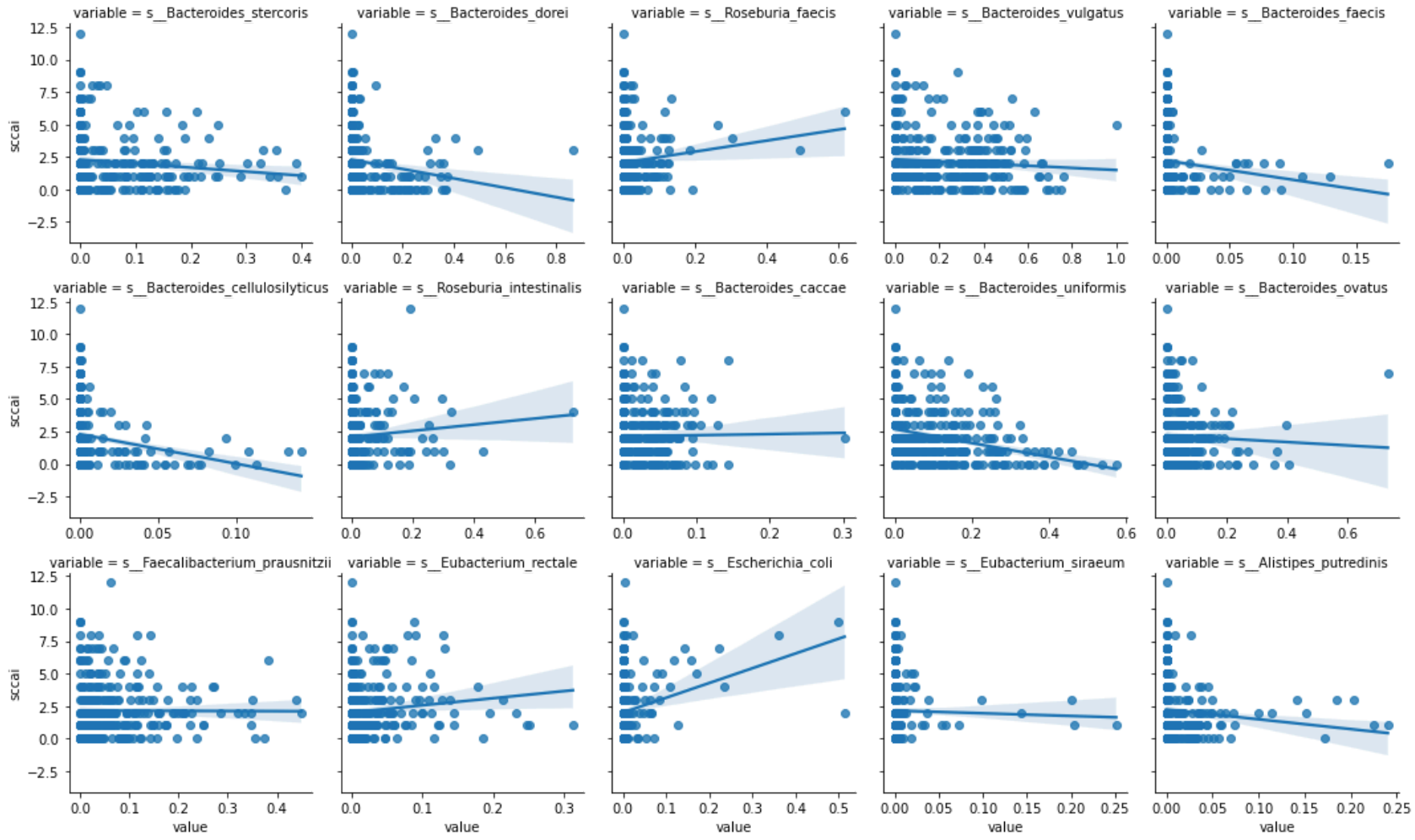
Appendix 2: Supplementary Chapter 3



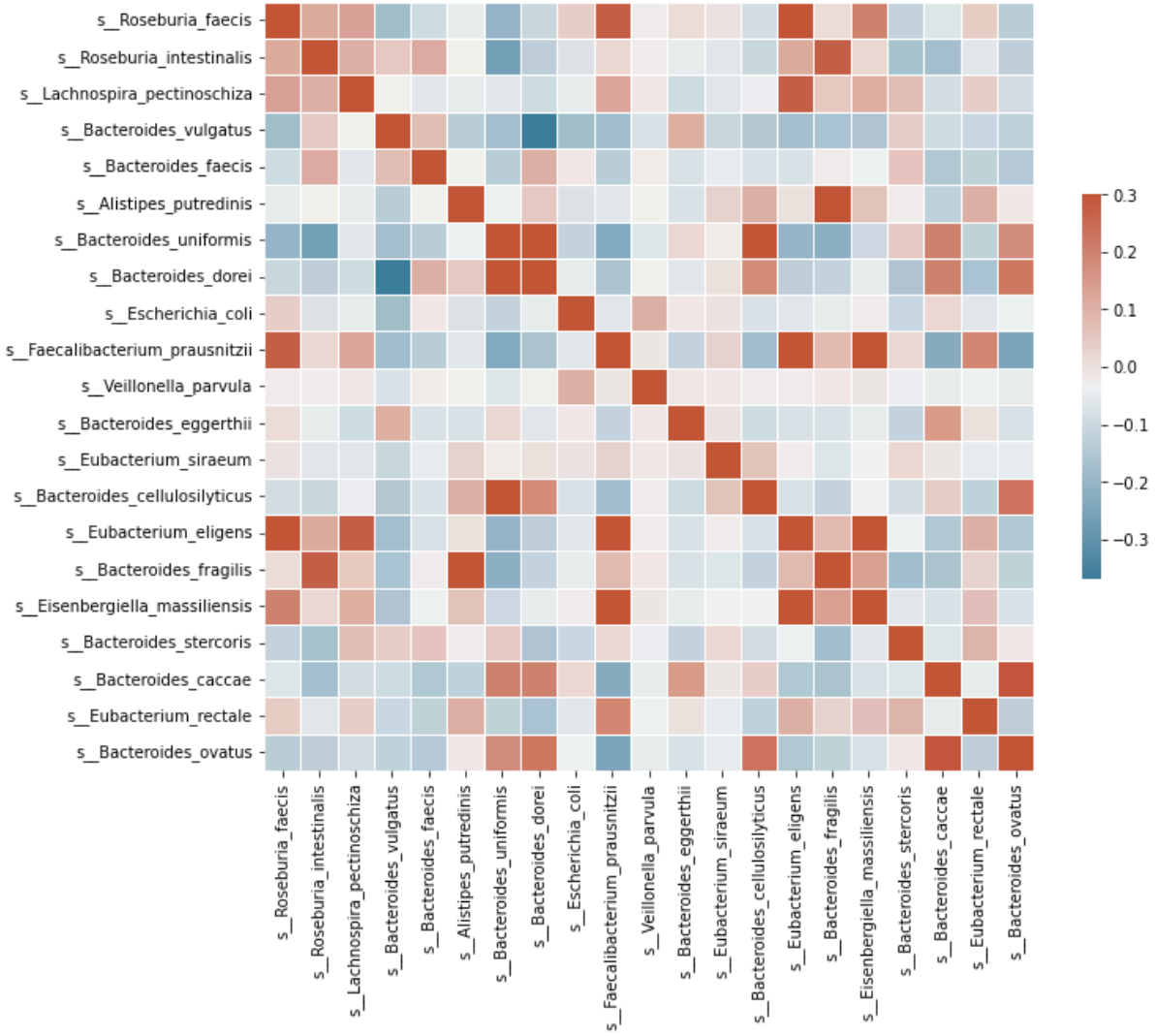
Supplementary Figure 3.1. IBD patients in remission vs patients that flared over the course of the study. The top 15 most unstable species in UC (left blue) and CD (right red) patients between remission and flare states. Δμ shows the current sample subtracted by the baseline regression for that patient. The solid line shows a patient in remission (SCCAI < 2, HBI < 5); the dashed line shows a patient who experiences a flare during the course of the study (SCCAI ≥ 2, HBI ≥ 5).



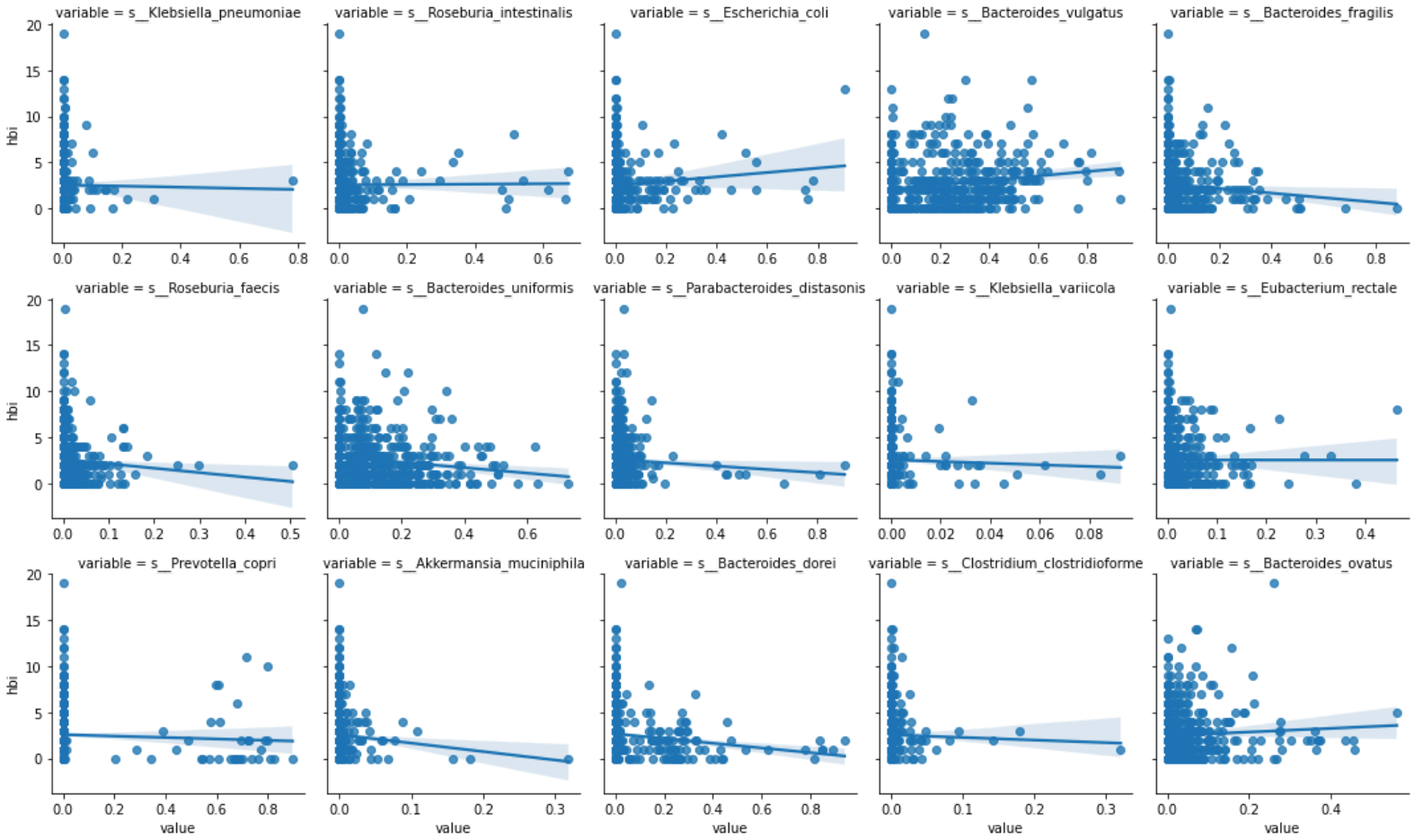
Supplementary Figure 3.2. SPM model selected top species in UC active regression against disease activity and relative abundance.



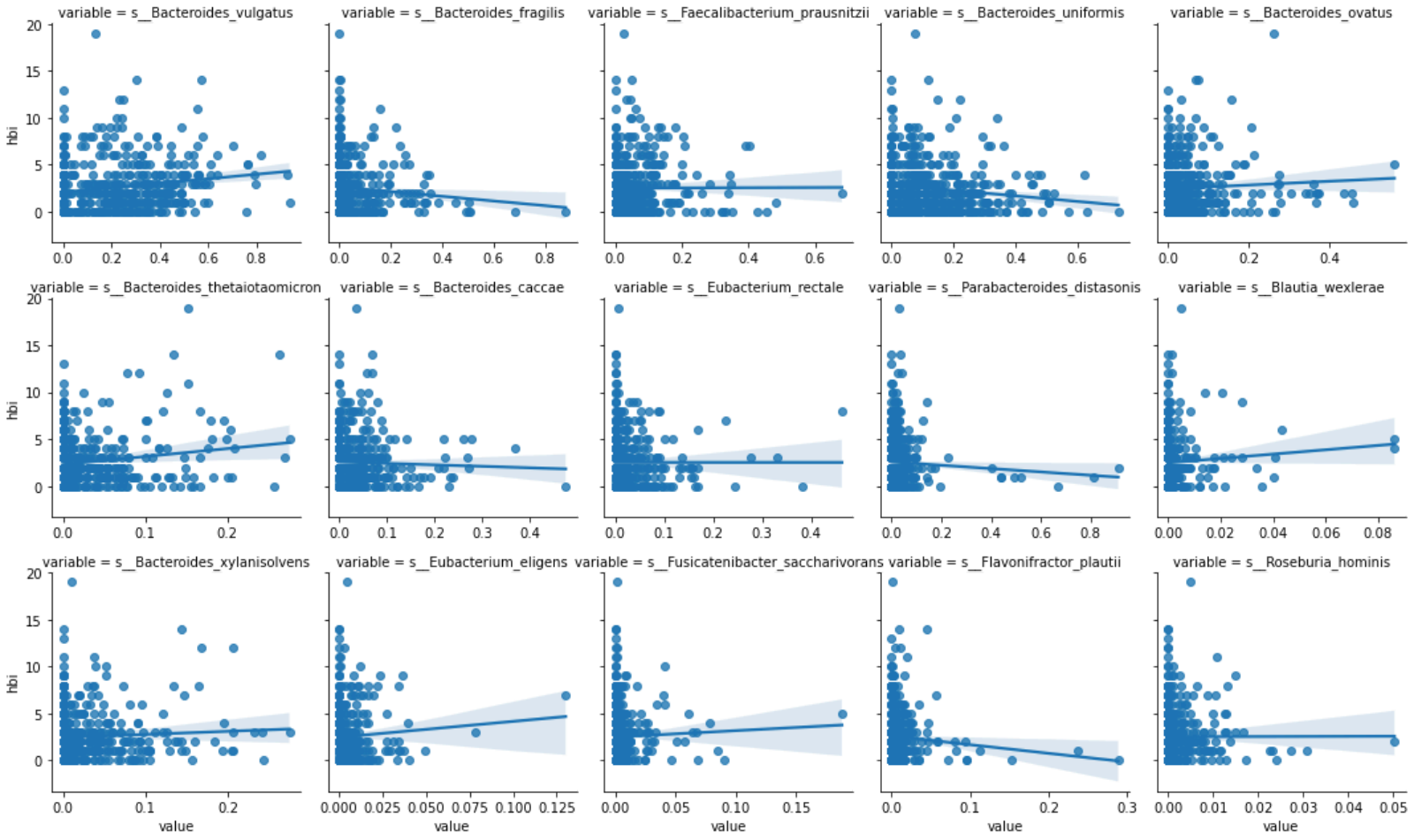
Supplementary Figure 3.3. SPM model selected top species in UC inactive regression against disease activity and relative abundance.



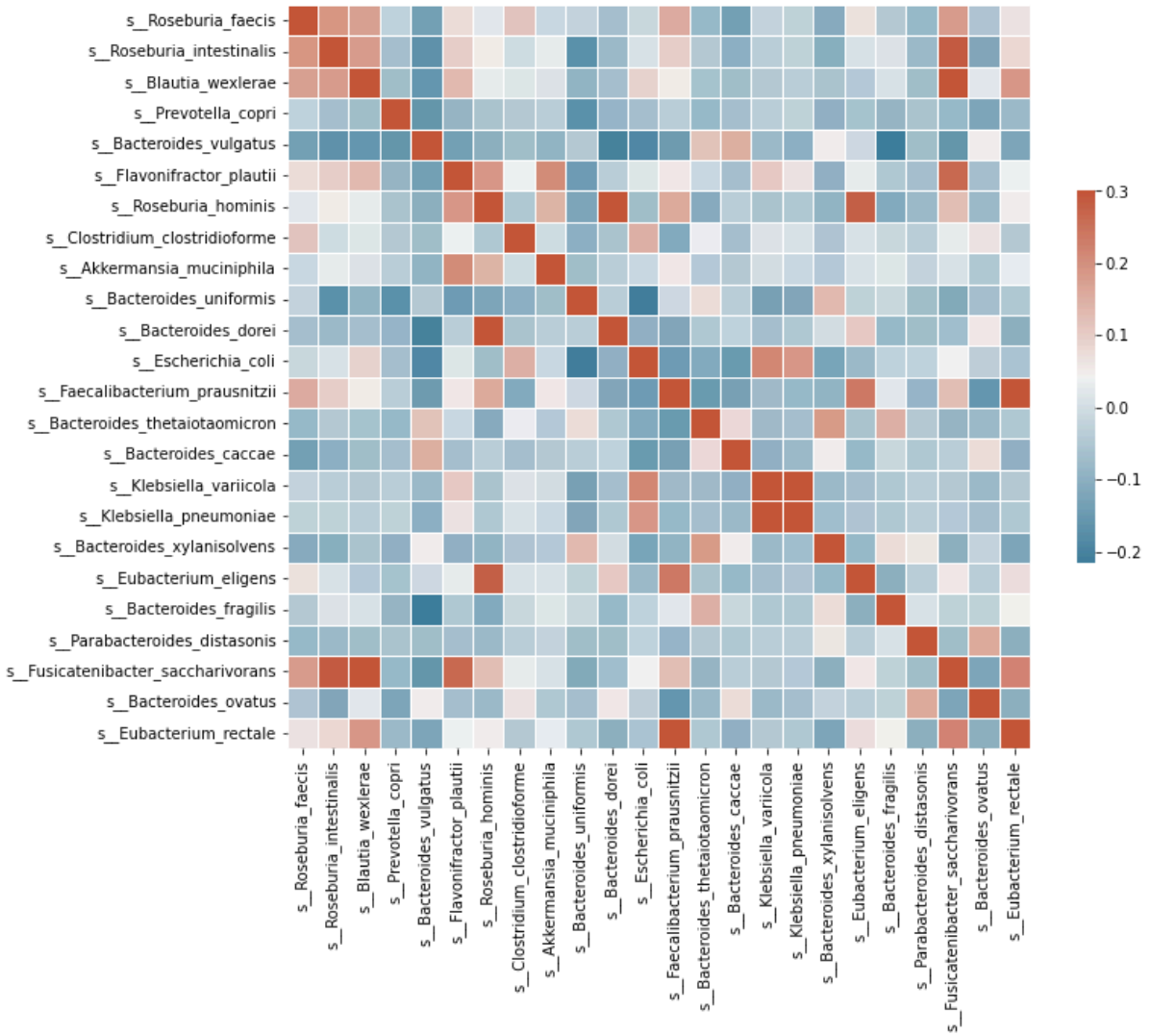
Supplementary Figure 3.4. SPM identified species in both active and inactive correlation based on their abundances in CD.



Supplementary Figure 3.5. SPM model selected top species in CD inactive regression against disease activity and relative abundance.

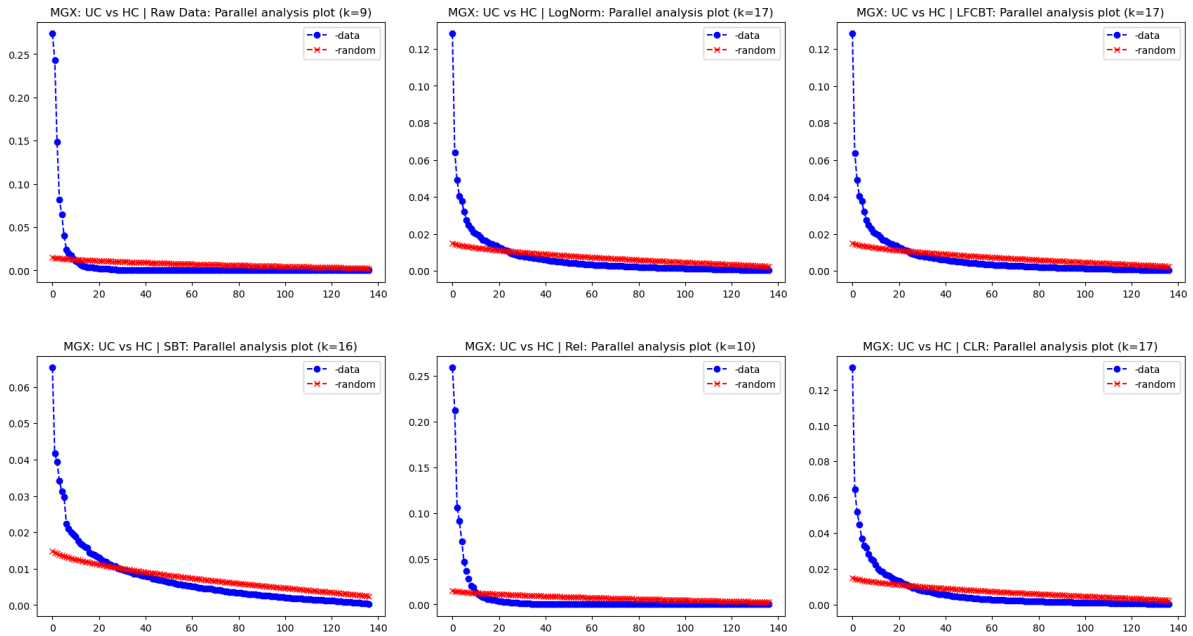


Supplementary Figure 3.6. SPM model selected top species in CD active regression against disease activity and relative abundance.

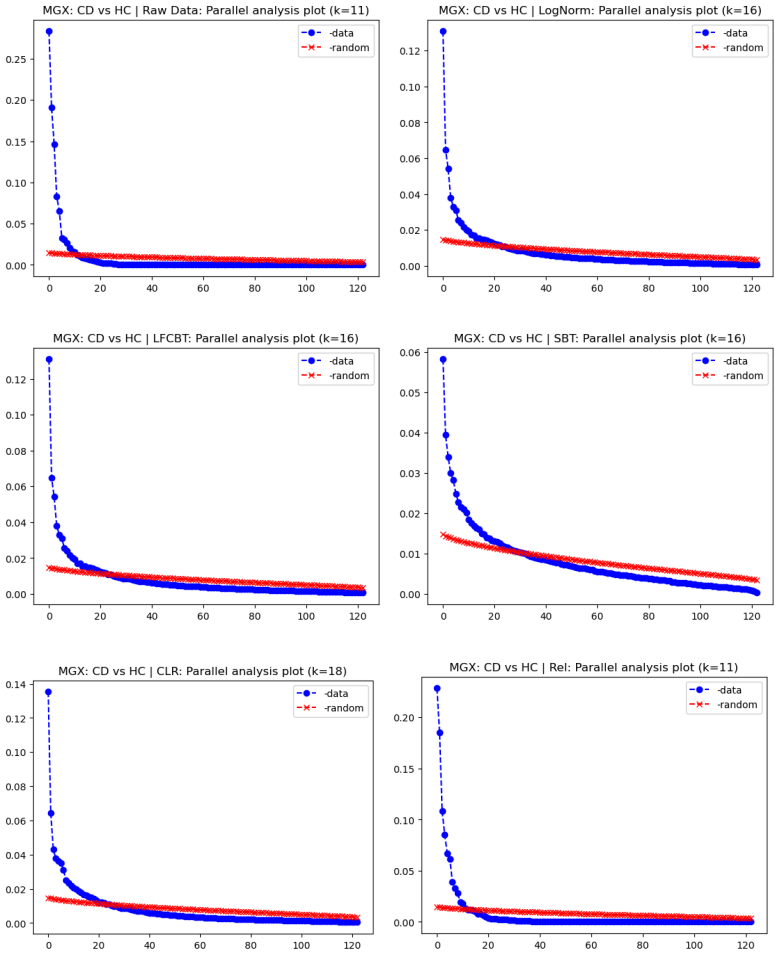


Supplementary Figure 3.7. SPM identified species in both active and inactive correlation based on their abundances in CD.

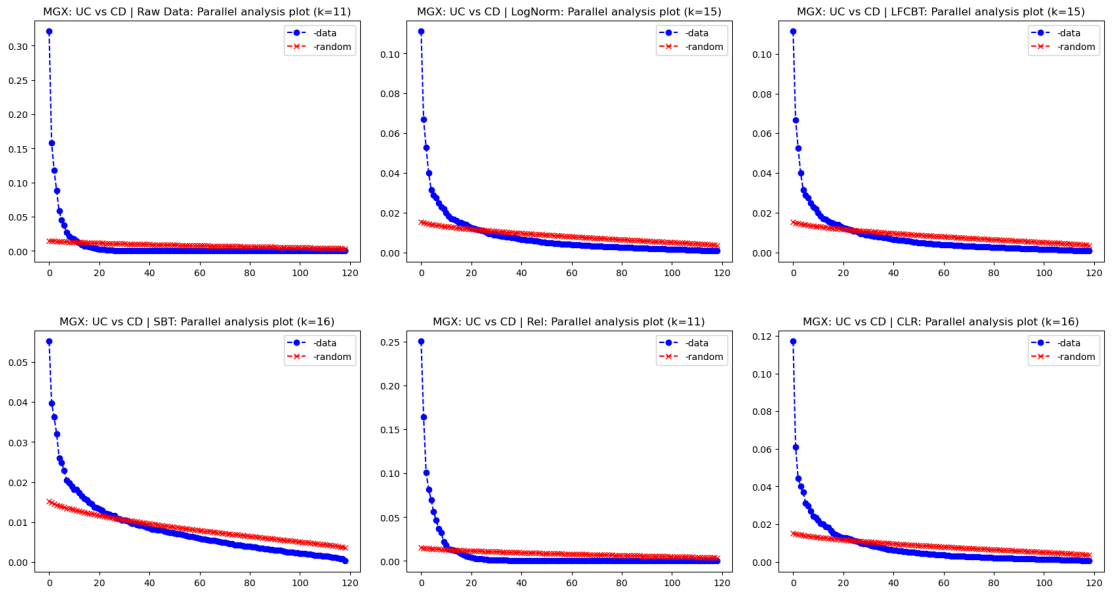
Appendix 3: Supplementary Chapter 4



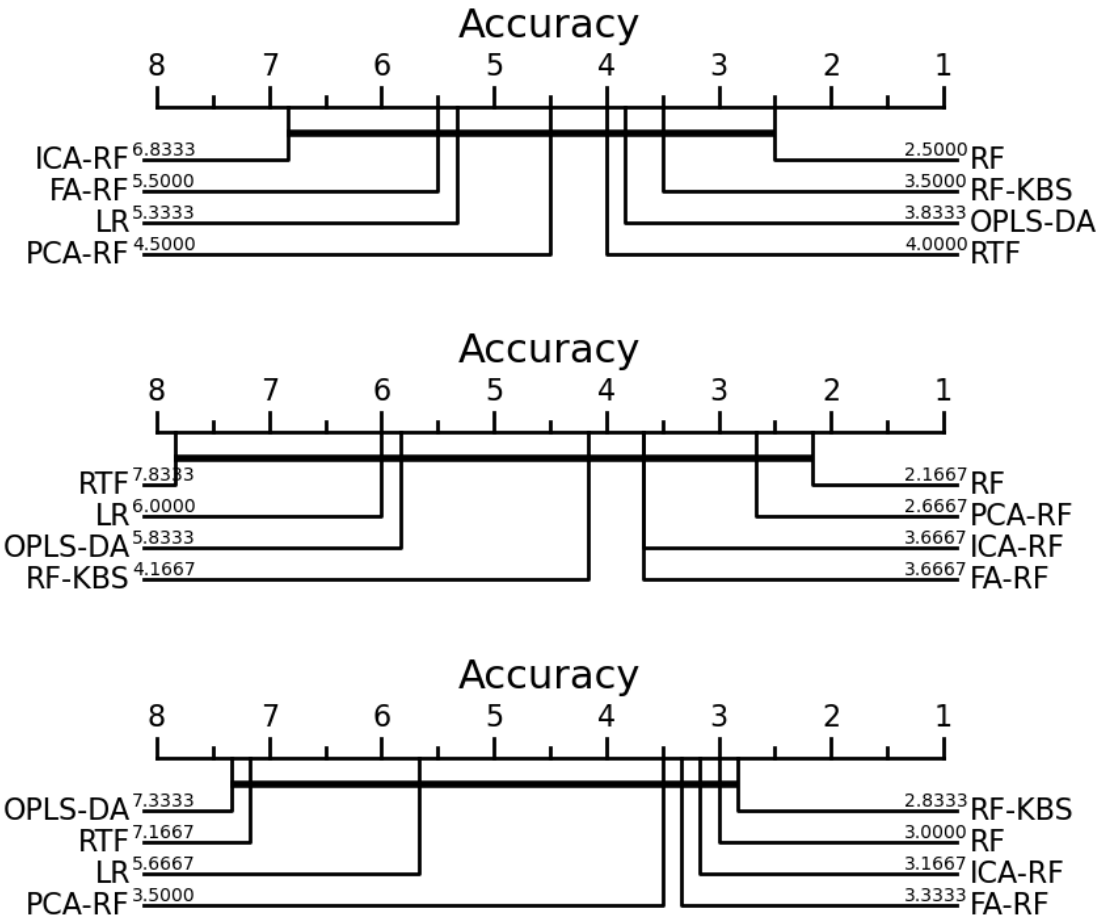
Supplementary Figure 4.1. Metagenomic UC vs HC Scree plot for assessing the number of components needed for matrix factorisation algorithms based on Horn's parallel analysis.



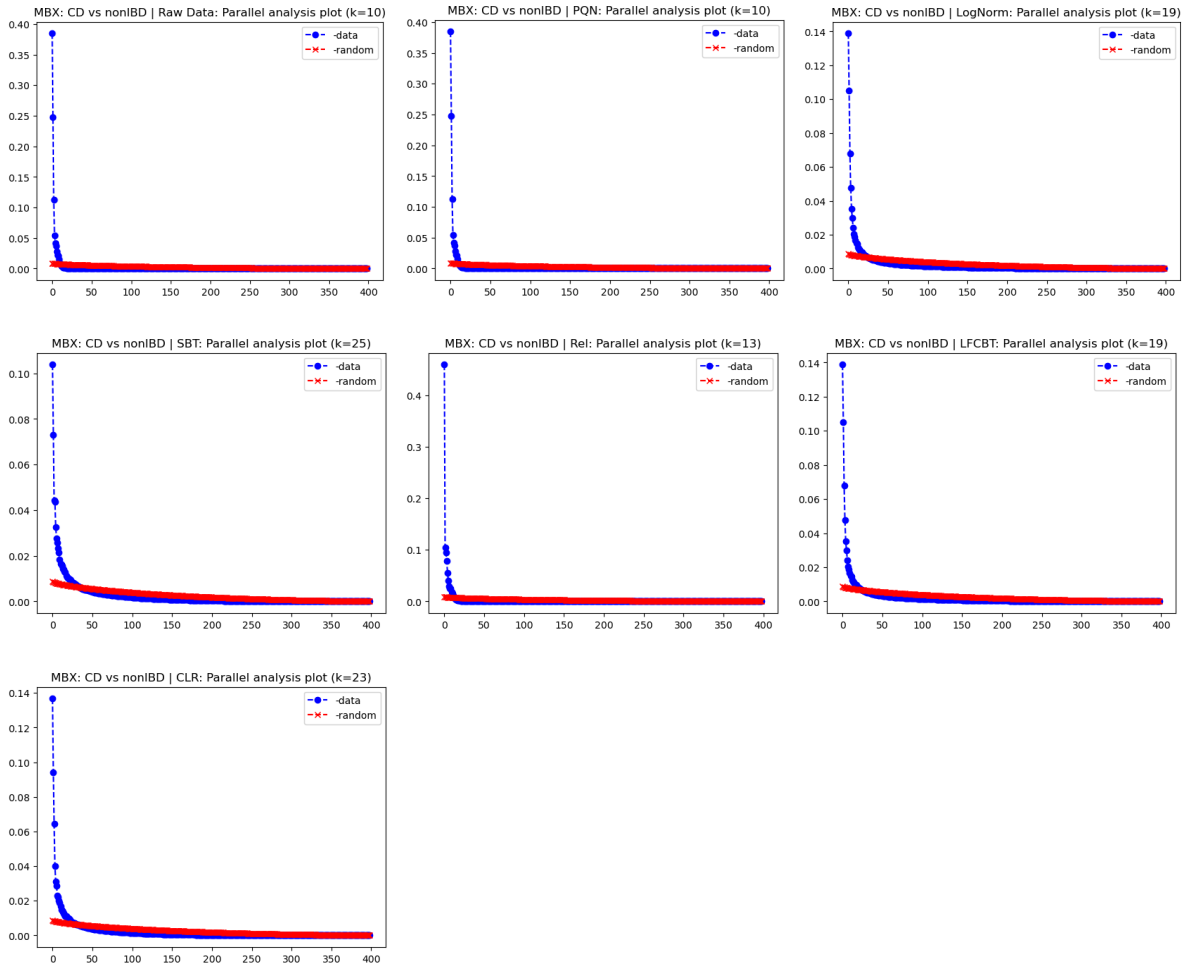
Supplementary Figure 4.2. Metagenomic CD vs HC Scree plot for assessing the number of components needed for matrix factorisation algorithms based on Horn's parallel analysis.



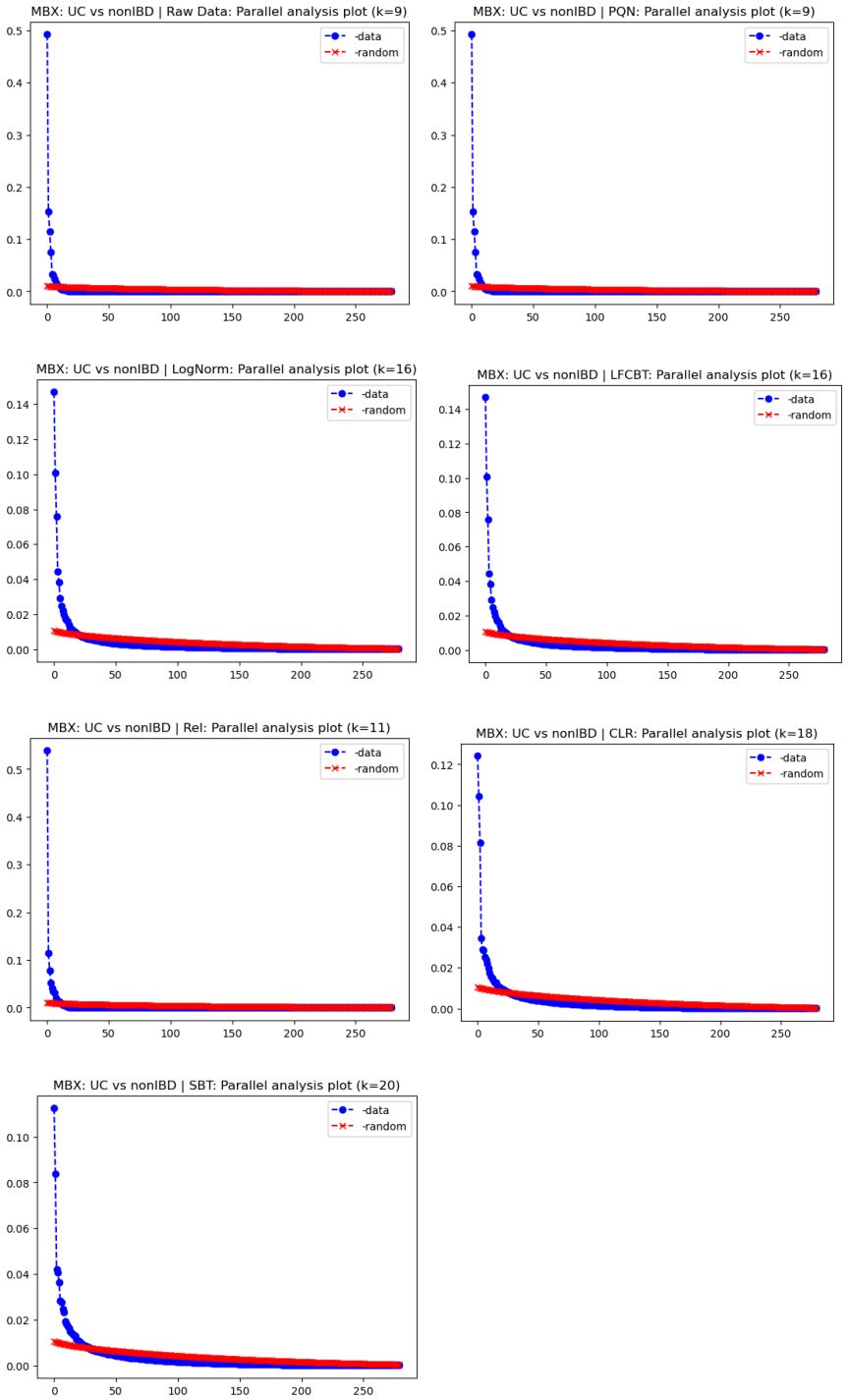
Supplementary Figure 4.3. Metagenomic UC vs CD Scree plot for assessing the number of components needed for matrix factorisation algorithms based on Horn's parallel analysis.



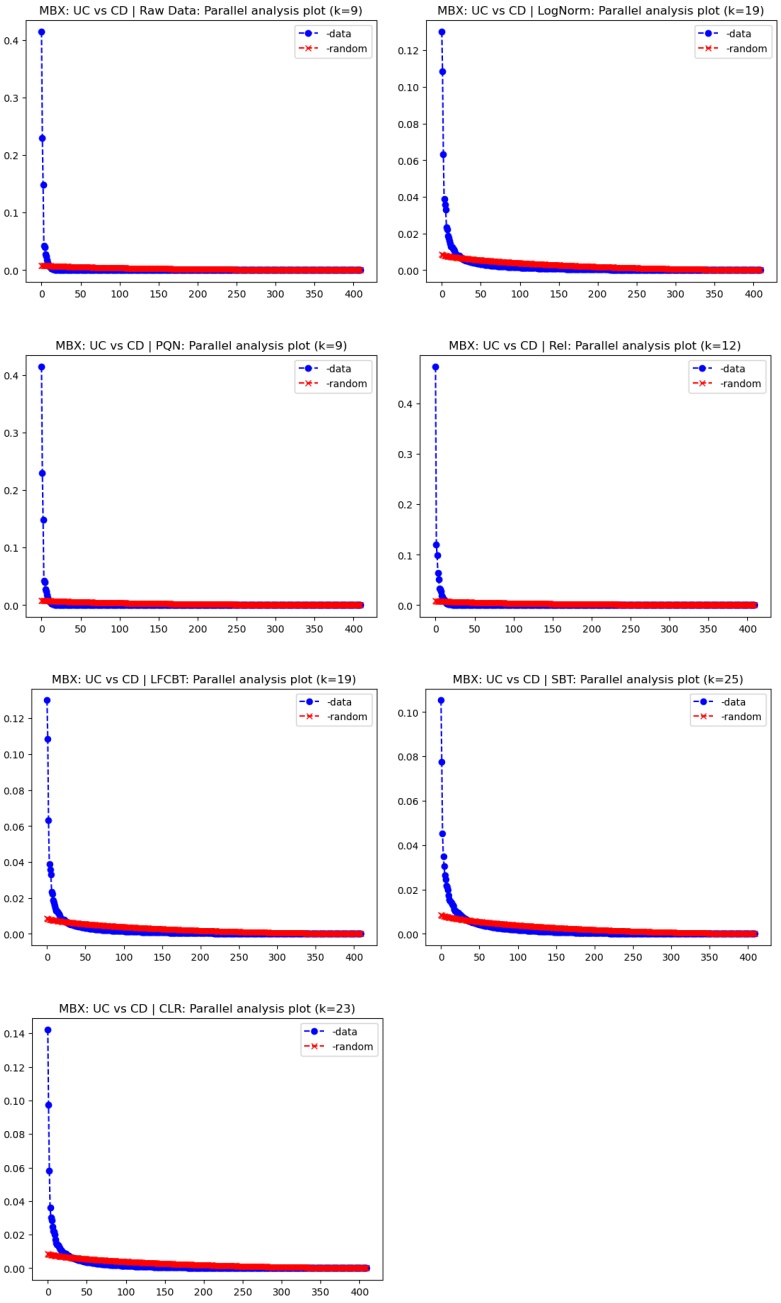
Supplementary Figure 4.4. Critical difference between models build for predicting phenotype based on metagenomic profiles.



Supplementary Figure 4.5. Metabolomic CD vs HC Scree plot for assessing the number of components needed for matrix factorisation algorithms based on Horn's parallel analysis.



Supplementary Figure 4.6. Metabolomic UC vs HC Scree plot for assessing the number of components needed for matrix factorisation algorithms based on Horn's parallel analysis.



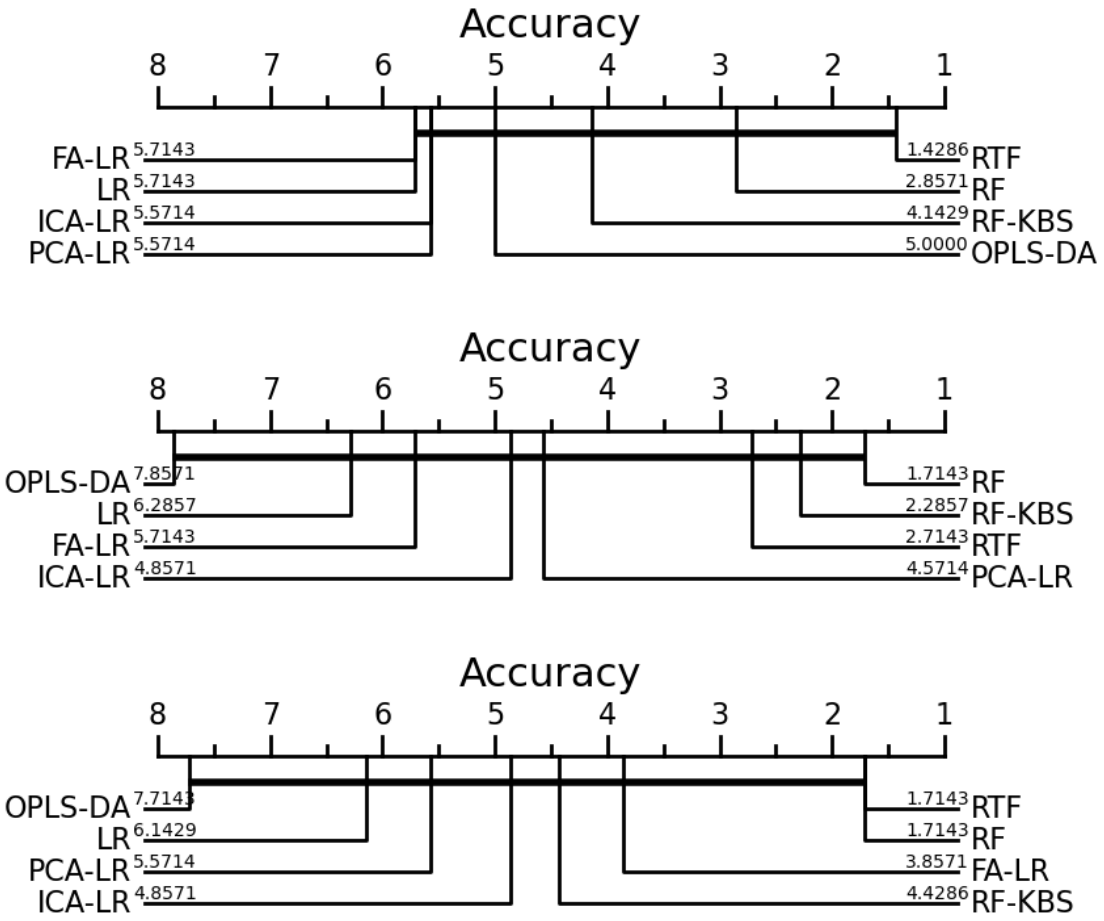
Supplementary Figure 4.7. Metagenomic UC vs CD Scree plot for assessing the number of components needed for matrix factorisation algorithms based on Horn's parallel analysis.

Supplementary Table 4.1. Table of results for Metagenomic classifiers using different normalisation methods. The table only includes prediction based sources and not probability based scores (i.e. log loss and Biers score).

		avg(Metric)	avg(Metric)	avg(Metric)	avg(Metric)	avg(Metric)	std(Metric)	std(Metric)	std(Metric)	std(Metric)	std(Metric)
Metric		**f1_score**	**matthews_corrcoef**	**precision_score**	**recall_score**	**roc_auc_score**	**f1_score**	**matthews_corrcoef**	**precision_score**	**recall_score**	**roc_auc_score**
Preprocessing	**Model**										
MBX: UC vs nonIBD | CLR	FA-LR	0.7736	0.5398	0.7730	0.7800	0.7670	0.1185	0.2308	0.1052	0.1474	0.1129
MBX: UC vs nonIBD | CLR	ICA-LR	0.7441	0.4887	0.7516	0.7533	0.7415	0.1166	0.2059	0.1126	0.1509	0.1043
MBX: UC vs nonIBD | CLR	LR	0.7625	0.5251	0.7868	0.7538	0.7601	0.0840	0.1706	0.1145	0.1021	0.0865
MBX: UC vs nonIBD | CLR	OPLS-DA	0.7777	0.5602	0.7904	0.7862	0.7735	0.0914	0.1624	0.0947	0.1444	0.0813
MBX: UC vs nonIBD | CLR	PCA-LR	0.7307	0.4579	0.7548	0.7200	0.7275	0.0981	0.2136	0.1404	0.1045	0.1062
MBX: UC vs nonIBD | CLR	RF	0.8107	0.6073	0.8125	0.8167	0.8033	0.1347	0.2908	0.1562	0.1327	0.1460
MBX: UC vs nonIBD | CLR	RF-KBS	0.7893	0.5613	0.7844	0.8033	0.7774	0.1326	0.2630	0.1450	0.1401	0.1330
MBX: UC vs nonIBD | CLR	RTF	0.8080	0.6019	0.8163	0.8090	0.8003	0.1423	0.3116	0.1699	0.1337	0.1563
MBX: UC vs nonIBD | FCBT	FA-LR	0.7570	0.5439	0.7953	0.7386	0.7708	0.1480	0.2269	0.1185	0.1711	0.1166
MBX: UC vs nonIBD | FCBT	ICA-LR	0.8141	0.6460	0.8327	0.8162	0.8181	0.1183	0.2187	0.1254	0.1668	0.1102
MBX: UC vs nonIBD | FCBT	LR	0.7920	0.5785	0.8019	0.7943	0.7872	0.0951	0.1943	0.1171	0.1113	0.0980
MBX: UC vs nonIBD | FCBT	OPLS-DA	0.7744	0.5505	0.7874	0.7795	0.7705	0.0908	0.1583	0.0971	0.1319	0.0796
MBX: UC vs nonIBD | FCBT	PCA-LR	0.8085	0.6298	0.8549	0.7743	0.8124	0.0952	0.1947	0.1200	0.1033	0.0971
MBX: UC vs nonIBD | FCBT	RF	0.7943	0.5697	0.8016	0.8038	0.7824	0.1333	0.3139	0.1673	0.1388	0.1559
MBX: UC vs nonIBD | FCBT	RF-KBS	0.7979	0.5880	0.7982	0.8090	0.7922	0.1318	0.2689	0.1470	0.1461	0.1357
MBX: UC vs nonIBD | FCBT	RTF	0.8149	0.6255	0.8267	0.8162	0.8105	0.1339	0.2856	0.1602	0.1452	0.1433
MBX: UC vs nonIBD | LogNorm	FA-LR	0.7675	0.5518	0.7950	0.7529	0.7746	0.1305	0.2356	0.1193	0.1533	0.1195
MBX: UC vs nonIBD | LogNorm	ICA-LR	0.8173	0.6261	0.7984	0.8505	0.8077	0.1049	0.2157	0.1164	0.1423	0.1069
MBX: UC vs nonIBD | LogNorm	LR	0.7892	0.5717	0.7965	0.7943	0.7836	0.0932	0.1903	0.1152	0.1113	0.0959

MBX: UC vs nonIBD | LogNorm	OPLS-DA	0.7744	0.5505	0.7874	0.7795	0.7705	0.0908	0.1583	0.0971	0.1319	0.0796
MBX: UC vs nonIBD | LogNorm	PCA-LR	0.8265	0.6547	0.8583	0.8019	0.8268	0.0823	0.1798	0.1135	0.0747	0.0906
MBX: UC vs nonIBD | LogNorm	RF	0.7910	0.5641	0.7906	0.8038	0.7808	0.1429	0.3147	0.1637	0.1511	0.1571
MBX: UC vs nonIBD | LogNorm	RF-KBS	0.7612	0.5173	0.7735	0.7610	0.7577	0.1510	0.3030	0.1684	0.1572	0.1528
MBX: UC vs nonIBD | LogNorm	RTF	0.8205	0.6401	0.8299	0.8233	0.8177	0.1355	0.2817	0.1548	0.1521	0.1412
MBX: UC vs nonIBD | PQN	FA-LR	0.7013	0.3711	0.7133	0.7119	0.6805	0.0788	0.1883	0.1294	0.1040	0.0961
MBX: UC vs nonIBD | PQN	ICA-LR	0.6606	0.2982	0.6601	0.6729	0.6479	0.1161	0.2203	0.1196	0.1360	0.1106
MBX: UC vs nonIBD | PQN	LR	0.6513	0.3120	0.6872	0.6419	0.6508	0.1566	0.3386	0.1845	0.1836	0.1677
MBX: UC vs nonIBD | PQN	OPLS-DA	0.7752	0.5527	0.8133	0.7529	0.7738	0.1223	0.2547	0.1494	0.1279	0.1275
MBX: UC vs nonIBD | PQN	PCA-LR	0.6707	0.3080	0.6595	0.7000	0.6508	0.1188	0.2398	0.1185	0.1496	0.1190
MBX: UC vs nonIBD | PQN	RF	0.7874	0.5567	0.7889	0.7971	0.7772	0.1445	0.3193	0.1640	0.1526	0.1591
MBX: UC vs nonIBD | PQN	RF-KBS	0.7698	0.5295	0.7727	0.7757	0.7640	0.1448	0.2942	0.1563	0.1530	0.1483
MBX: UC vs nonIBD | PQN	RTF	0.8077	0.6174	0.8290	0.7957	0.8080	0.1422	0.2891	0.1633	0.1436	0.1452
MBX: UC vs nonIBD | Raw Data	FA-LR	0.7013	0.3711	0.7133	0.7119	0.6805	0.0788	0.1883	0.1294	0.1040	0.0961
MBX: UC vs nonIBD | Raw Data	ICA-LR	0.6606	0.2982	0.6601	0.6729	0.6479	0.1161	0.2203	0.1196	0.1360	0.1106
MBX: UC vs nonIBD | Raw Data	LR	0.6513	0.3120	0.6872	0.6419	0.6508	0.1566	0.3386	0.1845	0.1836	0.1677
MBX: UC vs nonIBD | Raw Data	OPLS-DA	0.7752	0.5527	0.8133	0.7529	0.7738	0.1223	0.2547	0.1494	0.1279	0.1275
MBX: UC vs nonIBD | Raw Data	PCA-LR	0.6707	0.3080	0.6595	0.7000	0.6508	0.1188	0.2398	0.1185	0.1496	0.1190
MBX: UC vs nonIBD | Raw Data	RF	0.7874	0.5567	0.7889	0.7971	0.7772	0.1445	0.3193	0.1640	0.1526	0.1591
MBX: UC vs nonIBD | Raw Data	RF-KBS	0.7698	0.5295	0.7727	0.7757	0.7640	0.1448	0.2942	0.1563	0.1530	0.1483
MBX: UC vs nonIBD | Raw Data	RTF	0.8077	0.6174	0.8290	0.7957	0.8080	0.1422	0.2891	0.1633	0.1436	0.1452

MBX: UC vs nonIBD | Rel	FA-LR	0.7321	0.4684	0.7550	0.7262	0.7278	0.1470	0.2804	0.1564	0.1758	0.1379
MBX: UC vs nonIBD | Rel	ICA-LR	0.6651	0.2852	0.6531	0.6919	0.6408	0.1522	0.3332	0.1588	0.1672	0.1633
MBX: UC vs nonIBD | Rel	LR	0.7096	0.3822	0.6822	0.7586	0.6853	0.1616	0.3280	0.1677	0.1881	0.1634
MBX: UC vs nonIBD | Rel	OPLS-DA	0.7336	0.4596	0.7679	0.7110	0.7289	0.1668	0.3684	0.1968	0.1559	0.1833
MBX: UC vs nonIBD | Rel	PCA-LR	0.6441	0.2802	0.6267	0.6910	0.6346	0.2123	0.3864	0.1965	0.2604	0.1890
MBX: UC vs nonIBD | Rel	RF	0.7938	0.5652	0.7964	0.8033	0.7818	0.1346	0.3053	0.1659	0.1273	0.1533
MBX: UC vs nonIBD | Rel	RF-KBS	0.7958	0.6039	0.8140	0.8024	0.7974	0.1501	0.2968	0.1639	0.1850	0.1493
MBX: UC vs nonIBD | Rel	RTF	0.8190	0.6255	0.8125	0.8295	0.8127	0.1301	0.2666	0.1412	0.1285	0.1343
MBX: UC vs nonIBD | SBT	FA-LR	0.7684	0.5193	0.7614	0.7805	0.7565	0.1136	0.2219	0.1087	0.1347	0.1103
MBX: UC vs nonIBD | SBT	ICA-LR	0.7353	0.4742	0.7448	0.7405	0.7340	0.1438	0.2740	0.1396	0.1782	0.1359
MBX: UC vs nonIBD | SBT	LR	0.8119	0.6413	0.8492	0.7862	0.8193	0.1515	0.2642	0.1354	0.1689	0.1330
MBX: UC vs nonIBD | SBT	OPLS-DA	0.7113	0.4452	0.7785	0.6838	0.7174	0.1431	0.3109	0.1827	0.1675	0.1529
MBX: UC vs nonIBD | SBT	PCA-LR	0.7316	0.4476	0.7246	0.7462	0.7213	0.1162	0.2311	0.1176	0.1416	0.1135
MBX: UC vs nonIBD | SBT	RF	0.7874	0.5567	0.7889	0.7971	0.7772	0.1445	0.3193	0.1640	0.1526	0.1591
MBX: UC vs nonIBD | SBT	RF-KBS	0.7698	0.5295	0.7727	0.7757	0.7640	0.1448	0.2942	0.1563	0.1530	0.1483
MBX: UC vs nonIBD | SBT	RTF	0.8077	0.6174	0.8290	0.7957	0.8080	0.1422	0.2891	0.1633	0.1436	0.1452



Supplemenatry Figure 4.8. Critical difference between models build for predicting phenotype based on metabolomic profiles.

Supplementary Table 4.2. Top ICA loadings for UC vs nonIBD with metabolites as sources. The weights are taken to 2 standard deviations from the mean and transformed to their absolute value.

CD	Metabolite	IC	Weight	UC	Metabolite	IC	Weight
0	eicosatrienoate	IC1	0.1723400207	0	stearate	IC13	0.2251969265
1	sphingosine-isomer2	IC1	0.1706317201	1	nonadecanoate	IC13	0.2157857473
2	docosapentaenoate	IC1	0.168800648	2	arachidate	IC13	0.2148027687
3	palmithoylethanolamide	IC1	0.1631793961	3	palmitate	IC13	0.2023151499
4	palmitoylethanolamide	IC1	0.1605885992	4	17-methylstearate	IC13	0.1849058092
5	linoleoyl ethanolamide	IC1	0.1523316643	5	nervonic acid	IC13	0.1747971184
6	linoleoylethanolamide	IC1	0.1503143928	6	fumarate/maleate	IC13	0.1692744014
7	palmitoleate	IC1	0.136222285	7	eicosenoate	IC13	0.1691315205
8	sphingosine-isomer1	IC1	0.1335828701	8	oleate	IC13	0.1513342409
9	adrenate	IC1	0.1311870753	9	13-docosenoate	IC13	0.149570768
10	docosahexaenoate	IC1	0.129839287	10	malate	IC13	0.1472579823
11	eicosadienoate	IC1	0.1255578872	11	oxalate	IC13	0.1315086329
12	myristoleate	IC1	0.1228697361	12	10-nonadecenoate	IC13	0.1264219298
13	10-nonadecenoate	IC1	0.1193601686	13	phytanate	IC13	0.1188874959
14	phytanate	IC1	0.111757101	14	myristate	IC13	0.1136883641
15	arachidonate	IC1	0.1111453228	15	C20:0 LPE	IC13	0.1098077952
16	eicosapentaenoate	IC1	0.110879999	16	malonate	IC13	0.1026516257
17	eicosanedioate	IC1	0.1054568384	17	olmesartan	IC13	0.09880319401
18	12.13-diHOME	IC1	0.1033519056	18	linoleate	IC13	0.09454650606
19	5alpha-cholestan-3beta-ol	IC1	0.10312341	19	NH4_C16:1 MAG	IC13	0.09128807249
20	4-hydroxybenzeneacetonitrile	IC1	0.1022143958	20	tetradecanedioate	IC13	0.08979551146
21	suberate	IC1	0.1017912639	21	eicosadienoate	IC13	0.08952790936
22	phenylacetylglutamine	IC1	0.09888166298	22	3-hydroxyoctanoate	IC13	0.08840556137
23	2-hydroxyhexadecanoate	IC1	0.09425140684	0	ketodeoxycholate	IC2	0.2356543607
24	5-aminolevulinic acid	IC1	0.09346680892	1	chenodeoxycholate	IC2	0.2291600505
25	phytosphingosine	IC1	0.09186094122	2	cholate	IC2	0.2251047858

26	3-methyladipate/pimelate	IC1	0.08971929416	3	hyodeoxycholate/ursodeoxycholate	IC2	0.2059582606
27	9.10-diHOME	IC1	0.0885473599	4	lithocholate	IC2	0.166259314
0	NH4_C52:5 TAG	IC3	0.1911791761	5	N-methylproline	IC2	0.15908302
1	C52:5 TAG	IC3	0.1795588321	6	taurohyodeoxycholate/tauroursodeoxycholate	IC2	0.1527717886
2	C52:6 TAG	IC3	0.174479064	7	taurocholate	IC2	0.14196063
3	C54:6 TAG	IC3	0.1684031795	8	alpha-muricholate	IC2	0.1418591161
4	NH4_C56:6 TAG	IC3	0.1663153952	9	taurolithocholate	IC2	0.1184172052
5	NH4_C54:6 TAG	IC3	0.165646534	10	taurochenodeoxycholate	IC2	0.1158424758
6	NH4_C52:4 TAG	IC3	0.1623569544	11	glycoursodeoxycholate	IC2	0.1110751257
7	C50:4 TAG	IC3	0.159439191	12	atenolol	IC2	0.1097205733
8	NH4_C50:4 TAG	IC3	0.1587339027	13	4-pyridoxate	IC2	0.1056992169
9	C54:5 TAG	IC3	0.1490317046	14	glycochenodeoxycholate	IC2	0.1043793403
10	C50:3 TAG	IC3	0.1436045982	15	trigonelline	IC2	0.1037521252
11	NH4_C52:6 TAG	IC3	0.1410476493	16	ectoine	IC2	0.1030338937
12	NH4_C50:2 TAG	IC3	0.1401401674	17	deoxycholate	IC2	0.1014881474
13	C56:6 TAG	IC3	0.1379793018	18	pantothenate	IC2	0.1009645073
14	NH4_C50:3 TAG	IC3	0.1341326041	19	pipecolic acid	IC2	0.09943683225
15	C52:3 TAG	IC3	0.1339140171	20	12.13-diHOME	IC2	0.09884696306
16	C54:4 TAG	IC3	0.1297446469	21	oxymetazoline	IC2	0.0980670256
17	C50:2 TAG	IC3	0.1296710048	22	acetylcholine	IC2	0.09776819624
18	NH4_C52:3 TAG	IC3	0.1286981227	23	methylimidazole acetic acid	IC2	0.09593914516
19	C52:4 TAG	IC3	0.1192901566	24	riboflavin	IC2	0.09305898528
20	C51:2 TAG	IC3	0.1172967672	25	1.2.3.4-tetrahydro-beta-carboline-1.3-dicarboxylate	IC2	0.08926144481
21	NH4_C54:5 TAG	IC3	0.1171569299	0	C18:1 CE	IC7	0.2007309474
22	C50:5 TAG	IC3	0.1157058544	1	NH4_C16:0 CE	IC7	0.1905340177
23	NH4_C56:4 TAG	IC3	0.1092549142	2	NH4_C18:1 CE	IC7	0.1785183952
24	C52:2 TAG	IC3	0.1054155338	3	C22:4 CE	IC7	0.1776111072
25	NH4_C52:2 TAG	IC3	0.09933603515	4	C16:0 CE	IC7	0.1760505765

26	C54:3 TAG	IC3	0.0984088897 1	5	C18:2 CE	IC7	0.1745011035
27	NH4_C56:3 TAG	IC3	0.0962495159 1	6	C18:0 CE	IC7	0.1709378368
28	C54:2 TAG	IC3	0.09272716167	7	NH4_C18:0 CE	IC7	0.1708924111
29	C51:3 TAG	IC3	0.0925117874	8	NH4_C22:6 CE	IC7	0.1642264553
30	C48:2 TAG	IC3	0.0921782660 9	9	C20:3 CE	IC7	0.1629294982
31	C44:1 TAG	IC3	0.0882439549 3	10	C18:3 CE	IC7	0.1608066661
0	C22:1 SM	IC6	0.1714234574	11	C20:4 CE	IC7	0.1601831066
1	C20:0 SM	IC6	0.1672335699	12	C22:6 CE	IC7	0.1590207646
2	heptadecanoate	IC6	0.1615576824	13	C16:1 CE	IC7	0.1568889782
3	C22:0 SM	IC6	0.1562406137	14	C22:5 CE	IC7	0.155469127
4	hydroxymyristate	IC6	0.1481921518	15	NH4_C16:1 CE	IC7	0.1552822665
5	2-hydroxyhexadecanoate	IC6	0.1424553177	16	NH4_C18:3 CE	IC7	0.1520640791
6	pentadecanoate	IC6	0.1377203121	17	NH4_C18:2 CE	IC7	0.1485851279
7	hyodeoxycholate/ursodeoxycholate	IC6	0.1376130385	18	NH4_C20:4 CE	IC7	0.1456293264
8	alloisoleucine	IC6	0.128175241	19	NH4_C22:4 CE	IC7	0.136513491
9	C16:0 SM	IC6	0.1249952397	20	NH4_C22:5 CE	IC7	0.1271931792
10	C16:1 SM	IC6	0.1236673779	21	NH4_C20:3 CE	IC7	0.1175927704
11	C14:0 SM	IC6	0.1191233848	22	C20:5 CE	IC7	0.1169912185
12	glycolithocholate	IC6	0.1181332621	23	C14:0 CE	IC7	0.1145628173
13	acetylcholine	IC6	0.1175144571	24	NH4_C20:5 CE	IC7	0.1115988762
14	C16:1 MAG	IC6	0.1134412628	25	NH4_C53:3 TAG	IC7	0.09451443514
15	hippurate	IC6	0.1126672814	26	NH4_C53:2 TAG	IC7	0.09193711721
16	NH4_C49:1 TAG	IC6	0.1105021852	27	NH4_C51:2 TAG	IC7	0.0878859856
17	C24:1 SM	IC6	0.109720217	28	C56:7 TAG	IC7	0.08770020279
18	phytanate	IC6	0.108244512	0	C34:1 PC	IC17	0.1648996997
19	NH4_C50:5 TAG	IC6	0.107613969	1	C36:1 PC	IC17	0.1617996355
20	5alpha-cholestan-3beta-ol	IC6	0.105622764	2	C32:0 PC	IC17	0.1595068826
21	glycodeoxycholate	IC6	0.1045984754	3	C36:2 PC	IC17	0.1584505303
22	chenodeoxycholate	IC6	0.1041166062	4	C34:1 PC plasmalogen	IC17	0.157769903
23	C24:0 SM	IC6	0.1023389856	5	C36:2 PC plasmalogen	IC17	0.1539508206
24	C18:0 SM	IC6	0.101163135	6	C32:1 PC	IC17	0.1493422782
25	C34:4 PC plasmalogen	IC6	0.0986871674 8	7	C38:2 PC	IC17	0.1416527131

26	metronidazole	IC6	0.0984395647	8	C16:0 SM	IC17	0.1414927676
27	ketodeoxycholate	IC6	0.09320444165	9	C18:0 LPE	IC17	0.1383652931
28	indoleacetate	IC6	0.090741180772	10	C16:1 SM	IC17	0.137540197
29	valerate/isovalerate	IC6	0.08947277771	11	C34:2 PC plasmalogen	IC17	0.1358876752
30	2-aminoheptanoic acid	IC6	0.087791773688	12	C22:0 SM	IC17	0.1328138091
0	N-methylproline	IC2	0.1732328731	13	C38:4 PC plasmalogen	IC17	0.1317636956
1	taurohyodeoxycholate/tauroursodeoxycholate	IC2	0.1653491454	14	C36:3 PC	IC17	0.1313663523
2	taurocholate	IC2	0.1619438974	15	C36:1 PC plasmalogen	IC17	0.1254489674
3	taurochenodeoxycholate	IC2	0.1592194826	16	C36:4 PC plasmalogen	IC17	0.1245386363
4	taurodeoxycholate	IC2	0.1588787361	17	C36:5 PC plasmalogen	IC17	0.1237390002
5	alloisoleucine	IC2	0.1533717983	18	C18:0 SM	IC17	0.1233155677
6	tauro-alpha-muricholate/tauro-beta-muricholate	IC2	0.1484505661	19	C38:4 PC	IC17	0.122688541
7	acetylcholine	IC2	0.1384824974	20	C34:0 PC	IC17	0.1208025627
8	hexadecanedioate	IC2	0.1352478998	21	C34:3 PC	IC17	0.1185750056
9	taurolithocholate	IC2	0.1280712948	22	C34:2 PC	IC17	0.1184449847
10	pterin	IC2	0.1244410105	23	C30:0 PC	IC17	0.1181826314
11	N-acetylhistidine	IC2	0.1159922966	24	C24:1 SM	IC17	0.1176472406
12	hippurate	IC2	0.1140534845	25	C18:0 LPC	IC17	0.1164154175
13	glycocholate	IC2	0.1138567355	26	C14:0 SM	IC17	0.1159214735
14	glycochenodeoxycholate	IC2	0.1136217042	27	C24:0 SM	IC17	0.1155899642
15	dimethylglycine	IC2	0.1129830638	28	C20:0 SM	IC17	0.1091444296
16	5-aminolevulinic acid	IC2	0.1095677945	29	C22:0 LPE	IC17	0.1071643288
17	pipecolic acid	IC2	0.1089564682	30	C18:1 LPE	IC17	0.1041956225
18	taurine	IC2	0.1083012845	31	C20:3 LPC	IC17	0.1022587489
19	N6-acetyllysine	IC2	0.103869968	32	deoxycholate	IC17	0.09671476666
20	creatinine	IC2	0.1020016269	33	C36:4 PC-B	IC17	0.09671135541
21	C20:4 carnitine	IC2	0.0999049732	34	4-pyridoxate	IC17	0.09429848645
22	C3 carnitine	IC2	0.098880172224	35	C9 carnitine	IC17	0.09391792716
23	glycoursodeoxycholate	IC2	0.097489400558	36	C18:1 SM	IC17	0.09267960553

24	tetrahydro-1-methyl-beta-carboline-3-carboxylate	IC2	0.09714603872	37	C22:1 SM	IC17	0.08910133844
25	olmesartan	IC2	0.09523355004	0	4-guanidinobutanoic acid	IC18	0.1849762577
26	trimethylamine-N-oxide	IC2	0.094651296115	1	cortisol	IC18	0.1602664299
27	imidazoleacetic acid	IC2	0.094399873982	2	C16 carnitine	IC18	0.1474481336
28	C20 carnitine	IC2	0.0940911539	3	C18 carnitine	IC18	0.135657182
29	glycodeoxycholate	IC2	0.091444870884	4	suberate	IC18	0.1328012152
30	serotonin	IC2	0.09024932589	5	metformin	IC18	0.1290957698
31	chenodeoxycholate	IC2	0.0887472313	6	metronidazole	IC18	0.1279809024
0	C34:1 DAG	IC10	0.1734706299	7	gabapentin	IC18	0.1257528136
1	C34:2 DAG	IC10	0.1701702701	8	21-deoxycortisol	IC18	0.1239567698
2	NH4_C36:1 DAG	IC10	0.1678197529	9	azelate	IC18	0.123597865
3	C36:2 DAG	IC10	0.1676386823	10	C18:1 carnitine	IC18	0.1161172516
4	C36:3 DAG	IC10	0.1642294709	11	4-hydroxybenzaldehyde	IC18	0.1153016278
5	C32:2 DAG	IC10	0.1616643834	12	shikimate	IC18	0.1149799878
6	NH4_C36:3 DAG	IC10	0.1614093317	13	C18:1 LPE	IC18	0.1137089446
7	NH4_C34:2 DAG	IC10	0.1610646115	14	homovanillate	IC18	0.1130027166
8	C36:1 DAG	IC10	0.1596390214	15	hexadecanedioate	IC18	0.1110736494
9	C36:4 DAG	IC10	0.1578744997	16	pyridoxine	IC18	0.1099179282
10	NH4_C32:1 DAG	IC10	0.1575201141	17	cholesterol	IC18	0.1082352912
11	C34:3 DAG	IC10	0.1547739396	18	N-acetylglutamine	IC18	0.1059054258
12	NH4_C34:3 DAG	IC10	0.1533284134	19	masilinate	IC18	0.1053687996
13	C32:1 DAG	IC10	0.1510448075	20	C18:1-OH carnitine	IC18	0.105282717
14	NH4_C36:2 DAG	IC10	0.1505554696	21	4-hydroxybenzeneacetonitrile	IC18	0.1050962488
15	NH4_C32:2 DAG	IC10	0.1493148436	22	C18:0 MAG	IC18	0.1045362571
16	NH4_C34:1 DAG	IC10	0.1452657641	23	NH4_C18:0 MAG	IC18	0.1041398937
17	NH4_C36:4 DAG	IC10	0.1439608335	24	4-aminophenol	IC18	0.103815935

18	3-methylglutaconate	IC10	0.1429828029	25	trimethylbenzene	IC18	0.09894614044
19	C30:0 DAG	IC10	0.1357052405	26	acetyl-galactosamine	IC18	0.09717170553
20	NH4_C18:0 MAG	IC10	0.1291626968	27	hippurate	IC18	0.09564303683
21	C18:0 MAG	IC10	0.1291479495	28	C18:2 carnitine	IC18	0.09555173563
22	C32:0 DAG	IC10	0.1098921319	29	2-aminoadipate	IC18	0.09493140186
23	2-hydroxyglutarate	IC10	0.1092001002	30	C18:2 LPE	IC18	0.09426297674
24	3-hydroxymethylglutarate	IC10	0.1084628347	31	C16:1 LPE	IC18	0.09416043979
25	C34:0 DAG	IC10	0.1062751718	32	13-cis-retinoic acid	IC18	0.09293804626
26	10-heptadecenoate	IC10	0.093879600 91	33	C16-OH carnitine	IC18	0.08997261999
27	linoleate	IC10	0.092709465 04	34	myristoleate	IC18	0.08994670672
0	hypoxanthine	IC16	0.1960222681	35	C56:7 TAG	IC18	0.08906457228
1	2'-O-methyladenosine	IC16	0.1949465598	0	methionine	IC1	0.2045420393
2	8-hydroxy-deoxyguanosine	IC16	0.1873478179	1	glutamine	IC1	0.1986374793
3	5-acetylamino-6-amino-3-me thyluracil	IC16	0.175932787	2	leucine	IC1	0.1955400544
4	1-methylguanine	IC16	0.1651955836	3	isoleucine	IC1	0.1899661623
5	1-methylguanosine	IC16	0.1522032787	4	threonine	IC1	0.1832168747
6	xanthine	IC16	0.1494644609	5	phenylalanine	IC1	0.1829898498
7	1-3-7-trimethylurate	IC16	0.1484802391	6	citrulline	IC1	0.1769021063
8	hydroxycotinine	IC16	0.1385520194	7	tryptophan	IC1	0.1715365634
9	caffeine	IC16	0.1375654145	8	tyrosine	IC1	0.1698616556
10	2-deoxyadenosine	IC16	0.1359403779	9	lysine	IC1	0.169847571
11	shikimate	IC16	0.1355995912	10	serine	IC1	0.1689354676
12	theophylline	IC16	0.1350725754	11	alanine	IC1	0.1626914258
13	porphobilinogen	IC16	0.1299797038	12	methionine sulfoxide	IC1	0.1225882452
14	acetytyrosine	IC16	0.1292386949	13	histidine	IC1	0.1183561492
15	uracil	IC16	0.1259877352	14	saccharin	IC1	0.1126574125
16	inosine	IC16	0.1233405866	15	glycine	IC1	0.1124380402
17	xylose	IC16	0.1225787342	16	aspartate	IC1	0.1084524295
18	guanine	IC16	0.1207669801	17	acetytyrosine	IC1	0.1078335983
19	13-cis-retinoic acid	IC16	0.1126240602	18	valine	IC1	0.105903855

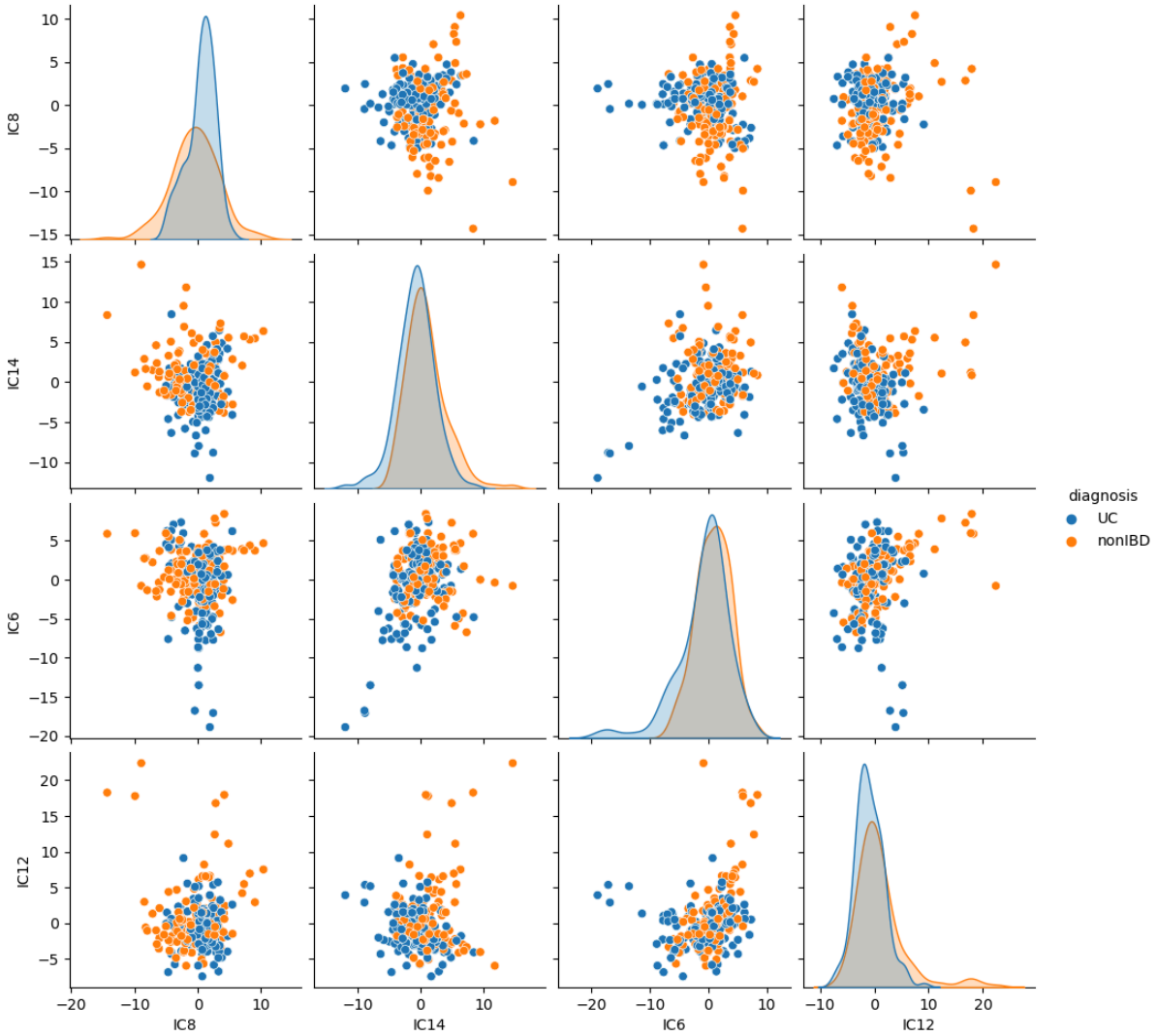
20	norepinephrine	IC16	0.1125912905	19	pyrocatechol	IC1	0.1024795046
21	phytanate	IC16	0.1113842117	20	4-hydroxybenzaldehyde	IC1	0.1016324301
22	fructose/glucose/galactose	IC16	0.1110316143	21	proline	IC1	0.09835692928
23	N-acetylglutamine	IC16	0.1052783951	22	4-aminophenol	IC1	0.09831498875
24	cotinine	IC16	0.1052192394	23	asparagine	IC1	0.09783403407
25	corticosterone	IC16	0.0963487515 7	24	5-hydroxytryptophan	IC1	0.09354204049
26	thymine	IC16	0.09567176187	0	imidazoleacetic acid	IC8	0.1990694621
27	oxypurinol	IC16	0.0956564773 6	1	guanosine	IC8	0.1956044667
28	masilinate	IC16	0.0952949461 9	2	2'-O-methyladenosine	IC8	0.18543885
29	guanosine	IC16	0.0921749244	3	3-methylxanthine	IC8	0.171287925
30	tetrahydro-1-methyl-beta-carboline-3-carboxylate	IC16	0.0910998076 5	4	hypoxanthine	IC8	0.1635291179
31	4-nitrophenol	IC16	0.0901315108 8	5	8-hydroxy-deoxyguanosine	IC8	0.1630872985
32	glycerate	IC16	0.0882866184 6	6	N-acetylhistamine	IC8	0.1557581016
0	isoleucine	IC7	0.2095094687	7	inosine	IC8	0.1518006001
1	leucine	IC7	0.2038344678	8	histamine	IC8	0.1505165189
2	methionine	IC7	0.1982446152	9	7-methylxanthine	IC8	0.150239602
3	phenylalanine	IC7	0.188082129	10	pipecolic acid	IC8	0.1285162652
4	tyrosine	IC7	0.1747896742	11	cytosine	IC8	0.1261630794
5	alanine	IC7	0.1723949071	12	cytidine	IC8	0.124183716
6	threonine	IC7	0.1687410011	13	uridine	IC8	0.1235508932
7	tryptophan	IC7	0.1680189831	14	N-acetyalanine	IC8	0.1202770161
8	metformin	IC7	0.1556539536	15	oxypurinol	IC8	0.1172169358
9	methionine sulfoxide	IC7	0.147669256	16	spermidine	IC8	0.1142324412
10	glutamine	IC7	0.1474372076	17	uracil	IC8	0.1108153538
11	citrulline	IC7	0.1472559078	18	7-methylguanine	IC8	0.1049965435
12	serine	IC7	0.1367012312	19	2-deoxyadenosine	IC8	0.1006407008
13	urate	IC7	0.1252377572	20	N-methylproline	IC8	0.1001613552
14	proline	IC7	0.1207700195	21	taurolithocholate	IC8	0.09541828516
15	thymine	IC7	0.1202703178	22	diaminopimelate	IC8	0.09502067252
16	glycine	IC7	0.1134225094	23	quinine	IC8	0.09068178499

17	glutamate	IC7	0.1126004323	24	cadaverine	IC8	0.09060339549
18	lysine	IC7	0.1105432724	25	tauro-alpha-muricholate/tauro-beta-muricholate	IC8	0.08981373665
19	acetylalanine	IC7	0.1067432574	26	ribothymidine	IC8	0.08880309368
20	pyrocatechol	IC7	0.1061495207	27	taurohyodeoxycholate/tauroursodeoxycholate	IC8	0.08834971643
21	2-hydroxy-3-methylpentanoate	IC7	0.1042616488	0	phenylacetylglutamine	IC14	0.1976087623
22	2-aminobutyrate	IC7	0.09813400324	1	urate	IC14	0.1700876636
23	C20:4 carnitine	IC7	0.09461190597	2	C18:3 LPC	IC14	0.168198117
24	4-methylcatechol	IC7	0.09308534691	3	pseudouridine	IC14	0.1526471439
25	trimethylamine-N-oxide	IC7	0.09074545124	4	C16:0 LPC	IC14	0.1479415027
26	hyodeoxycholate/ursodeoxycholate	IC7	0.08995221943	5	erythronate	IC14	0.1463516069
27	beta-guanidinopropionic acid	IC7	0.08984432972	6	2-aminoheptanoic acid	IC14	0.1429996398
0	taurodeoxycholate	IC5	0.2209295176	7	C16:1 LPC plasmalogen	IC14	0.1407843493
1	glycodeoxycholate	IC5	0.2142056326	8	C22:6 LPE	IC14	0.1399964665
2	taurolithocholate	IC5	0.1999994324	9	C18:1 LPC plasmalogen	IC14	0.1386548952
3	ketodeoxycholate	IC5	0.1877490543	10	N-carbamoyl-beta-alanine	IC14	0.1360623441
4	caffeine	IC5	0.1770555862	11	mandelate	IC14	0.1359088287
5	lithocholate	IC5	0.1655287623	12	norepinephrine	IC14	0.1311528765
6	deoxycholate	IC5	0.1575044322	13	C14:0 LPC	IC14	0.13104104
7	N-carbamoyl-beta-alanine	IC5	0.1417898376	14	deoxycholate	IC14	0.1309800216
8	glycolithocholate	IC5	0.1406571036	15	C18:0 LPE-B	IC14	0.1281075464
9	alpha-muricholate	IC5	0.1359328204	16	C18:2 LPC	IC14	0.1215154908
10	trigonelline	IC5	0.1337312279	17	alanylalanine	IC14	0.1210992583
11	kynurenic acid	IC5	0.1267396464	18	C20:1 LPC	IC14	0.1202093292
12	2-hydroxy-3-methylbutyrate	IC5	0.1251401579	19	C18:0 LPC	IC14	0.1130938182
13	2-hydroxy-3-methylpentanoate	IC5	0.1241005493	20	C18:1 LPC	IC14	0.1092960868
14	cholate	IC5	0.1202743218	21	C16:1 LPC	IC14	0.1081349495
15	adipate	IC5	0.1182802112	22	tetrahydro-1-methyl-beta-carboline-3-carboxylate	IC14	0.1078944861

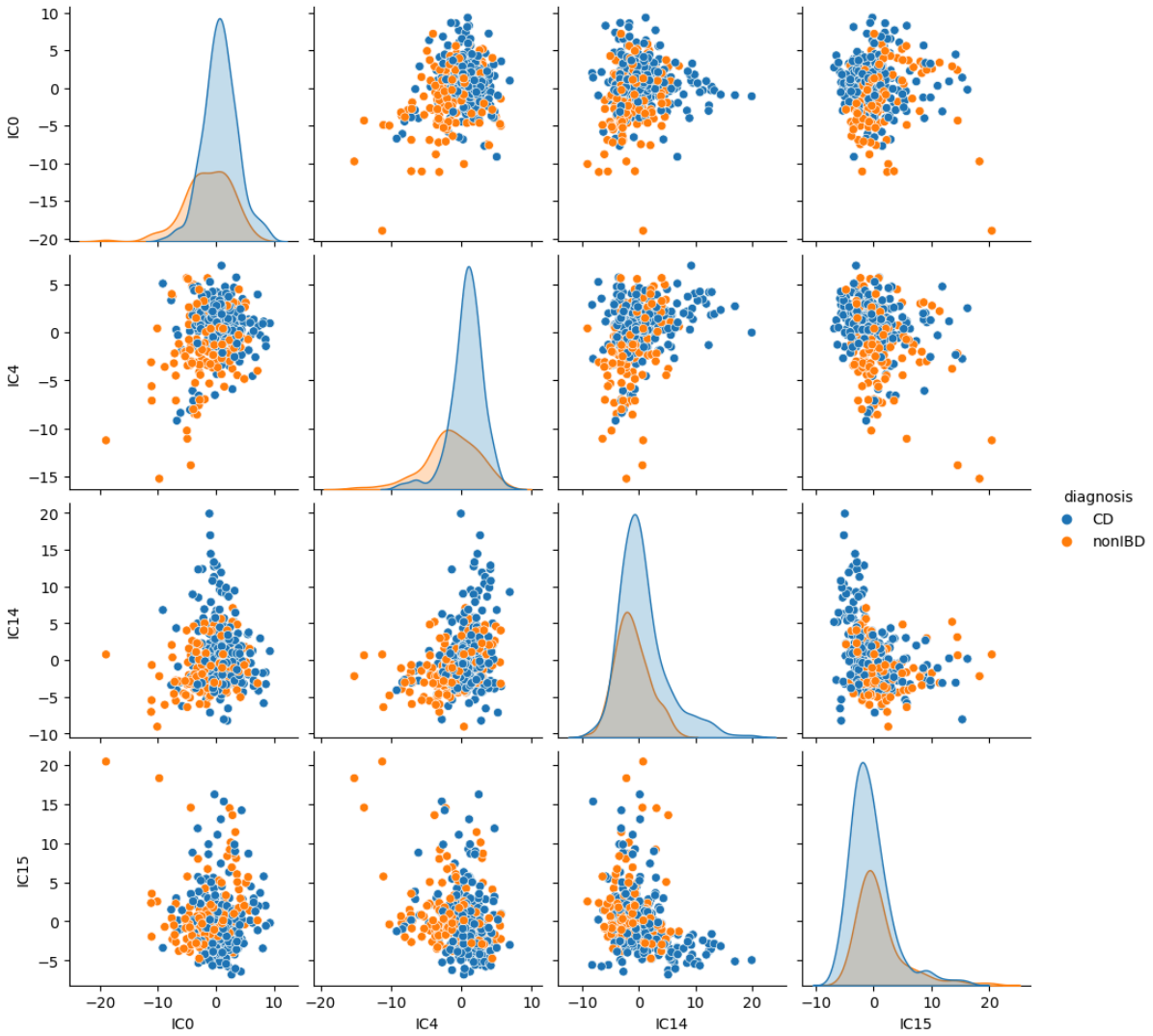
16	chenodeoxycholate	IC5	0.1126006781	23	lithocholate	IC14	0.1076641073
17	shikimate	IC5	0.1114744004	24	C20:3 LPC	IC14	0.1033516351
18	xanthurenate	IC5	0.1079547659	25	3-methylglutaconate	IC14	0.1012529913
19	p-hydroxyphenylacetate	IC5	0.1050846174	26	S-methylcysteine-S-oxide	IC14	0.1009959096
20	4-nitrophenol	IC5	0.1023208863	27	choline	IC14	0.09854785563
21	pyrocatechol	IC5	0.094256844 09	28	C18:0 LPE-A	IC14	0.09850188212
22	imidazole propionate	IC5	0.092986096 57	29	biliverdin	IC14	0.09520749592
23	urate	IC5	0.0901961378 5	30	pterin	IC14	0.09024622926
24	methylimidazole acetic acid	IC5	0.088609888 72	31	theophylline	IC14	0.08843301766
0	salicylate	IC17	0.2209026361	0	5-acetylamino-6-amino-3-methyluracil	IC6	0.1840580242
1	N-alpha-acetylarginine	IC17	0.1792701847	1	1-3-7-trimethylurate	IC6	0.1769212728
2	N-acetylglutamate	IC17	0.1587539618	2	1-methylguanine	IC6	0.1743534441
3	porphobilinogen	IC17	0.1581015927	3	carnosol	IC6	0.1639238447
4	ethyl glucuronide	IC17	0.1512058366	4	N-acetylhistidine	IC6	0.1495801548
5	acetaminophen	IC17	0.150878326	5	2-hydroxy-3-methylbutyrate	IC6	0.1454594399
6	erythronate	IC17	0.147749398	6	caffeine	IC6	0.1379540514
7	N6-acetyllysine	IC17	0.1406568101	7	norepinephrine	IC6	0.137699602
8	hydroxycotinine	IC17	0.138495966	8	imidazolelactate	IC6	0.1244922972
9	dTMP	IC17	0.1360995876	9	2-aminoheptanoic acid	IC6	0.1231068676
10	xanthurenate	IC17	0.1348113527	10	2-hydroxy-3-methylpentanoate	IC6	0.1230202199
11	serine	IC17	0.1319732186	11	pseudouridine	IC6	0.1220092999
12	nicotinate	IC17	0.1290856449	12	theophylline	IC6	0.1217883867
13	kynurenic acid	IC17	0.1227004685	13	serotonin	IC6	0.1172168955
14	alpha-ketoglutarate	IC17	0.119865859	14	tetrahydro-1-methyl-beta-carboline-3-carboxylate	IC6	0.1168803937
15	biliverdin	IC17	0.1186126715	15	C20:4 carnitine	IC6	0.1154574415
16	N-acetylglutamine	IC17	0.1184953466	16	lactate	IC6	0.1121870199
17	asparagine	IC17	0.1117822362	17	acetytyrosine	IC6	0.1112592866
18	biotin	IC17	0.1091600943	18	dTMP	IC6	0.1064419311

19	oxypurinol	IC17	0.1085871503	19	homovanillate	IC6	0.1050358224
20	4-aminophenol	IC17	0.1040094302	20	piperine	IC6	0.1035415352
21	acetyl-galactosamine	IC17	0.1031414796	21	sorbitol	IC6	0.1021348977
22	arginine	IC17	0.1028772564	22	xylose	IC6	0.09513731607
23	1-methylguanosine	IC17	0.1027622174	23	succinate	IC6	0.09101632944
24	C38:4 PC	IC17	0.1011985166	24	alpha-glycerophosphate	IC6	0.09044120583
25	caffeine	IC17	0.0984519060 4	25	erythronate	IC6	0.0902295169
26	saccharin	IC17	0.09611271903	26	adenine	IC6	0.08926113957
27	C36:4 PC-B	IC17	0.094072995 08	27	carnosol_isomer	IC6	0.08853305944
28	cinnamoylglycine	IC17	0.092859503 38	28	3-hydroxyoctanoate	IC6	0.08783003064
29	metronidazole	IC17	0.0917895080 1	0	homocitrulline	IC12	0.1579671004
30	N-acetylputrescine	IC17	0.0878867959 8	1	ADMA/SDMA	IC12	0.1460405531
31	uridine	IC17	0.0878395730 5	2	lithocholate	IC12	0.1440882848
0	C50:0 TAG	IC12	0.2457595287	3	deoxycholate	IC12	0.1429128907
1	NH4_C50:0 TAG	IC12	0.2237102778	4	NMMA	IC12	0.1401886124
2	C48:0 TAG	IC12	0.2082827263	5	homoarginine	IC12	0.1383160657
3	C52:0 TAG	IC12	0.206969148	6	N1,N12-diacetylspermine	IC12	0.1366533253
4	NH4_C48:0 TAG	IC12	0.1981949635	7	N6,N6-dimethyllysine	IC12	0.1359540814
5	NH4_C52:0 TAG	IC12	0.1831394892	8	dimethylglycine	IC12	0.1350646939
6	C46:0 TAG	IC12	0.1735384956	9	glycodeoxycholate	IC12	0.1334214427
7	C36:0 DAG	IC12	0.1669822711	10	putrescine	IC12	0.1291538199
8	NH4_C51:0 TAG	IC12	0.1654342683	11	alpha-glycerophosphocholine	IC12	0.1273922328
9	C34:0 DAG	IC12	0.1610701438	12	diacetylspermine	IC12	0.1247130381
10	gemfibrozil	IC12	0.1608483626	13	4-hydroxybenzaldehyde	IC12	0.118001211
11	NH4_C36:0 DAG	IC12	0.152509291	14	1-methylhistamine	IC12	0.1170059668
12	C32:0 DAG	IC12	0.1449096666	15	N6,N6,N6-trimethyllysine	IC12	0.1160654141
13	NH4_C46:0 TAG	IC12	0.1443376134	16	agmatine	IC12	0.1140126616
14	C51:0 TAG	IC12	0.1430519521	17	hydroxyproline	IC12	0.1133091012
15	C22:1 MAG	IC12	0.1157329135	18	myristoleate	IC12	0.1128160611

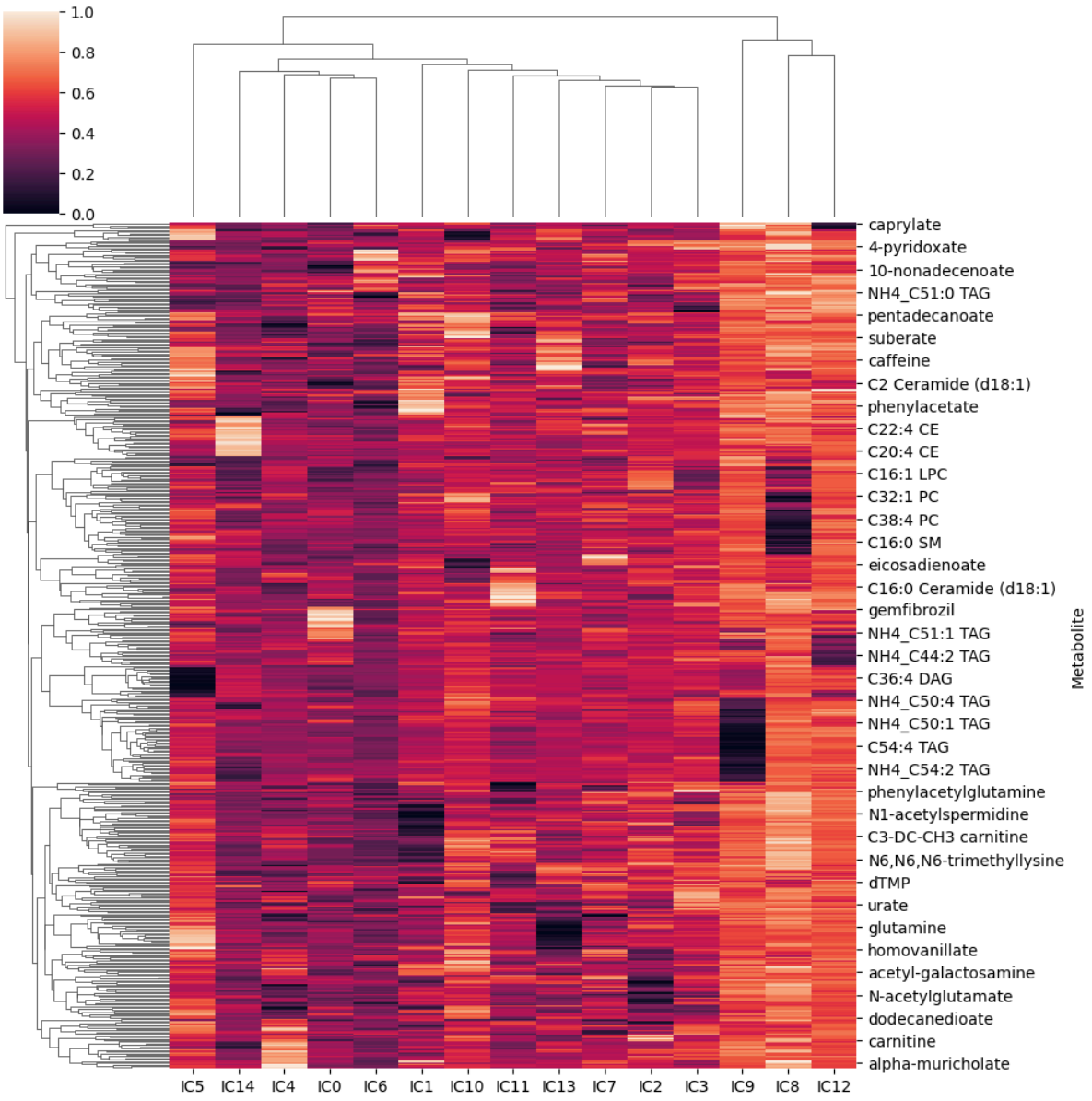
16	NH4_C22:1 MAG	IC12	0.1152327292	19	5-dodecenoate	IC12	0.1115506484
17	C2 Ceramide (d18:1)	IC12	0.1149915492	20	3-methylhistidine	IC12	0.1097459208
18	4-hydroxy-3-methylacetophenone	IC12	0.1002786602	21	taurodeoxycholate	IC12	0.105214357
19	C50:1 TAG	IC12	0.09248598759	22	alpha-ketoisovalerate	IC12	0.1051941071
20	linoleate	IC12	0.08961810777	23	beta-sitosterol	IC12	0.0999120889
21	oleate	IC12	0.0884812753	24	13-cis-retinoic acid	IC12	0.09864426952
				25	ornithine	IC12	0.09863725918
				26	N1-acetylspermidine	IC12	0.09647018852
				27	1-methylnicotinamide	IC12	0.0943410783
				28	biliverdin	IC12	0.0941627089
				29	histidinol	IC12	0.09374906752
				30	cholesterol	IC12	0.09336737772
				31	N-carbamoyl-beta-alanine	IC12	0.09161696595
				32	N6-acetyllysine	IC12	0.09035057531
				33	butyrobetaine	IC12	0.08993982656
				34	anserine	IC12	0.08879974912



Supplemenatry Figure 4.9. Inverse Kurtosis from ICA with metabolites as sources after FCBT between UC and nonIBD patients.



Supplementary Figure 4.10. Inverse Kurtosis from ICA with metabolites as sources after FCBT between CD and nonIBD patients.



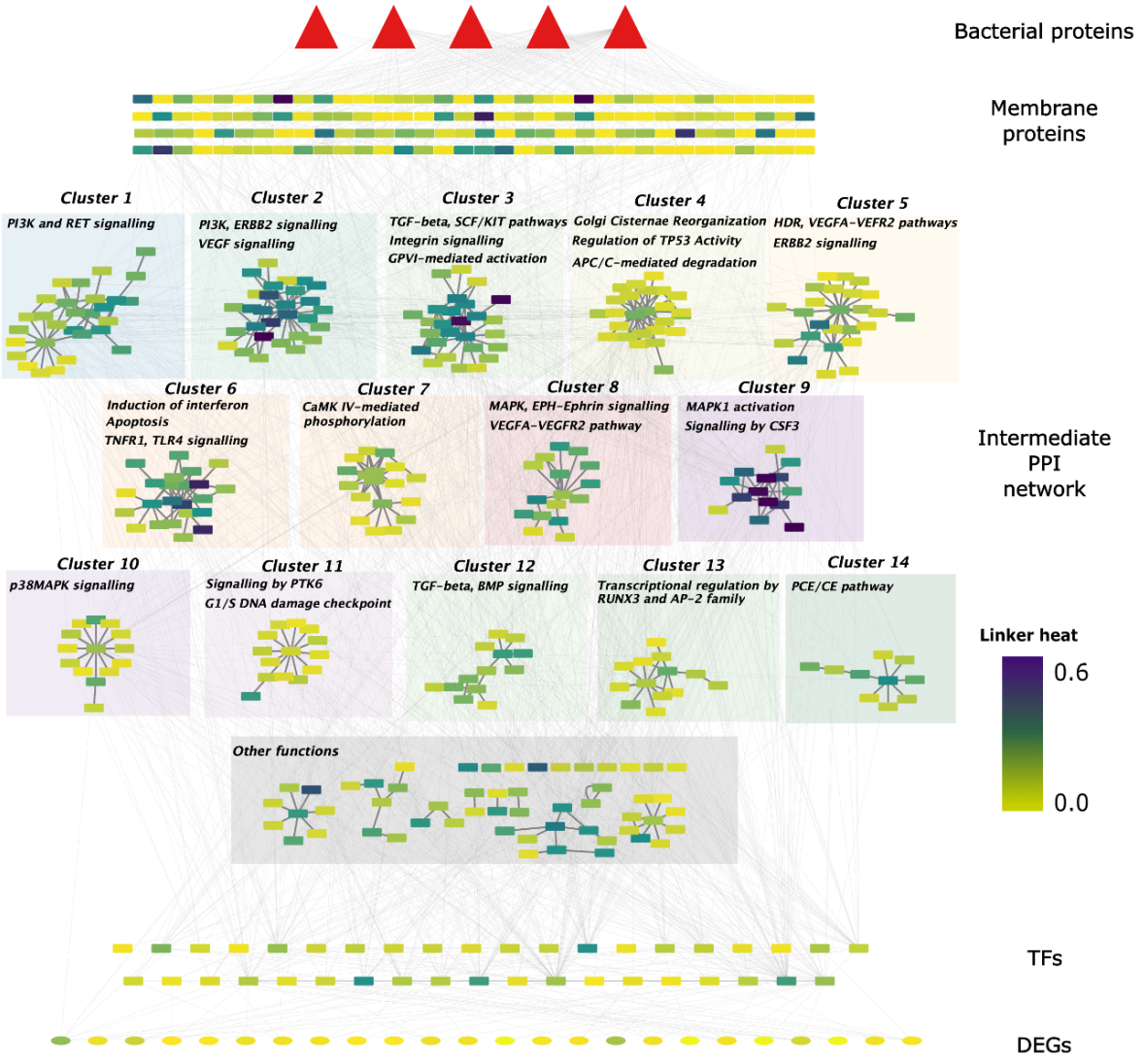
Supplementary Figure 4.11. Hierarchical clustering of the ICs extracted between UC and healthy controls when using Metabolites as the sources. The weights are then normalised to standard scale between 0 and 1. There are 4 clear clusters. Cluster 1 is presented by IC5; Cluster 2 by IC14, IC4, I0, and IC6; Cluster 3 IC1, IC10, IC11, IC13, IC7, IC2 and IC3; and Cluster 4 by IC9, IC8 and IC12. Hierarchical clustering was performed using the SciPy package, with the linkage method to use for calculating clusters set as average and the distance metric set as euclidean.

Appendix 4: Supplementary Chapter 5

Supplementary Table 5.1. 65 bacterial proteins that had domain-domain interactions. Protein identifiers that have been converted to their Uniprot IDs. Further information can be found at https://www.uniprot.org/

UniProt ID	UniProt ID contin.
B3JG17	J9GDD5
S8FI70	Q8A9B8
A0A0P0M566	W4UTT1
K9E3T6	A0A0K6BYZ4
A6KXF4	D7VKC1
A6L2K1	B2RIW6
G8UNS2	A0A0A2VU39
S3Z591	R9I6Y5
Q8A9V4	Q8AAU8
F3PQK3	A0A0P0GBI7
A6L8N5	B3JKV0
F4KUX7	I9S454
F3PQB4	A0A0K6BPJ8
Q8A1G0	E1WS50
W1I596	Q64TC3
A0A0K6BUV2	Q68H09
K1SR17	E6MPJ8
D4VDI0	E4T6E5
A0A0P0FT08	A0A0P0GT86
A0A078QIK9	F9P7L9
E6MQG1	F3QWA7
A0A0N7IF12	A0A127T5P2
D4V5A1	R5V370
A6KYX9	K1TGG7
S0GGR1	A0A0P0G4X6
R7AE42	D2QSG1
B5CXY1	Q9X4N3
A0A0P0M1P2	A0A132GWK0
R6E4C6	A0A108T7M9

J9C2S6	J9FKP8
Q89ZV6	A0A0P0M0P9
D1W412	C3QMG6
A0A076IWM7	



Supplementary Figure 5.1. Inferred multi-layer host-microbe protein-protein interaction (PPI) network from the source *Bacteroides vulgatus* microbial proteins to the host target proteins in ulcerative colitis. The full resulting PPI network from Microbiolink2 pipeline annotated for the both functional clusters in intermediate PPI network and heat (signal propagation) value for each protein in the network. The larger the value of heat propagation the more influence that protein has on the network.

Appendix 5: Peer-reviewed Publications

Attached peer-reviewed paper that appears in this thesis in Chapter 1.

Recent advances in clinical practice

Big data in IBD: big progress for clinical practice

Nasim Sadat Seyed Tabib ,[1] Matthew Madgwick,[2,3] Padhmanand Sudhakar,[1,2,3] Bram Verstockt ,[4,5] Tamas Korcsmaros,[2,3] Séverine Vermeire[1,5]

[1]Department of Chronic Diseases, Metabolism and Ageing, TARGID, KU Leuven, Leuven, Belgium
[2]Organisms and Ecosystems, Earlham Institute, Norwich, UK
[3]Gut microbes in health and disease, Quadram Institute Bioscience, Norwich, UK
[4]Translational Research in GastroIntestinal Disorders, KU Leuven, Leuven, Belgium
[5]Department of Gastroenterology and Hepatology, KU Leuven University Hospitals Leuven, Leuven, Belgium

Correspondence to
Dr Séverine Vermeire, Department of Chronic Diseases, Metabolism and Ageing - TARGID, KU Leuven, Leuven B-3000, Belgium; severine.vermeire@uzleuven.be

Received 11 October 2019
Revised 5 February 2020
Accepted 6 February 2020
Published Online First 28 February 2020

To cite: Seyed Tabib NS, Madgwick M, Sudhakar P, *et al*. *Gut* 2020;**69**:1520–1532.

ABSTRACT

IBD is a complex multifactorial inflammatory disease of the gut driven by extrinsic and intrinsic factors, including host genetics, the immune system, environmental factors and the gut microbiome. Technological advancements such as next-generation sequencing, high-throughput omics data generation and molecular networks have catalysed IBD research. The advent of artificial intelligence, in particular, machine learning, and systems biology has opened the avenue for the efficient integration and interpretation of big datasets for discovering clinically translatable knowledge. In this narrative review, we discuss how big data integration and machine learning have been applied to translational IBD research. Approaches such as machine learning may enable patient stratification, prediction of disease progression and therapy responses for fine-tuning treatment options with positive impacts on cost, health and safety. We also outline the challenges and opportunities presented by machine learning and big data in clinical IBD research.

INTRODUCTION

Precision medicine holds great promise to improve the landscape of IBD course of care for an individual patient, providing the most beneficial therapy while minimising the risk. The ultimate goals of precision medicine include stratifying patients based on disease subtypes and severity, disease progression and treatment response using personal and clinical data coupled with molecular profiling data of patients.[1 2] IBD, with its two main subtypes, Crohn's disease (CD) and UC, is a complex inflammatory disease with a wide range of contributing factors including host genetics, immune system, environmental exposures and the gut microbiome.[3–5] The inherent complexity of the disease introduces a large number of confounding factors, which stand in the way of accurate diagnosis and precision medicine.[6]

The term 'big data' is generally referred to as large volume of rapidly produced data from variable sources, known as the three 'V's (volume, velocity and variety).[7] Over the past decades, the production and availability of data that could inform healthcare has increased remarkably mainly due to technological advancements and falling costs of data generation. Most important sources of data in IBD comprise study cohorts, clinical trials, administrative and electronic health record databases, patient-reported outcomes databases, medical imaging databases and omics datasets (including genomics, transcriptomics, proteomics and metabolomics, as well as environmental omics) (figure 1). The use of big data in IBD allows medical researchers to reveal disease-related trends, associations and patterns to propel our understanding of IBD forward and to inform clinical practice.[2 8] However, due to the high complexity of big data and the long list of confounding factors, interpreting these data is not trivial and warrants approaches that can uncover hidden patterns in these large and complex datasets.[9]

Recent developments in computational biology have driven the integration of big data and molecular networks using the principles of systems biology and machine learning. Systems biology centres around the holistic and mathematical modelling of complex biological system.[10] Machine learning is a subset of artificial intelligence, which refers to the ability of algorithms to learn from data in order to detect patterns and make decisions (without explicitly being programmed what to do) (Box 1).[11] Machine learning algorithms provide the means and opportunity to investigate large amounts of data and thus help identify patterns behind complex medical conditions. These analytical approaches allow categorisation of patients based on their specific differences through screening a patient's genome, transcriptome, proteome, epigenome, immunome and microbiome. Integrating the omics datasets using systems biology-based approaches may advance understanding of the underlying causative factors in individual patients. The arrival of systems biology and machine learning into IBD clinical research has allowed researchers to capture complex associations andincreased understanding of disease mechanisms in IBD. In this narrative review, we provide an overview of the sources of big data in IBD. We discuss how artificial intelligence could help us better understand IBD pathogenesis and how some components of it have already begun to shape our knowledge of IBD. We address how artificial intelligence could contribute to the diagnosis and prognosis of IBD, and whether it could assist with predictions of therapy efficacy and adverse effects. As a final point, we argue the potential that artificial intelligence provides for personalised medicine in IBD and evaluate the feasibility of big data in IBD disease management.

ROLE OF MACHINE LEARNING AND SYSTEMS BIOLOGY IN THE INTERPRETATION OF BIG DATA IN IBD RESEARCH

The main challenge faced by many scientists is to extract meaningful information through integrating different sources of data and thereby discover disease association patterns. Classical statistical methods are not powerful enough to explain the

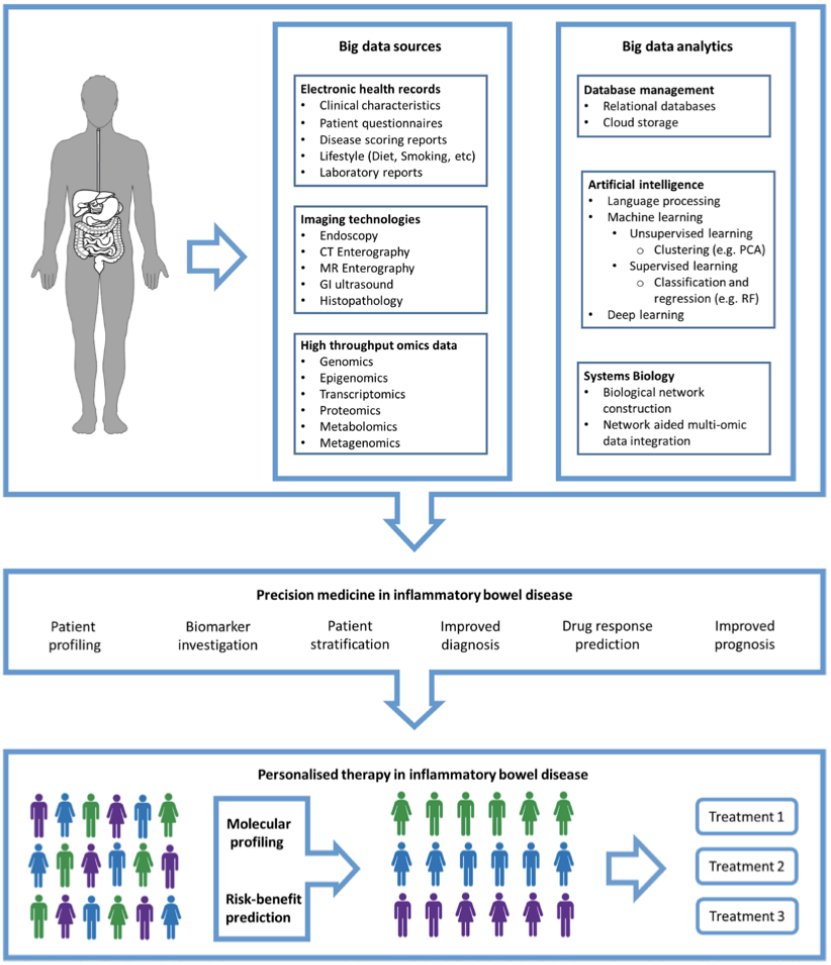


Figure 1 Precision medicine in IBD. Generation of big data from thousands of individuals, along with analytical advancements such as machine learning and systems biology, assists the application of precision medicine and therefore allows patient stratification for personalised therapeutic intervention and disease management strategies. MR, magnetic resonance; PCA, principal component analysis; RF, random forest.

underlying milieu of pathogenic and causative factors in IBD. Hence, scientists have adopted different analytical methodologies. Generally, such analytical methodologies are categorised into two main groups, namely, systems biology and machine learning. These are more powerful and flexible methods in biomedical data science and have the potential to uncover novel insights into disease pathogenesis.[12 13]

Systems biology paves the way for data integration and analysis from a functional perspective, and it has assisted in identifying the pathophysiological mechanisms of IBD. The approach of systems biology typically involves the use of networks (mostly molecular networks such as protein–protein interaction networks, regulatory networks involving transcription factors and metabolic networks) to capture the physical and signalling interactions and to interpret contextual measurements such as expression of genes, proteins and metabolites. This approach thereby provides a framework to identify key components and/or pathways which mediate the pathogenesis of the disease. Brooks *et al* identified different clusters of patients with UC using network footprints created by combining mutation data, protein–protein interaction networks and gene expression data.[14]

In the past decade, machine learning has attracted much attention from groups engaged in IBD research, owing to its ability to learn complex patterns and make prediction. With machine learning as a framework, several attempts have been made to use different types of omics and clinical datasets to improve our understanding of disease mechanism. Given that omics datasets, such as RNAseq data, comprise expression information of thousands of genes (features) with far few samples, feature selection is of great importance. Machine learning algorithms take

Box 1

Artificial intelligence terminology

Artificial intelligence: the field of computer science which concerns the theory and development of computers to perform tasks which usually requires human intelligence, such as imagine classification, speech recognition and decision-making.

Machine learning: a field of artificial intelligence which refers to the computers' ability to learn to make decisions or detect patterns (without explicitly being programmed) from data.

Deep learning: a subfield of machine learning that exploits many layers of non-linear information processing for supervised or unsupervised feature extraction and transformation, and for pattern analysis and classification using various neural network frameworks.

Supervised learning: the task of an algorithm learning a function that maps an input to an output based on provided example data.

Unsupervised learning: the task of a machine learning algorithm to learn the underlying data structure of unlabeled example data, for example, finding commonalities, leading to insights and therefore a greater understanding of the example data.

Classification: the process of predicting a class/subcategory of given data points from known example data.

Generalisation: refers to how well the machine learning model learns the underlying data and the model's ability to apply this to specific examples not seen by the model during training.

Ensemble learning: the union of homologous or heterogeneous machine learning algorithms whose predictions are combined to achieve greater performance than just the individual machine learning algorithm could achieve alone.

Support vector machine: this is a discriminative classifier which determines classes from a separating hyperplane. Through the use of a kernel, SVMs can be adapted to suit non-linear problems.

Random forests: a homologous ensemble algorithm which constructs a great number of decision trees at training.

Matrix factorisation: an algorithm which extracts meaningful association from an incomplete data matrix and transforms them in a lower dimensional latent space, also known as recommender systems.

advantage of data-dependent automatic feature learning, while systems biology approaches need to be manually programmed. Machine learning algorithms can learn how to integrate several predictors to identify a representative subset of input data[15]; for example, a machine learning algorithm using the concept of random forests identified a panel of 50 faecal bacteria capable of distinguishing active and remission states in patients with CD.[16]

Genomics

IBD is considered as a polygenic disease, with the exception of rare monogenic cases.[17] The notable example of research into the genetic basis of IBD is the introduction of *NOD2* as the first CD susceptibility gene.[18] To date, the continued search for genetic determinants of IBD identified 242 variants associated with IBD,[17] of which 45 have been fine mapped to statistically significant causal variants. Interestingly, associated regions indicate that there is a profound overlap between IBD and other immune-mediated inflammatory diseases. However, merely a small percentage of heritability is explained by the identified loci.[17]

To further resolve the genetic architecture of IBD, machine learning and data integration could be employed to propel the gene discoveries. The main issue with association studies is the imbalance between the number of patient samples and the number of single-nucleotide polymorphisms (SNPs) that are being analysed. In addition, the classical genotype–phenotype association at high statistical confidence neglects a considerable fraction of genetic variation. Machine learning could be used to detect meaningful patterns containing thousands of DNA variants, regardless of the statistical significance level.[19–22] This could result in predictions of genetic markers and variants with greater accuracy. An exemplary study was conducted using data from the International IBD Genetics Consortium's Immunochip project. To reduce the number of SNPs, Wei and colleagues applied a less rigid statistical confidence limit (p values of $<10^4$ and minor allele frequency of <0.01) followed by a machine learning classifier-based feature selection method (the penalised logistic regression model). The authors defined 573 SNP-based CD and 366 SNP-based UC predictive models with superior area under the receiver operating characteristic curve (AUC) values than the log OR-based models (AUCs of 0.86 (95% CI 0.85 to 0.86) and 0.826 (95% CI 0.81 to 0.83) for CD and UC, respectively).[23] Another interesting study was conducted using the UK Inflammatory Bowel Disease Genetics Consortium and UK10K consortium for the controls, which cumulatively comprises approximately 8000 individuals (4280 patients and 3652 controls). In this study, a machine learning model, a support vector machine (SVM), was used to hunt for novel genetic variants, which resulted in the identification of a missense variant in *ADCY7* associated with UC with a frequency of 0.6%.[24] A recent study reanalysed the Immunochip dataset using different machine learning models, including random forests and neural networks. Romagnoni *et al* identified new variants with minor effects, in addition to almost all of the previously known variants among the best predictors of CD.[25]

Advancements in sequencing technologies allow a more in-depth genomic screening. Scientists have used whole genome/exome sequencing particularly to discover rare genetic variants, such as *NOX1*, contributing to very early-onset IBD.[26] Machine learning methodologies, particularly deep learning, are resourceful tools for not only making predictions but also extracting biomedical insights.[27–30] In a notable publication, Zou *et al* provided a primer on deep learning for genomic data analysis accompanied with practical guidelines for the discovery of DNA-binding motifs.[31]

Transcriptomics and proteomics

Investigating the downstream effects of genomic aberrations, namely, on the transcriptome and proteome, provides additional molecular details to unravelling IBD pathogenesis. Differential gene expression analysis has been used to identify key genes and pathways underlying IBD pathogenesis. Transcriptomic analyses of human ileum and colonic samples have helped to uncover the roles of different pathways driving inflammation in IBD. For example, inflamed and non-inflamed tissues have altered gene expression in CD and UC. To investigate the functional significance of these modifications and to characterise their molecular signatures in colonic tissue, an integrated systems approach has highlighted significant enrichment in proteasome and apoptosis pathways.[32] With protein–protein interaction network analysis, Li *et al* identified *MAPK3*, *NDRG1* and *HLA-DRA* as key players in disease pathogenesis. Following a similar approach, Hong *et al* identified altered gene expression profiles and key cellular

pathways in patients with inflamed and non-inflamed intestinal mucosa with CD, including immune response, chemokine signalling and cell adhesion.[33]

Weighted gene coexpression network analysis allows researchers to detect genes that are upregulated or downregulated in tandem under specific conditions.[34 35] For example, Lin *et al* revealed important pathogenic roles for *IL-8* and *MMP-9* in the colonic tissues of patients with UC by combining gene coexpression and protein–protein interaction networks.[36] A similar study in the context of gene expression alteration in different stages of CD by Verstockt *et al* pinpointed that dysregulation of the coexpression network is more evident in newly diagnosed and late-stage CD compared with recurrent CD.[37] Likewise, this network approach can elucidate biological mechanisms driving treatment resistance to biological therapies, such as with tumour necrosis factor (TNF) inhibitor agents.[38] Another functional approach to explore the gene expression data is metabolism-level interpretation using Recon 2,[39] the model of the human metabolic network. Using this model, critical pathways such as cellular transport of thiamine and bile acid metabolism have been identified.[40]

Yuan and colleagues reported 41 discriminatory IBD-related genes by combining machine learning and systems biology. In searching for novel candidate genes, the authors used a two-step feature selection on microarray data from patients with CD, UC and control individuals. First, they ranked thousands of genes according to their correlation to diagnosis and the redundancies between each gene related to all other genes in the ranked list. Then, using an SVM as a machine learning classifier, they identified a feature set containing 21 genes, which yield the highest prediction accuracy. Additionally, based on the concept of functional similarity among closely related proteins, the authors used the protein–protein interaction network of the proteins encoded by those 21 genes and applied the shortest path approach (typically defined as the path with the least number of links between two proteins in a network) to find an additional 20 candidate genes.[41] In another interesting study by Isakov *et al*, novel candidate genes were identified by developing a machine learning model trained on expression values of known IBD susceptibility genes and their functional annotations. The authors used the feature importance of a machine learning classifier as the feature selection method.[42]

Environmental 'omics'

The gut microbiota, which comprise intestinal bacteria, fungi, archaea and viruses, is an essential part of the human GI tract and plays a pivotal role in human health. In homeostatic condition, there is a state of immunological tolerance to the commensal intestinal microbiota. It has been established that perturbation of composition, function and structure of the gut microbiota, known as dysbiosis, is one of the key players in IBD pathogenesis.[43] However, it is still not clear whether this dysbiosis is the cause or consequence in patients with IBD.

There is a decline in both species diversity and richness in patients with IBD. Several studies have reported an increase in the abundance of certain species from the Proteobacteria phylum, such as *Escherichia coli*, and a decline in anti-inflammatory butyrate-producing bacteria species, such as *Faecalibacterium prausnitzii*, belonging to the Firmicutes phylum. Additionally, a longitudinal study suggested an increase in dynamic fluctuation of the gut microbiome composition in patients with IBD.[44]

Much less is currently known on the role played by viruses in the dysbiotic state in patients with IBD. Recent advances in sequencing technologies and data analytic techniques have enabled in-depth characterisation of microbiota communities to investigate IBD pathogenesis using meta-level omics datasets, namely, metagenomics, metatranscriptomics, metaproteomics and metabolomics. Deep metagenomics paved the way to study gut resident fungi, archaea and viruses in both healthy and disease states. Different stool virome profiles have been observed in patients with IBD compared with healthy individuals.[45] Zuo and colleagues used machine learning-based clustering to define viral metacommunities in rectal mucosa derived from patients with UC. The predominant viral community among patients with UC showed decreased viral diversity, richness and evenness, particularly among *Caudovirales* species. However, two species of *Caudovirales* (*Escherichia phage* and *Enterobacteria phage*) were much more common among patients with UC compared with healthy controls. This suggests a loss of corelationship between the viruses and bacteria, which can cause microbiota dysbiosis and intestinal inflammation.[46]

The interplay between the microbial composition and metabolism of the gut is an interesting nexus in IBD. While much of the previous research on this interaction level has been interpretive in nature, most of the studies on the gut protein and metabolic composition used shotgun metagenomic technique. Thus, by comparing the abundance of enzymatic genes across samples, scientists have been able to infer the effect of variations in microbial composition on the protein and metabolic levels. An example of this is the study by Greenblum *et al* in which they used faecal metagenomics to build metabolic networks. They demonstrate topological differences by which IBD-associated metabolic networks interact with the gut environment and the host.[47] There is a growing number of investigations applying the approaches of metaproteomics and metabolomics. Particularly, there are two avenues in which metaproteomics-based investigations have been employed, the mucosal–luminal interface analysis and the stool metaproteome profiling. Li *et al* investigated the protein co-occurrence network at the mucosal surface of six different colonic regions. Employing weighted correlation network analysis and multiple clustering methods such as hierarchical clustering, they identified distinct functional protein modules (protein clusters that alter together) in association with non-IBD, UC and CD disease states.[48] In addition to systems biology methods, machine learning could be applied to define relevant protein clusters. Profiling of stool samples revealed that metaproteomic signatures in patients with CD differ from those of healthy individuals. By integrating metagenomics and metaproteomics, and applying a hierarchical clustering method, Erickson *et al* reported a depletion of several microbial proteins in patients with CD with ileal involvement, such as proteins in the butyrate pathway which corresponded to a reduction in the Firmicutes phylum.[49]

Multiomics data integration

In more recent investigations, researchers have been collecting different levels of omics data from patients with IBD to investigate the crosstalk between the key players in IBD pathogenesis. An interesting area in which multiomics data integration has been applied is to characterise the dysregulated multifaceted interactions between various host and microbial factors in IBD. For example, Häsler *et al* studied the transition of intestinal homeostasis to dysbiosis by integrating multiple levels of data, namely, the mucosal transcriptomic, post-transcriptional alterations and the mucosal microbiome of patients with UC and CD in comparison with healthy individuals. The authors identified

the enrichment of host transcript splicing events as a result of the interplay between microbial and host factors which probably mediate the transition of intestinal homeostasis to dysbiosis in patients with IBD.[50] In order to investigate the dysbiosis at the functional level, Lloyd-Price *et al* followed up 132 patients with IBD for 1 year and performed extensive molecular profiling of all patients. The authors revealed a distinctive upsurge in the ratio of facultative anaerobes to obligate anaerobes, along with disruptions at the molecular level, including microbial transcription division (within clostridia) and metabolite disruptions (acylcarnitines, bile acids, and short-chain fatty acids). Additionally, they reported noticeable alterations in the composition and function of microbiota with regard to different disease activity states.[51]

CURRENT PARADIGM OF IBD DISEASE MANAGEMENT AND ITS LIMITATIONS

The scope of IBD treatment is extending swiftly, with the introduction of new biologics and small molecules as a result of the improved understanding of the disease pathophysiology. With novel treatment options (targeting different aspects of IBD pathophysiology) such as anticytokine or chemokine agents, antiadhesion molecules, stem cell therapy and manipulation of the gut microbiota becoming increasingly available, it is time to move beyond the 'one-size-fits-all' approach.[52]

IBD management (figure 2) encompasses three different stages, starting with diagnosis, followed by the assessment of disease and the choice of therapy regimens, follow-up assessments and associated treatment changes, if necessary. Disease monitoring is key and is currently carried out by tracking different markers like faecal calprotectin, serum C reactive protein, also colonoscopy and/or medical imaging technologies such as abdominal ultrasound and MRI.[53 54] Hitherto, the clinical decision on the choice of therapeutic strategy depended on the response and tolerability of treatment in patients. However, in light of recent innovative therapies in IBD, a more accurate method is warranted to assist and complement existing management.[55] In recent years, there has been an increasing interest in the application of machine learning in IBD clinical research. Using machine learning for personalised predictions will not only strengthen medical care and improve outcomes but also considerably decrease healthcare expenditure. Despite the importance of health economics, there are little published data on the cost-effectiveness of artificial intelligence in healthcare. An interesting example is the study conducted by Bremer *et al*, who deployed a machine learning methodology to predict the individual outcome and

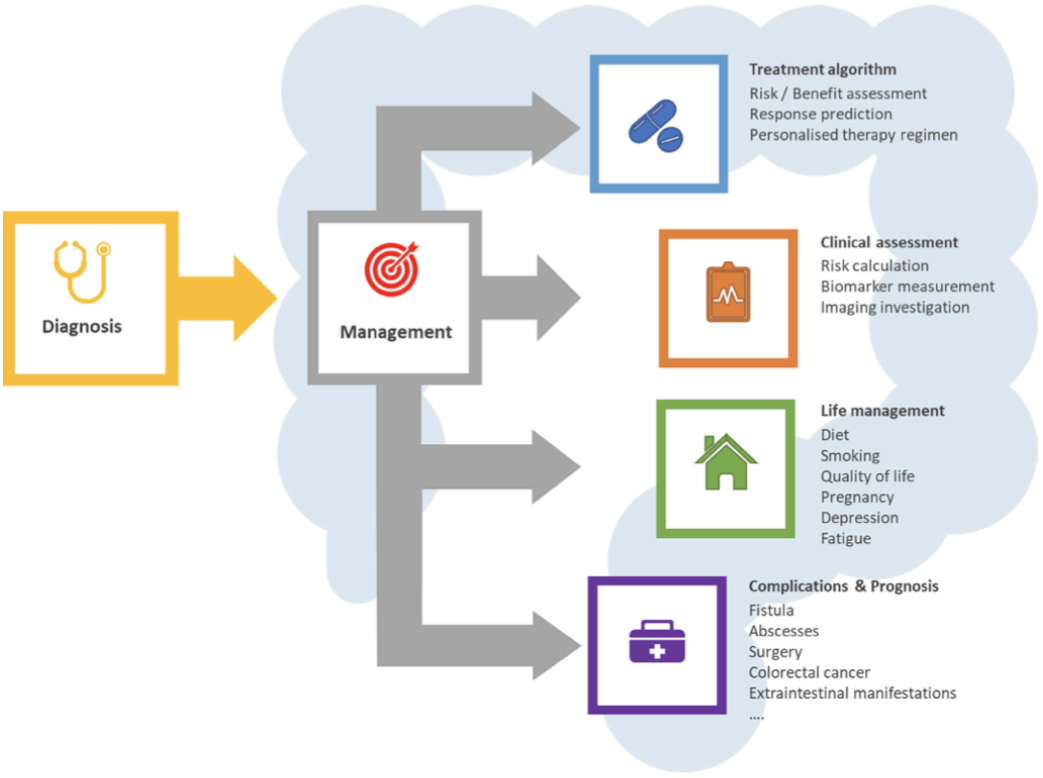


Figure 2 Clinical management of IBD from the point of diagnosis to life-term monitoring and follow-up. Each stage of the disease management process can potentially be subjected to precision medicine-aided improvement of patient care to reduce the socioeconomic burden on patients, clinicians and the healthcare system.

costs for patients with depressive disorders prior to the start of intervention in order to allocate patients to the most beneficial treatment.[56] In the field of gynaecology, Wang *et al* proposed a machine learning-based strategy for urinalysis which significantly increased the detection rate of the pathogen *Trichomonas vaginalis* in a cost-effective manner.[57]

Diagnosis and risk stratification

The current paradigm of IBD classification, which relies on invasive ileocolonoscopy and biopsies, does not adequately capture the broad spectrum of phenotypes of the disease or the patient-specific manifestations of its comorbid conditions. Recent research has focused on identifying and evaluating potential non-invasive diagnostic markers to diagnose IBD, differentiate it from other disorders and potentially improve its classification. There is great interest in the diagnostic value of genomics data, with over 240 IBD-associated risk loci already identified using genome wide association study (GWAS) data. A genotype–phenotype study associated three loci, *NOD2*, *MHC* and *MST*, with subphenotypes of IBD, particularly disease location.[58] Exome sequencing has arisen with the promise of unravelling the genetics of complex diseases. However, extracting disease-associated sequence variants is challenging due to inherited diversity of genomic variation. By incorporating exome sequence data with biological knowledge, such as functional interaction networks, into a matrix factorisation-based machine learning model, Jeong and Kim were able to distinguish patients with CD from healthy individuals (AUC=0.81).[59]

Likewise, molecular and cellular signatures can enable stratification of patients based on underlying pathways that drive their disease. Gene expression profiling is a major area of interest in the search of clinically associated signatures for IBD class prediction. To identify a set of genes distinguishing between UC and CD, novel machine learning-based methods have been used. Two examples which stand out are the PROPhet software package,[60] which automatically selects the best classifier and the optimal selection of genes to distinguish disease subtypes. Montero-Meléndez *et al* used this technique with microarray gene expression profiling of colonic biopsies to identify predictive transcriptional signatures associated with either CD or UC.[61] The second example is the Probabilistic Pathway Score, which is a pathway-based machine learning model that uses gene interactions to identify molecular pathways affected by the disease of interest and identify similarities and differences between them.[62] Proteomic signature is another promising nexus in biomarker research. Machine learning models have also been used with proteomic data to stratify patients with IBD. For example, Seeley *et al* investigated the protein signatures from colonic tissues using an SVM machine learning classifier trained on 25 peaks from histology-based mass spectrometry data. The model was able to discriminate patients with CD and UC from each other with an accuracy rate of 76.9%.[63] Another interesting area of biomarker research in IBD is microRNAs (miRNAs), a group of small noncoding RNA molecules which control gene expression and protein production and are detectable in many sources such as blood and urine. Hence, miRNAs hold great promise as potential non-invasive diagnostic markers. miRNAs are dysregulated in IBD.[64] Therefore, researchers have attempted to demonstrate the diagnostic value of circulating miRNAs signatures in the blood as diagnostic biomarkers using machine learning modelling, including random forests and SVM.[65 66]

An interesting example of exploring the diagnostic value of a set of biomarkers is the study conducted by Plevy *et al* combining genetic variants, serological and inflammatory markers to establish a diagnostic model to distinguish patients with IBD from those without IBD (healthy individuals or other diseases) and to separate patients with CD from UC. Based on the data from 1520 individuals, the authors selected 17 statistically significant markers and trained a random forest classifier, a machine learning algorithm, to differentiate the clinical groups.[67]

Machine learning approaches also hold great promise in unravelling disease-specific microbial signatures. Multiple machine learning-based microbiome frameworks have been established such as Multivariate Association with Linear Models (MaAsLin),[68] Metagenomic prediction Analysis based on Machine Learning[69] and phylogenetic convolutional neural networks[70] which incorporate patient clinical data, knowledge of microbial strains and knowledge of phylogenetic structure, respectively. Integrating additional information is expected to enhance the classification performance of microbiome-based machine learning models. As an example, Gevers *et al* were able to use rectal mucosa-associated microbiome signatures to distinguish paediatric patients with CD from patients with other GI tract conditions by integrating patient clinical data age, gender and past antibiotic use with the microbiome profiles using MaAsLin workflow.[71]

While the initial results of biomarker identification are promising, there is still a long way to go before these biomarkers can be applied in clinical practice, mainly due to the heterogeneity of the disease, diverse comorbidity factors and, importantly, lack of validation. The emergence of big data and big data analytics has led to a pile of studies and hypotheses. Although these approaches show great potential in a study-by-study basis, to translate these findings to a clinical setting, it is crucial to distinguish true discoveries from red herrings. Therefore, replication and validation studies in much larger cohort sizes are required. To achieve this, large and up-to-date clinical biobanks with a variety of different data types, including molecular, clinical and host characteristics, will be required to fully leverage these analytical methodologies. In precision medicine era, many national and international collaborative efforts are under way aimed at improving clinical research (figure 3).[72–84]

Advances in imaging technologies

Image recognition is one of the major applications of artificial intelligence, particularly deep learning, and holds great promise in assisting the fields of biological and medical imaging. Deep learning is a collection of algorithms in the field of machine learning with an outstanding ability to decode the contents of images. This has led to a proliferation of studies with an attempt to automate the interpretation and the evaluation of medical images, such as endoscopy, histopathology, and CT/MRI. Evaluating endoscopic inflammation, characterisation of lesions and assessment of mucosal healing is essential for proper management in IBD. However, endoscopic assessment of inflammation in IBD is highly subjective with high interobserver variability. Computer-aided scores would be much more objective for the interpretation of the endoscopic images.[85] For example, a deep learning-based model showed performance comparable to those of experienced gastroenterologists for the classification of endoscopic severity of UC into two groups: remission (Mayo 0 or 1 endoscopic score) and moderate to severe (Mayo 2 or 3 endoscopic score).[86] A novel objective computer-based score to assess UC disease activity based on endoscopic images has been developed. In particular, deep learning has been used to extract different layers of pixel data, such as measuring the redness

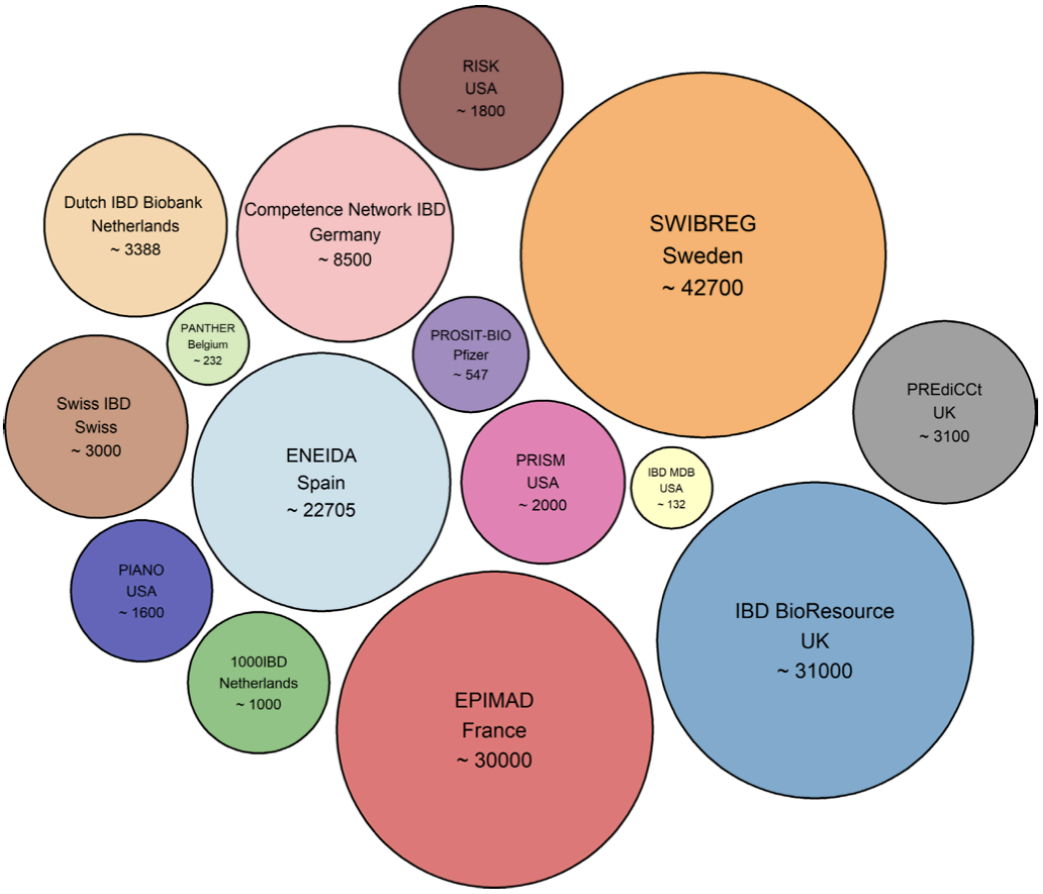


Figure 3 Academic initiatives with cohorts/biobanks in IBD. The numbers in each circle represent the approximate patient cohort size.

degree through extraction of the intensity and distribution of red pixels in the red density score in UC.[87 88] Similarly, assessment of CT/MRI images in IBD is extremely subjective; therefore, computer-aided scores could potentially overcome interobserver variation. A semiautomated image analysis software showed a performance similar to those of experienced radiologists for the assessment of CD structural bowel damage in abdominal CT-enterography data.[89] Also, machine learning methods and algorithms have been applied to predict the grading of severity of CD in abdominal MRI data.[90 91] Additionally, machine learning algorithms could assist with the time-consuming assessment of wireless capsule endoscopy data. It paves the way for automated analysis of wireless capsule endoscopy images to detect CD lesions via detection of predefined structural and textural characteristics, as well as enhancement of the underlying pixel information.[92 93]

Machine learning may also improve the analysis of histopathology and possibly tackle the unmet need of patients with unclassified IBD. Raman microspectroscopy as a cell and tissue diagnostics approach has been investigated to distinguish different IBD subtypes. Bielecki *et al* proposed that a machine learning-based workflow is capable of distinguishing tissue morphology among healthy subjects, CD and UC with great accuracy.[94] Ultimately, artificial intelligence is promising in medical imaging and will undoubtedly have a considerable impact on endoscopy practice in the future (figure 4).

Predicting prognosis

Predicting disease progression and severity is pivotal to the design of appropriate disease management strategies for individual patients. Machine learning has the potential to assist with this. Extraction of information from routinely collected electronic medical records (EMRs), such as physician's clinical observations and endoscopy reports, will allow researchers to perform prognostic research on longitudinal data. A machine learning model trained on codified information (International Classification of Diseases, Ninth Revision (ICD-9)) retrieved from EMRs, including a set of baseline laboratory parameters, patient demographics and clinical characteristics, accurately (AUC= 0.93) predicted disease severity in patients with CD.[95] Similarly, Waljee *et al* constructed a random forests machine learning model to

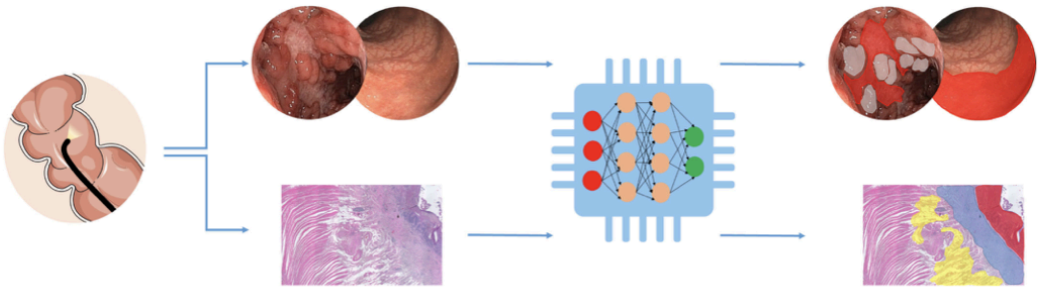


Figure 4 Artificial intelligence in medical imaging. Graphical representation of a simple deep learning-based image segmentation approach to predict boundaries of inflamed areas. The top section of the figure represents the endoscopic image of colonic CD demonstrating the 'cobblestone' appearance and ulceration. Using a simple deep learning-based image segmentation method inflamed boundaries could be predicted: cobblestone in grey and inflamed ulcer in red. The bottom section of the figure illustrates a histopathology image of inflamed stenosis from ileal CD. A deep learning-based method could be used for image segmentation and predicting boundaries of inflamed areas: acute infiltration (ulcer) in red, muscolari mucosae thickening in blue and adipocytes hyperplasia in yellow. CD, Crohn's disease.

predict IBD-related hospitalisation and outpatient steroid use, as surrogate markers of disease flares (AUC=0.87, 95% CI 0.87 to 0.88). The authors pointed out older age, high serum albumin, platelet counts, immunosuppressive medication, history of corticosteroid use and hospitalisation as risk predictors.[96] One way to improve and facilitate data extraction from plain text in medical records is by employing natural language processing (NLP), another field of artificial intelligence.[97] For example, an NLP-based model showed superior performance in comparison to an ICD-9-based model for extracting extraintestinal manifestation data from EMRs.[98] In the IBD therapeutic space, Cai *et al* applied NLP to clinical notes in identifying the risk of arthralgia in two groups of IBD patients: one treated with vedolizumab and another with TNF inhibitor.[98] Hou *et al* examined the performance of NLP-based software to classify the endoscopy procedure in patients with IBD that was performed in a diagnostic or follow-up context by mining the pathology reports.[99]

Most investigation in prognostic research has centred on investigating the diversity of underlying disease pathophysiology aiming to identify predictive correlates, which shed light onto the factors prompting disease progress, severity and clinical manifestations. Genome-wide association studies in patients with CD pinpointed the distinct genetic bases of susceptibility and prognosis and hence separate biology. These prognosis-associated SNPs are enriched for pathways involved in the regulation of innate and adaptive immune responses and responses to microorganisms. Among those, four loci have been identified to be significantly associated with prognosis in CD, namely, *FOXO3*, *XACT*, a region upstream of *IGFBP1* and *MHC*. This serves as the point of departure for better understanding of the biology that determines disease prognosis.[100] Advancement of omics techniques and data analytics have led to molecular and functional-based disease classification. For example, on combining mucosal gene expression, metagenomics and CD4 +T cell population signatures, Tang *et al* employed a machine learning approach to define a list of 26 predictors, which were effective in distinguishing between normal intestinal regions and those with active inflammation in IBD patients. Using network analysis to further interpret the inferred predictors, the authors pinpointed the role of *SAA1* in the induction of IL17 and IL22 secretion by CD4+ T cells in relation to *Bacteroides* abundance.[101]

To date, various studies have assessed the predictive value of gut microbiota. Machine learning models, especially random forests, are used extensively in microbiome research due to their ease of understanding, excellent performance and incorporated feature selection (via estimating feature importance). Douglas and colleagues studied microbial taxa and their inferred function in intestinal biopsies of 20 treatment-naive paediatric patients with CD and 20 control patients. The authors pointed to the predictive value of microbiome profiling using 16s rRNA sequencing for the disease state, whereas metagenomic-based identified markers performed best for classifying treatment response.[102]

When large integrated EMRs and multiomics datasets are combined with a powerful and robust machine learning framework, they can achieve exceptional results. Cushing *et al* identified a unique expression profile in anti-TNF-naive and anti-TNF-exposed patients with CD that could predict postoperative disease recurrence. The authors uncovered 30 influential transcripts in anti-TNF-naive patients using random forests-based machine learning models built on demographic and clinical data extracted from the EMR and transcriptomic profile of non-inflamed ileal tissue.[103]

These methodologies provide a promising initiative to the application of machine learning to predict IBD disease course and outcome, a research scope demanding comprehensive and longitudinal investigations. By expansion of data resources as well as advancement in analytic approaches, prediction of prognosis and identifying low-risk and high-risk patients doubtlessly become feasible. Future studies should aim at mining health records and integrating them with multiomics data.

Predicting drug response

In the past decades, enormous efforts have been made to predict the response to medications. Since prospective indicators of drug responses are expected to have a big impact on pharmacoeconomics, machine learning approaches have been applied to dissect the underlying complexities and predict responses to drugs used in IBD treatments. Integration of clinical and laboratory data has been used for monitoring drugs with narrow therapeutic window, such as thiopurine, to assess the risk of developing adverse events. Currently, evaluation of clinical efficacy and risk management of thiopurine is either through blood count or measuring and monitoring of the level of its metabolites 6-thioguaninenucleotide, as an indicator of response, and

6-methylmercaptopurine, which is associated with the risk of hepatotoxicity. Waljee and colleagues studied the predictive value of a set of clinical and laboratory data to differentiate clinical responders from non-responders using a machine learning model, random forests. The proposed model has an AUC of 0.85, in contrast to the conventional model with an AUC of 0.59.[104] Subsequent work has shown significant clinical benefits, including decreased steroid prescriptions, hospitalisations and surgeries.[105]

Using clinical trial data from the GEMINI I and GEMINI II studies with vedolizumab, Waljee and colleagues developed a machine learning model, random forests, incorporating demographic data, clinical data and laboratory tests to predict the likelihood of achieving week 52 corticosteroid-free endoscopic remission in patients with UC[106] and CD[107] treated with vedolizumab. Interestingly, the strongest positive prognostic markers in patients with UC were low levels of faecal calprotectin and albumin; and those in patients with CD were low levels of serum C reactive protein and albumin.

An example of efforts to generate and integrate molecular and clinical data to guide treatment relates to identifying biomarkers predictive of drug response. In an interesting study, Zarringhalam and colleagues searched for predictive biomarkers for response to infliximab for refractory UC. First, an in-house algorithm incorporating causal prior knowledge (relationships between genes defined from the literature) with gene expression data was used to define upstream gene regulators. The newly defined features were subsequently used in a machine learning model (panelised logistic regression) to predict patient's response to infliximab (accuracy=70%). The authors pinpointed interferon gamma (IFNG), lipopolysaccharide (LPS) and TNF as key regulators. They inferred that the lack of response could be due to higher expression of the TNF pathway components, enzymatic dysregulation in the IFNG pathway and activation of the LPS–TLR4 pathway triggered by the presence of Gram-negative bacteria.[108]

Given that the human gut hosts billions of microorganisms, the gut microbiome is increasingly known to be a contributor of drug efficacy.[109] Doherty and colleagues used a machine learning model using the concept of random forests to predict the therapeutic response to ustekinumab in patients with CD.[110] The model helped in the identification of microbial signatures such as altered levels of *Faecalibacterium* that were predictive of remission. Similarly, Shaw *et al* performed an analysis using a similar classifier model based on longitudinal microbiome data derived from 19 treatment-naive paediatric individuals diagnosed with IBD and exposed to biologics.[111] The authors were able to achieve a 76.5% accuracy in predicting responders based on the pretreatment microbiome. These studies suggest that stratification of patients according to their molecular and clinical characteristics would be beneficial for evaluating therapeutic efficacy. Multiomics data integration could prove useful in biomarker discovery for treatment response. Recently, our group identified 10-feature transcriptomic (accuracy of 98%) and 15-feature genomic (accuracy 96.6%) panels predicting endoscopic response to ustekinumab by incorporating genomics and transcriptomics data into a matrix factorisation-based machine learning model in patients with CD.[112]

KEY CHALLENGES AND OPPORTUNITIES

Big data and artificial intelligence represent a great step forward in precision medicine with a high reward stand-off. With the potential to simultaneously discover new therapies, make informed treatment decisions and identify disease subgroups, there is a massive effort towards making artificial intelligence commonplace in clinical and biomedical research. The increasing availability of big data, especially multiomics datasets from large IBD cohorts, development of machine learning-based algorithms and systems biology-based tools have enabled the discovery of biological knowledge relevant to IBD. However, key challenges remain especially in the realm of how such datasets become useful in clinical translation and precision medicine (figure 5). Even though existing datasets have yielded interesting biological insights, the number of cases of such datasets resulting in direct clinical benefits, has been few and far in between. This is striking especially given the fact that there is a call for personalised therapies.

This translational gap is not unsurprising since the causality axis for IBD has not yet been established. In part, this could be attributed to the temporal nature (cross sectional or longitudinal) and/or the composition (type of multiomics data types) of datasets. Longitudinal profiling of multiomics datasets even from smaller cohorts may have higher performance and information richness than larger cohorts without longitudinal profiling. This has been demonstrated in other complex diseases such as diabetes and obesity.[113 114] The cross-sectional nature of most IBD datasets tends to limit their usefulness in inferring causal mechanisms.

Missing data are also a key challenge since these leave researchers with a choice of having to leave out particular samples or imputing missing data points, which results in reduced data and unintended errors, respectively. Also unbalanced distribution of clinical or phenotypical heterogeneity is a real-world issue affecting the interpretation of any integrative analyses. There is also a dearth of omics datasets such as proteomics, which are closer to phenotypical manifestations than other data types such as genotyping or transcriptomics. The availability of already assembled large IBD cohorts with stored biomaterial throws open multiple opportunities for improving and delivering on the research front. Sampling the biomaterials for generating the missing datatypes provides new opportunities to explore complete datasets. Thus, coordination between lead researchers and funding agencies to generate coherent multilayered datasets from the same patient samples is a major requirement. Harmonised collection, storage and usage of patient metadata and medical records are also a key challenge for inferring knowledge and clinical translation.

The contribution of disease complexity to the usefulness of multiomics datasets also extends to the composition and completeness of these datasets. The specific roles of distinct cellular populations and lineages in driving and contributing to specific phenotypes are becoming increasingly clear in IBD.[115–118] Adding to the complexity is the recently discovered fact that mutations occur in a cell type-specific manner.[119] Most of the datasets from organised cohorts have either profiled expression and genotyping from bulk RNA and DNA extracted from biopsy material or whole blood respectively, making it difficult to investigate the role of specific cell types in the aetiology and pathogenesis of IBD. As a case in point, Smillie *et al* demonstrated the power of profiling the expression of more than 50 cell types to pinpoint intercellular circuits which distinguish UC and healthy states.[120]

The implementation of big data and artificial intelligence approaches into clinical practice and meaningful benefits for patients is the ultimate challenge. On one hand, the deployment and operationalisation of big data are challenging, which are being addressed using computational sciences and algorithmic

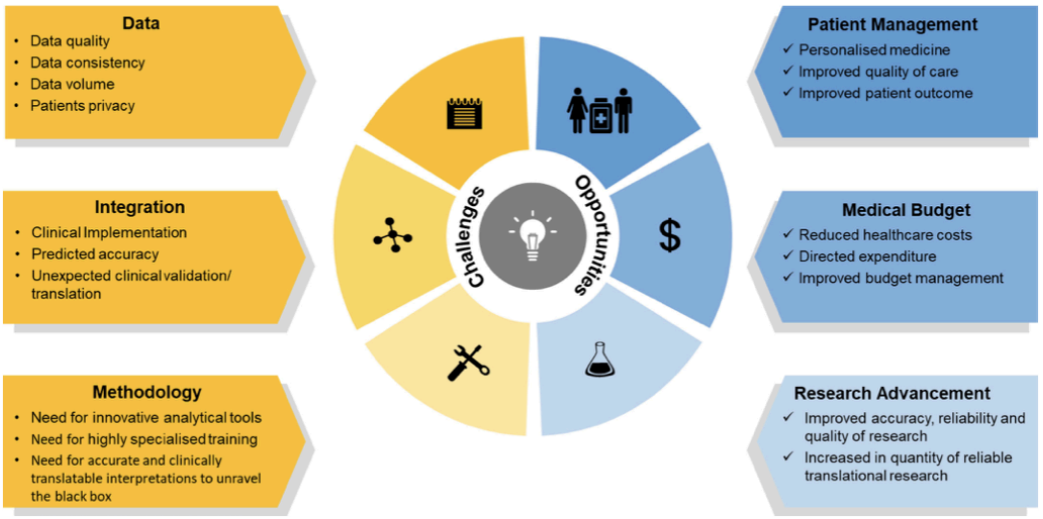


Figure 5 Opportunities and challenges in the use of machine learning and data integration to achieve improved and personalised healthcare in IBD. While challenges exist in generating good quality data in a standardised manner and at a volume deemed suitable for ensuring baseline performance of machine learning models, there remain difficulties in terms of the expertise needed to identify and employ appropriate tools for data integration and interpretation. However, with emerging advances in the data integration field, the incentives and opportunities to advance precision medicine with clinical implications are expected to drive integrative IBD research forward.

frameworks to manage problems related to storage, analysis, integration and interpretation of big data. Most of the infrastructures are being explored and adopted from the computer science field into healthcare. These include cloud-based data storage and analysis, and massively parallel processing hardware to tackle the rapid increase in the volumes of data from EMR, imaging and omics measurements, for example. Moreover, there is a need for user-friendly software and workflows to facilitate the integration of big data analytics into clinical practice. For instance, there have been efforts into developing NLP-based software to assist medical investigators with extracting data from plain text, such as clinical reports.[121 122]

On the other hand, many clinicians are cautious of artificial intelligence approaches mainly because most of these approaches are essentially black boxes and do not link predictions to underlying mechanisms, nor provide functional explanations for the discovered associations, correlations and recommended decisions. However, causal mechanistic insights are key for clinical applicability so as to enhance reliability and thereby patient safety, especially in a complex heterogeneous disease such as IBD. Furthermore, as poorly validated models could do more harm than good, in depth experimental and clinical validation is crucial for machine learning-based models before implementation in clinical setting. From the analytics point, interpretable machine learning models should be developed.[123] Besides, there is a need to benchmark performance indices and parameters to evaluate the performance of machine learning techniques.[124] Other challenges include the uncertainties associated with analyses involving the use of biological networks despite the functional context provided by the networks. Even though high-quality manually curated and benchmarked networks exist,[125 126] analytical methods which take into account the uncertainties of individual interactions and their contextuality need to be developed. Clinical validation is fundamental for the implementation of artificial intelligence-based approaches. In one of the first randomised clinical trials using artificial intelligence, Lin *et al* compared the efficacies of childhood cataracts diagnosed by senior ophthalmologists with those from CC-Cruiser, a previously developed artificial intelligence platform for risk stratification and treatment guidance. This trial showed that regardless of the inferior accuracy of CC-Cruiser compared with senior ophthalmologists, artificial intelligence had the capacity to assist doctors in decision-making.[127 128] All in all, clinicians are right to be sceptical of the implementation of these otherwise inexplicable approaches in clinical practice, and although there have been considerable advances in the implementation of big data, there still remain many technological, translational and cultural barriers for the assimilation of artificial intelligence approaches into clinical practice.

CONCLUSION

By enabling data integration and assisting the discovery of non-trivial patterns and translatable knowledge in the integrated datasets, machine learning and systems biology offer unique opportunities to study and investigate the aetiology of complex diseases such as IBD. Machine learning guided IBD research has great potential to accelerate the formulation of cutting-edge precision medicine applications with clinical relevance and utility. However, for the promise of machine learning to come to translational fruition, there remain many stumbling blocks. However, almost all of the challenges also come with a huge potential for discovering knowledge and translating it to IBD clinical practice. It is expected that, with the availability of large IBD initiatives such as national biobanks with stored biomaterial, datasets can be made more coherent and complete, thus filling

the biological gap for systems biology and the statistical gap for machine learning to produce knowledge which is closer to clinical practice and translation.

SEARCH STRATEGY

Articles were retrieved from PubMed after employing the following search criteria. Two key-word groups were created, with the first one comprising "Inflammatory Bowel Disease", "Crohn's disease" and "ulcerative colitis" and second one comprising "machine learning", "Artificial Intelligence", "deep learning", "-omics", "big data", "systems biology", "network biology", "genomics", "transcriptomics", "GWAS", "proteomics" and "microbiome". Pairwise combination of keywords from the two groups was used to search for articles published until July 2019. Only articles written in English were included.

Contributors NSST performed the literature search, wrote the individual sections, compiled the visual objects and was involved with the overall manuscript generation. PS wrote an individual section of the manuscript and made a critical review of the manuscript. MM performed the literature survey and critical review of the manuscript. TK performed a critical review of the manuscript. BV performed a critical review of the manuscript. SV formulated the idea and read and edited the manuscript. All authors discussed and revised the draft and approved the final version of the manuscript.

Funding This work was supported by European Research Council (ERC). NSST and PS were supported by the ERC Advanced Grant (ERC-2015-AdG, 694679, CrUCCial). BV is a doctoral fellow and SV is a senior clinical investigator of the Research Foundation Flanders (FWO), Belgium. MM is supported by Biotechnological and Biosciences Research Council (BBSRC) Norwich Research Park Biosciences Doctoral Training Partnership (grant number BB/S50743X/1), as an NPIF Award. TK was supported by a fellowship in computational biology at the Earlham Institute (Norwich, UK) in partnership with the Quadram Institute (Norwich, UK) and was strategically supported by the BBSRC (BB/J004529/1, BB/P016774/1 and BB/CSP17270/1).

Competing interests BV received lecture fees from Abbvie, Ferring Pharmaceuticals, Janssen, R-Biopharm and Takeda; consultancy fees from Janssen and Sandoz. SV: research grant: MSD, AbbVie, Takeda, Pfizer, J&J; lecture fee: MSD, AbbVie, Takeda, Ferring, Centocor, Hospira, Pfizer, J&J, Genentech/Roche; consultancy: MSD, AbbVie, Takeda, Ferring, Centocor, Hospira, Pfizer, J&J, Genentech/Roche, Celgene, Mundipharma, Celltrion, SecondGenome, Prometheus, Shire, Prodigest, Gilead, Galapagos. SV is a senior clinical investigator of the Research Foundation–Flanders (FWO). The work of MM and TK is supported by BenevolentAI, and TK's work is also supported by Unilever.

Patient consent for publication Not required.

Provenance and peer review Commissioned; externally peer reviewed.

ORCID iDs
Nasim Sadat Seyed Tabib http://orcid.org/0000-0001-9612-0012
Bram Verstockt http://orcid.org/0000-0003-3898-7093

REFERENCES

1 Korcsmaros T, Schneider MV, Superti-Furga G. Next generation of network medicine: interdisciplinary signaling approaches. *Integr Biol* 2017;9:97–108.
2 Weersma RK, Xavier RJ, Vermeire S, *et al*. Multiomics analyses to deliver the most effective treatment to every patient with inflammatory bowel disease. *Gastroenterology* 2018;155:e1–4.
3 Ananthakrishnan AN. Epidemiology and risk factors for IBD. *Nat Rev Gastroenterol Hepatol* 2015;12:205–17.
4 de Souza HSP, Fiocchi C. Immunopathogenesis of IBD: current state of the art. *Nat Rev Gastroenterol Hepatol* 2016;13:13–27.
5 Ananthakrishnan AN, Bernstein CN, Iliopoulos D, *et al*. Environmental triggers in IBD: a review of progress and evidence. *Nat Rev Gastroenterol Hepatol* 2018;15:39–49.
6 Gecse KB, Vermeire S. Differential diagnosis of inflammatory bowel disease: imitations and complications. *Lancet Gastroenterol Hepatol* 2018;3:644–53.
7 Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst* 2014;2:3.
8 Brooks J, Watson A, Korcsmaros T. Omics approaches to identify potential biomarkers of inflammatory diseases in the focal adhesion complex. *Genomics Proteomics Bioinformatics* 2017;15:101–9.
9 Gligorijević V, Pržulj N. Methods for biological data integration: perspectives and challenges. *J R Soc Interface* 2015;12:20150571.
10 Ideker T, Galitski T, Hood L. A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet* 2001;2:343–72.
11 Camacho DM, Collins KM, Powers RK, *et al*. Next-Generation machine learning for biological networks. *Cell* 2018;173:1581–92.
12 Hood L, Tian Q. Systems approaches to biology and disease enable translational systems medicine. *Genomics Proteomics Bioinformatics* 2012;10:181–5.
13 Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol* 2019;20:e262–73.
14 Brooks J, Modos D, Sudhakar P, *et al*. A systems genomics approach to uncover patient-specific pathogenic pathways and proteins in a complex disease. *bioRxiv* 2019;692269.
15 Sorzano COS, Vargas J, Pascual-Montano AD. A survey of dimensionality reduction techniques. *ArXiv* 2014:abs/1403.2.
16 Tedjo DI, Smolinska A, Savelkoul PH, *et al*. The fecal microbiota as a biomarker for disease activity in Crohn's disease. *Sci Rep* 2016;6:35216.
17 Mirkov MU, Verstockt B, Cleynen I. Genetics of inflammatory bowel disease: beyond NOD2. *Lancet Gastroenterol Hepatol* 2017;2:224–34.
18 Ogura Y, Bonen DK, Inohara N, *et al*. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* 2001;411:603–6.
19 Shah SC, Kusiak A. Data mining and genetic algorithm based gene/SNP selection. *Artif Intell Med* 2004;31:183–96.
20 Phuong TM, Lin Z, Altman RB. Choosing SNPs using feature selection. In: 2005 IEEE Computational Systems Bioinformatics Conference (CSB'05). *IEEE* 2005:301–9.
21 Long N, Gianola D, Rosa GJM, *et al*. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. *J Anim Breed Genet* 2007;124:377–89.
22 Bermingham ML, Pong-Wong R, Spiliopoulou A, *et al*. Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Sci Rep* 2015;5:10312.
23 Wei Z, Wang W, Bradfield J, *et al*. Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. *Am J Hum Genet* 2013;92:1008–12.
24 Luo Y, de Lange KM, Jostins L, *et al*. Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7. *Nat Genet* 2017;49:186–92.
25 Romagnoni A, Jégou S, Van Steen K, *et al*. Comparative performances of machine learning methods for classifying Crohn disease patients using genome-wide genotyping data. *Sci Rep* 2019;9:10351.
26 Schwerd T, Bryant RV, Pandey S, *et al*. Nox1 loss-of-function genetic variants in patients with inflammatory bowel disease. *Mucosal Immunol* 2018;11:562–74.
27 Ghandi M, Lee D, Mohammad-Noori M, *et al*. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput Biol* 2014;10:e1003711.
28 Alipanahi B, Delong A, Weirauch MT, *et al*. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* 2015;33:831–8.
29 Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat Methods* 2015;12:931–4.
30 Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res* 2016;26:990–9.
31 Zou J, Huss M, Abid A, *et al*. A primer on deep learning in genomics. *Nat Genet* 2019;51:12–18.
32 Li XL, Zhou CY, Sun Y, *et al*. Bioinformatic analysis of potential candidates for therapy of inflammatory bowel disease. *Eur Rev Med Pharmacol Sci* 2015;19:4275–84.
33 Hong SN, Joung J-G, Bae JS, *et al*. Rna-Seq reveals transcriptomic differences in inflamed and Noninflamed intestinal mucosa of Crohn"s disease patients compared with normal mucosa of healthy controls. *Inflamm Bowel Dis* 2017;23:1098–108.
34 Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 2005;4:Article17.
35 Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;9:559.
36 Lin X, Li J, Zhao Q, *et al*. WGCNA reveals key roles of IL8 and MMP-9 in progression of involvement area in colon of patients with ulcerative colitis. *Curr Med Sci* 2018;38:252–8.
37 Verstockt S, De Hertogh G, Van der Goten J, *et al*. Gene and Mirna Regulatory Networks During Different Stages of Crohn's Disease. *J Crohn's Colitis* 2019;13:916–30.
38 Verstockt B, Verstockt S, Creyns B, *et al*. Mucosal IL13RA2 expression predicts nonresponse to anti-TNF therapy in Crohn's disease. *Aliment Pharmacol Ther* 2019;49:572–81.
39 Thiele I, Swainston N, Fleming RMT, *et al*. A community-driven global reconstruction of human metabolism. *Nat Biotechnol* 2013;31:419–25.
40 Knecht C, Fretter C, Rosenstiel P, *et al*. Distinct metabolic network states manifest in the gene expression profiles of pediatric inflammatory bowel disease patients and controls. *Sci Rep* 2016;6:32584.

41 Yuan F, Zhang Y-H, Kong X-Y, *et al*. Identification of candidate genes related to inflammatory bowel disease using minimum redundancy maximum relevance, incremental feature selection, and the shortest-path approach. *Biomed Res Int* 2017;2017:1–15.
42 Isakov O, Dotan I, Ben-Shachar S. Machine Learning-Based gene prioritization identifies novel candidate risk genes for inflammatory bowel disease. *Inflamm Bowel Dis* 2017;23:1516–23.
43 Manichanh C, Borruel N, Casellas F, *et al*. The gut microbiota in IBD. *Nat Rev Gastroenterol Hepatol* 2012;9:599–608.
44 Halfvarson J, Brislawn CJ, Lamendella R, *et al*. Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat Microbiol* 2017;2:17004.
45 Norman JM, Handley SA, Baldridge MT, *et al*. Disease-Specific alterations in the enteric Virome in inflammatory bowel disease. *Cell* 2015;160:447–60.
46 Zuo T, Lu X-J, Zhang Y, *et al*. Gut mucosal virome alterations in ulcerative colitis. *Gut* 2019;68:1169–79.
47 Greenblum S, Turnbaugh PJ, Borenstein E. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc Natl Acad Sci U S A* 2012;109:594–9.
48 Li X, LeBlanc J, Elashoff D, *et al*. Microgeographic proteomic networks of the human colonic mucosa and their association with inflammatory bowel disease. *Cell Mol Gastroenterol Hepatol* 2016;2:567–83.
49 Erickson AR, Cantarel BL, Lamendella R, *et al*. Integrated Metagenomics/Metaproteomics reveals human Host-Microbiota signatures of Crohn's disease. *PLoS One* 2012;7:e49138.
50 Häsler R, Sheibani-Tezerji R, Sinha A, *et al*. Uncoupling of mucosal gene regulation, mRNA splicing and adherent microbiota signatures in inflammatory bowel disease. *Gut* 2017;66:2087–97.
51 Lloyd-Price J, Arze C, Ananthakrishnan AN, *et al*. Multi-Omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 2019;569:655–62.
52 Verstockt B, Ferrante M, Vermeire S, *et al*. New treatment options for inflammatory bowel diseases. *J Gastroenterol* 2018;53:585–90.
53 Maaser C, Sturm A, Vavricka SR, *et al*. ECCO-ESGAR guideline for diagnostic assessment in IBD Part 1: initial diagnosis, monitoring of known IBD, detection of complications. *J Crohn's Colitis* 2019;13:144–64.
54 Sturm A, Maaser C, Calabrese E, *et al*. ECCO-ESGAR guideline for diagnostic assessment in IBD Part 2: IBD scores and general principles and technical aspects. *J Crohn's Colitis* 2019;13:273–84.
55 Colombel J-F, Panaccione R, Bossuyt P, *et al*. Effect of tight control management on Crohn's disease (calm): a multicentre, randomised, controlled phase 3 trial. *Lancet* 2017;390:2779–89.
56 Bremer V, Becker D, Kolovos S, *et al*. Predicting therapy success and costs for personalized treatment recommendations using baseline characteristics: data-driven analysis. *J Med Internet Res* 2018;20:e10275.
57 Wang H-Y, Hung C-C, Chen C-H, *et al*. Increase Trichomonas vaginalis detection based on urine routine analysis through a machine learning approach. *Sci Rep* 2019;9.
58 Cleynen I, Boucher G, Jostins L, *et al*. Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: a genetic association study. *Lancet* 2016;387:156–67.
59 Jeong C-S, Kim D. Inferring Crohn's disease association from exome sequences by integrating biological knowledge. *BMC Med Genomics* 2016;9:35.
60 Medina I, Montaner D, Tarraga J, *et al*. Prophet, a web-based tool for class prediction using microarray data. *Bioinformatics* 2007;23:390–1.
61 Montero-Meléndez T, Llor X, García-Planella E, *et al*. Identification of novel predictor classifiers for inflammatory bowel disease by gene expression profiling. *PLoS One* 2013;8:e76235.
62 Han L, Maciejewski M, Brockel C, *et al*. A probabilistic pathway score (PROPS) for classification with applications to inflammatory bowel disease. *Bioinformatics* 2018;34:985–93.
63 Seeley EH, Washington MK, Caprioli RM, *et al*. Proteomic patterns of colonic mucosal tissues delineate Crohn's colitis and ulcerative colitis. *Proteomics Clin Appl* 2013;7:541–9.
64 Cao B, Zhou X, Ma J, *et al*. Role of miRNAs in inflammatory bowel disease. *Dig Dis Sci* 2017;62:1426–38.
65 Duttagupta R, DiRienzo S, Jiang R, *et al*. Genome-Wide maps of circulating miRNA biomarkers for ulcerative colitis. *PLoS One* 2012;7:e31241.
66 Hübenthal M, Hemmrich-Stanisak G, Degenhardt F, *et al*. Sparse modeling reveals miRNA signatures for diagnostics of inflammatory bowel disease. *PLoS One* 2015;10:e0140155.
67 Plevy S, Silverberg MS, Lockton S, *et al*. Combined serological, genetic, and inflammatory markers differentiate Non-IBD, Crohn's disease, and ulcerative colitis patients. *Inflamm Bowel Dis* 2013;19:1139–48.
68 Morgan XC, Tickle TL, Sokol H, *et al*. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol* 2012;13:R79.
69 Pasolli E, Truong DT, Malik F, *et al*. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput Biol* 2016;12:e1004977.
70 Fioravanti D, Giarratano Y, Maggio V, *et al*. Phylogenetic convolutional neural networks in metagenomics. *BMC Bioinformatics* 2018;19:49.
71 Gevers D, Kugathasan S, Denson LA, *et al*. The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* 2014;15:382–92.
72 | Pfizer for Professionals. PROSIT-BIO. Available: https://www.pfizerpro.co.uk/product/inflectra/ulcerative-colitis/support/prosit-bio-0 [Accessed 23 Aug 2019].
73 1000 IBD. Available: https://1000ibd.org/ [Accessed 23 Aug 2019].
74 Spekhorst LM, Imhann F, Festen EAM, *et al*. Cohort profile: design and first results of the Dutch IBD Biobank: a prospective, nationwide Biobank of patients with inflammatory bowel disease. *BMJ Open* 2017;7:e016695.
75 Chaparro M, Ramas M, Benítez JM, *et al*. Extracolonic cancer in inflammatory bowel disease: data from the GETECCU Eneida registry. *Am J Gastroenterol* 2017;112:1135–43.
76 Beaulieu DB, Ananthakrishnan AN, Martin C, *et al*. Use of biologic therapy by pregnant women with inflammatory bowel disease does not affect infant response to vaccines. *Clin Gastroenterol Hepatol* 2018;16:99–105.
77 IBDMDB. Home IBDMDB. Available: https://ibdmdb.org/ [Accessed 23 Aug 2019].
78 PREdiCCt. Home. Available: https://www.predicct.co.uk/ [Accessed 23 Aug 2019].
79 CSIBD PRISM registry Hospital, Boston, MA. Available: https://www.massgeneral.org/csibd/cores/clinical.aspx [Accessed 9 Jan 2020].
80 IBD BioResource. Translating today's science into tomorrow's treatments. Available: https://www.ibdbioresource.nihr.ac.uk/ [Accessed 9 Jan 2020].
81 Home - SWISS IBDcohort. Available: http://www.ibdcohort.ch/ [Accessed 9 Jan 2020].
82 Swibreg. Patient. Available: http://www.swibreg.se/ [Accessed 9 Jan 2020].
83 EPIMAD : le plus grand registre au monde – Observatoire National des MICI. Available: http://www.observatoire-crohn-rch.fr/epimad-le-plus-grand-registre-de-malades-au-monde/ [Accessed 9 Jan 2020].
84 Study management - Competence Network for Bowel Diseases. Available: http://www.kompetenznetz-darmerkrankungen.de/Studienmanagement [Accessed 9 Jan 2020].
85 Bossuyt P, Vermeire S, Bisschops R. Scoring endoscopic disease activity in IBD: artificial intelligence sees more and better than we do. *Gut* 2020;69:788–9.
86 Stidham RW, Liu W, Bishu S, *et al*. Performance of a deep learning model vs human reviewers in grading endoscopic disease severity of patients with ulcerative colitis. *JAMA Netw Open* 2019;2:e193963.
87 Bossuyt P, Nakase H, Vermeire S, *et al*. 436 - Automated Digital Calculation of Endoscopic Inflammation in Ulcerative Colitis: Results of the Red Density Study. *Gastroenterology* 2018;154:S98–9.
88 Bossuyt P, Nakase H, Vermeire S, *et al*. Automatic, computer-aided determination of endoscopic and histological inflammation in patients with mild to moderate ulcerative colitis based on red density. *Gut* 2020. doi: 10.1136/gutjnl-2019-320056. [Epub ahead of print: 8 Jan 2020]. doi:10.1136/gutjnl-2019-320056. [Epub ahead of print:
89 Stidham RW, Enchakalody B, Waljee AK, *et al*. Assessing Small Bowel Stricturing and Morphology in Crohn's Disease Using Semi-automated Image Analysis. *Inflamm Bowel Dis*;11.
90 Tielbeek JAW, Vos FM, Stoker J. A computer-assisted model for detection of MRI signs of Crohn's disease activity: future or fiction? *Abdom Imaging* 2012;37:967–73.
91 Mahapatra D, Schüffler PJ, Tielbeek JAW, *et al*. Semi-supervised and active learning for automatic segmentation of Crohn's disease. *Med Image Comput Comput Assist Interv* 2013;16:214–21.
92 Kumar R, Qian Zhao Q, Seshamani S, *et al*. Assessment of Crohn's Disease Lesions in Wireless Capsule Endoscopy Images. *IEEE Trans Biomed Eng* 2012;59:355–62.
93 Charisis VS, Hadjileontiadis LJ. Potential of hybrid adaptive filtering in inflammatory lesion detection from capsule endoscopy images. *WJG* 2016;22:8641.
94 Bielecki C, Bocklitz TW, Schmitt M, *et al*. Classification of inflammatory bowel diseases by means of Raman spectroscopic imaging of epithelium cells. *J Biomed Opt* 2012;17:0760301.
95 Reddy BK, Delen D, Agrawal RK. Predicting and explaining inflammation in Crohn's disease patients using predictive analytics methods and electronic medical record data. *Health Informatics J* 2019;25:1201–18.
96 Waljee AK, Lipson R, Wiitala WL, *et al*. Predicting hospitalization and outpatient corticosteroid use in inflammatory bowel disease patients using machine learning. *Inflamm Bowel Dis* 2018;24:45–53.
97 Ananthakrishnan AN, Cai T, Savova G, *et al*. Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing. *Inflamm Bowel Dis* 2013;19:1411–20.
98 Cai T, Lin T-C, Bond A, *et al*. The association between arthralgia and vedolizumab using natural language processing. *Inflamm Bowel Dis* 2018;24:2242–6.
99 Hou JK, Chang M, Nguyen T, *et al*. Automated identification of surveillance colonoscopy in inflammatory bowel disease using natural language processing. *Dig Dis Sci* 2013;58:936–41.
100 Lee JC, Biasci D, Roberts R, *et al*. Genome-Wide association study identifies distinct genetic contributions to prognosis and susceptibility in Crohn's disease. *Nat Genet* 2017;49:262–8.
101 Tang MS, Bowcutt R, Leung JM, *et al*. Integrated analysis of biopsies from inflammatory bowel disease patients identifies SAA1 as a link between mucosal microbes with Th17 and Th22 cells. *Inflamm Bowel Dis* 2017;23:1544–54.

102 Douglas GM, Hansen R, Jones CMA, *et al*. Multi-omics differentially classify disease state and treatment outcome in pediatric Crohn's disease. *Microbiome* 2018;6:13.
103 Cushing KC, Mclean R, McDonald KG, *et al*. Predicting Risk of Postoperative Disease Recurrence in Crohn's Disease: Patients With Indolent Crohn's Disease Have Distinct Whole Transcriptome Profiles at the Time of First Surgery. *Inflamm Bowel Dis* 2019;25:180–93.
104 Waljee AK, Joyce JC, Wang S, *et al*. Algorithms Outperform metabolite tests in predicting response of patients with inflammatory bowel disease to thiopurines. *Clin Gastroenterol Hepatol* 2010;8:143–50.
105 Waljee AK, Sauder K, Patel A, *et al*. Machine learning algorithms for objective remission and clinical outcomes with thiopurines. *J Crohns Colitis* 2017;11:801–10.
106 Waljee AK, Liu B, Sauder K, *et al*. Predicting corticosteroid-free endoscopic remission with vedolizumab in ulcerative colitis. *Aliment Pharmacol Ther* 2018;47:763–72.
107 Waljee AK, Liu B, Sauder K, *et al*. Predicting Corticosteroid-Free Biologic Remission with Vedolizumab in Crohn's Disease. *Inflamm Bowel Dis* 2018;24:1185–92.
108 Zarringhalam K, Enayetallah A, Reddy P, *et al*. Robust clinical outcome prediction based on Bayesian analysis of transcriptional profiles and prior causal networks. *Bioinformatics* 2014;30:i69–77.
109 Vázquez-Baeza Y, Callewaert C, Debelius J, *et al*. Impacts of the human gut microbiome on therapeutics. *Annu Rev Pharmacol Toxicol* 2018;58:253–70.
110 Doherty MK, Ding T, Koumpouras C, *et al*. Fecal Microbiota Signatures Are Associated with Response to Ustekinumab Therapy among Crohn's Disease Patients. *MBio* 2018;9:e02120–17.
111 Shaw KA, Bertha M, Hofmekler T, *et al*. Dysbiosis, inflammation, and response to treatment: a longitudinal study of pediatric subjects with newly diagnosed inflammatory bowel disease. *Genome Med* 2016;8:75.
112 Verstockt B, Sudahakar P, Creyns B, *et al*. DOP70 An integrated multi-omics biomarker predicting endoscopic response in ustekinumab treated patients with Crohn's disease. *J Crohn's Colitis* 2019.
113 Piening BD, Zhou W, Contrepois K, *et al*. Integrative personal omics profiles during periods of weight gain and loss. *Cell Syst* 2018;6:157–70.
114 Zhou W, Sailani MR, Contrepois K, *et al*. Longitudinal multi-omics of host-microbe dynamics in prediabetes. *Nature* 2019;569:663–71.
115 Allez M, Tieng V, Nakazawa A, *et al*. CD4+NKG2D+ T cells in Crohn's disease mediate inflammatory and cytotoxic responses through MICA interactions. *Gastroenterology* 2007;132:2346–58.
116 Lee JC, Lyons PA, McKinney EF, *et al*. Gene expression profiling of CD8+ T cells predicts prognosis in patients with Crohn disease and ulcerative colitis. *J Clin Invest* 2011;121:4170–9.
117 Imam T, Park S, Kaplan MH, *et al*. Effector T helper cell subsets in inflammatory bowel diseases. *Front Immunol* 2018;9:1212.
118 Chapuy L, Bsat M, Rubio M, *et al*. IL-12 and mucosal CD14+ monocyte-like cells induce IL-8 in colonic memory CD4+ T cells of patients with Ulcerative colitis but not Crohn's disease. *J Crohn's Colitis*.
119 Yizhak K, Aguet F, Kim J, *et al*. RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science* 2019;364:eaaw0726.
120 Smillie CS, Biton M, Ordovas-Montanes J, *et al*. Intra- and Inter-cellular rewiring of the human colon during ulcerative colitis. *Cell* 2019;178:714–30.
121 D'Avolio LW, Nguyen TM, Farwell WR, *et al*. Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC). *J Am Med Inform Assoc* 2010;17:375–82.
122 Savova GK, Masanz JJ, Ogren PV, *et al*. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;17:507–13.
123 Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 2019;1:206–15.
124 Hand DJ. Classifier technology and the illusion of progress. *Stat Sci* 2006;21:1–14.
125 Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat Methods* 2016;13:966–7.
126 Huang JK, Carlin DE, Yu MK, *et al*. Systematic evaluation of molecular networks for discovery of disease genes. *Cell Syst* 2018;6:484–95.
127 Long E, Lin H, Liu Z, *et al*. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. *Nat Biomed Eng* 2017;1.
128 Lin H, Li R, Liu Z, *et al*. Diagnostic efficacy and therapeutic decision-making capacity of an artificial intelligence platform for childhood cataracts in eye clinics: a multicentre randomized controlled trial. *EClinicalMedicine* 2019;9:52–9.