
















Host-microbe multi-omics and succinotype profiling have prognostic value for future relapse in patients with inflammatory bowel disease

Jill O'Sullivan ^{a,b,c}, Shriram Patel ^{a,b,d}, Gabriel E. Leventhal ^e, Rachel S. Fitzgerald ^{a,b}, Emilio J. Laserna-Mendieta ^{a,b}, Chloe E. Huseyin ^{a,b}, Nina Konstantinidou ^{a,f}, Erica Rutherford ^f, Aonghus Lavelle ^{b,g}, Karim Dabbagh ^f, Todd Z. DeSantis ^f, Fergus Shanahan ^{b,h}, Andriy Temko ⁱ, Shoko Iwai ^f, and Marcus J. Claesson ^{a,b}

^aSchool of Microbiology, University College Cork, Cork, Ireland; ^bAPC Microbiome Ireland, University College Cork, Cork, Ireland; ^cSFI Centre for Research Training in Genomics Data Science, University of Galway, Galway, Ireland; ^dSeqBiome Ltd, Cork, Ireland; ^ePharmaBiome AG, Schlieren, Zurich, Switzerland; ^fDepartment of Informatics, Second Genome Inc, South San Francisco, California, USA; ^gDepartment of Anatomy and Neuroscience, University College Cork, Cork, County Cork, Ireland; ^hDepartment of Medicine, University College Cork, Cork, Ireland; ⁱDepartment of Electrical and Electronic Engineering, University College Cork, Cork, Ireland

ABSTRACT

Crohn's disease (CD) and ulcerative colitis (UC) are chronic relapsing inflammatory bowel disorders (IBD), the pathogenesis of which is uncertain but includes genetic susceptibility factors, immune-mediated tissue injury and environmental influences, most of which appear to act via the gut microbiome. We hypothesized that host-microbe alterations could be used to prognostically stratify patients experiencing relapses up to four years after endoscopy. We therefore examined multiple omics data, including published and new datasets, generated from paired inflamed and non-inflamed mucosal biopsies from 142 patients with IBD (54 CD; 88 UC) and from 34 control (non-diseased) biopsies. The relapse-predictive potential of 16S rRNA gene and transcript amplicons (standing and active microbiota) were investigated along with host transcriptomics, epigenomics and genetics. While standard single-omics analysis could not distinguish between patients who relapsed and those that remained in remission within four years of colonoscopy, we did find an association between the number of flares and a patient's succinotype. Our multi-omics machine learning approach was also able to predict relapse when combining features from the microbiome and human host. Therefore multi-omics, rather than single omics, better predicts relapse within 4 years of colonoscopy, while a patient's succinotype is associated with a higher frequency of relapses.

ARTICLE HISTORY

Received 16 July 2024
Revised 7 November 2024
Accepted 2 January 2025

KEYWORDS

Crohn's disease; ulcerative colitis; inflammatory bowel disease; gut microbiome; host-microbe interactions; machine learning


Introduction

The microbiome has been implicated in the pathogenesis of Crohn's disease and ulcerative colitis, collectively described as inflammatory bowel disease (IBD). Although alterations to the gut microbiome composition have been reported prior to the onset of clinically overt disease in subjects at increased risk of developing IBD by us and others,^{1–7} many of the microbiome compositional anomalies are linked with and may be secondary to the presence of inflammation. Thus, we previously found disturbances in the microbiome in a cross-sectional study of biopsies from inflamed and non-inflamed segments of the bowel in both forms of IBD.^{5,8} Moreover, in a separate longitudinal study, we showed that fecal microbiome disturbances

were associated with active disease rather than remission.⁴

While an undisputed link exists between the microbiome and IBD, this is just one of multiple factors involved in the disease.⁹ Therefore, multi-omics analysis of both host and microbiome data offers novel opportunities to further our understanding of these chronic disorders. We previously showed that by combining microbiota data with host features, it was possible to improve the classification of disease and inflammation status of samples from patients with IBD.⁵ Priya and coauthors also found associations between microbiota and host gene expression profiles, several of which were specific to IBD when compared to other gastrointestinal disorders,¹⁰ while another study

CONTACT Marcus J. Claesson  m.claesson@ucc.ie  School of Microbiology, University College Cork, Cork, Ireland

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/19490976.2025.2450207>

© 2025 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

identified context-specific mucosal host-microbe interactions within patients with IBD.¹¹

While the number of IBD studies using multi-omics data is increasing, few have applied these datasets to predict future disease outcomes in patients. Of those that have, the focus has been on multi-omics data from either the host¹² or the microbiome,¹³ rather than the integration of both. Here, we assess disease outcome four years after initial microbiome sampling⁵ to determine if a multi-omics profile consisting of both microbiome and host data might be of value in prognostically stratifying patients. The results suggest that a multi-omics strategy (rather than single omics) is more predictive of relapse four years after colonoscopy while a patient's succinotype is associated with a higher frequency of relapses.

Materials and methods

The subjects included in this study were recruited as described previously.⁵ Briefly, these subjects were all undergoing colonoscopy or sigmoidoscopy as part of their ongoing care and volunteered to provide biopsy material for research. For patients with CD, colonic biopsies were collected from inflamed and non-inflamed regions. In the case of UC subjects, biopsies were taken from the distal inflamed and proximal non-inflamed segment of the colon. Those controls included in the study consisted of subjects undergoing colonoscopy for cancer or other disease screening in which no significant colonic or gastrointestinal disorder was found.

Data generation

Nucleic acid extraction and sequence data generation

Biopsies were completely defrosted in RNA-later before performing DNA/RNA purification with AllPrep DNA/RNA/Protein Mini kit (Qiagen). Defrosted biopsies were transferred into a tube containing 350 μ L RLT buffer with β -mercaptoethanol (Sigma-Aldrich, St Louis, MO), three 3.5 mm glass beads and 0.25 mL of 0.1 mm glass beads (Biospec, Bartlesville, OK). Disruption and homogenization were carried out in a MagNA

Lyser (Roche, Penzberg, Germany) twice for 15 seconds at 3,500 or 6,500 rpm. PERMANOVA test confirmed the different centrifugation speed did not significantly affect microbiota (data not shown). Subsequent DNA/RNA purification was performed according to the kit manufacturer's instructions. DNA contaminations in RNA samples were removed by Turbo DNA-free kit following manufacturer's instructions (Ambion, Carlsbad, CA). DNA and RNA concentrations were measured using a Nano-Drop 2000 Spectrophotometer (Thermo Scientific, Waltham, MA). DNA and RNA integrity were checked on 1% agarose gel electrophoresis and 2100 Bioanalyzer system (Agilent Technologies, Santa Clara, CA), respectively. In addition, RNA quality was considered acceptable if RNA integrity number ≥ 6 and rRNA ratio ≥ 1.5 . Nucleic acid extracts were stored at -80°C until further downstream applications. For 16S cDNA analysis, total RNA was reverse transcribed using High Capacity cDNA Reverse Transcription kit following manufacturer's instructions (Applied Biosystems, Foster City, CA). The PCR was employed to amplify 16S rRNA V3-V4 hypervariable region using 341F and 805 R primer set with Nextera transposase adaptors¹⁴: 16S_V3_341F, TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG; 16S_V4_805R, GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC. Template DNA or cDNA was mixed with primers at a concentration of 0.2 μ M and Phusion High-Fidelity DNA Polymerase for a total volume of 30 μ L (Thermo Scientific). PCR conditions were 98 $^{\circ}\text{C}$ for 30 sec, 30 cycles of 98 $^{\circ}\text{C}$ for 10 sec, 55 $^{\circ}\text{C}$ for 15 sec and 72 $^{\circ}\text{C}$ for 20 sec, with final elongation at 72 $^{\circ}\text{C}$ for 5 minutes. PCR products were verified with a presence of a band on an agarose gel and purified using Agencourt AMPure XP magnetic beads (Beckman-Coulter, Brea, CA). Purified DNA product was eluted in 50 μ L EB buffer (Qiagen). Using 5 μ L of the PCR products as template, eight additional cycles of PCR was conducted with Illumina primers containing Nextera XT indexes (Illumina, San Diego, CA) and Phusion High-Fidelity DNA Polymerase in a final volume of 50 μ L, then purified using Agencourt AMPure XP magnetic beads. The amplicon concentrations were measured using

Quant-iT Picogreen dsDNA assay kit (Thermo Scientific). Libraries were pooled equimolar and sequenced by Illumina MiSeq for 2×300 bp reads at Eurofins Genomics (Ebersberg, Germany).

The quality of raw reads was visualized using FastQC v0.11.5¹⁵ followed by first-pass quality filtering using Trimmomatic v0.39¹⁶ with the following parameters: SLIDINGWINDOW:5:20 AVGQUAL:20 minLEN:200. A big data pipeline was used to infer Ribosomal Sequence Variants (RSVs) using the DADA2 v1.20 considering the following parameters: truncLen=c(265,220), trimLeft=c(17,21), maxEE=c(2,2), truncQ=c(2,2), maxN=0, rm.phix=TRUE.¹⁶ We carried out error correction for the samples sequenced across multiple sequencing runs separately until chimera removal step as the error rates might differ between runs. Resulting non-chimeric RSVs were again chimera filtered using reference-based chimera filtering implemented in USEARCH v11¹⁷ with the ChimeraSlayer Gold database v2011051967. Taxonomy was assigned to non-chimeric sequences using assignTaxonomy function using SILVA database v138 with a bootstrap confidence threshold of 80%.¹⁸ Additionally, we used SPINGO for species level classification with the same reference database whenever possible.¹⁹

Initial pre-processing of 16S RSV table was conducted using the CoDaSeq package,²⁰ whereby rare RSVs present in less than 5% of the samples in each of the datasets (gDNA biopsy, cDNA biopsy and stool) were removed using the codaseq.filter function. Overall, a total of 520 RSVs were retained with 318 RSVs from the gDNA biopsy, 435 RSVs from the cDNA biopsy and 361 RSVs from stool dataset. Except in the case of alpha diversity, this filtered RSV count table was used for all the downstream bioinformatic analysis.

Host RNA-Seq data was generated from mucosal biopsy samples as described previously.⁵ Trimmomatic v0.39 was used to trim adapters and remove low quality reads.²¹ Reads were quality checked using FastQC and multiQC before and after trimming. The quality filtered reads were then aligned to the human genome (GRCh38) using Hisat2 v2.1.0.²² A count table was then generated using featureCounts v1.5.0 using default parameters.²³

Host epigenome

Epigenetic data for CD samples (37 inflamed and 37 non-inflamed) and control samples ($n = 22$) was generated using the Illumina Infinium HumanMethylation 450 BeadChip array as described previously.⁵ In addition to this, epigenetic data is now also available for UC samples (30 inflamed and 46 non-inflamed) and some additional control samples ($n = 14$; 13/14 generated on both arrays) that were assayed using the Illumina Infinium HumanMethylation EPIC array. For both datasets, pre-processing and quality control was implemented using R libraries *minfi* and *minifiData*. Beta values were extracted and filtered using *BMIQ*²⁴ for normalization between probe types with R libraries *methylumi* and *wateRmelon*. Probes were removed if the probe sequence mapped to multiple positions in the genome, mapped to sex chromosomes, had missing data or mapped to SNPs.

Host genotype

Genotyping was carried out as described in Ryan et al.⁵ Due to the limited sample size, GWAS was not possible. Like in our previous study, we included 264 loci in our analysis which had been previously associated with IBD. This dataset was only considered for machine learning analysis.

16S G4 phylochip data

Libraries were prepared as previously described.²⁵ Briefly, full-length V1-V9 16S rRNA genes were amplified from extracted DNA and purified. Approximately, 3.0×10^{10} double-stranded DNA molecules (500ng) from each library were combined with non-16S DNA spike-in controls, digested, and labeled. Each library was denatured and hybridized to its own G4 PhyloChip (non-multiplexed) in a 96-well high-throughput format, fluorescently stained and then scanned in a GeneTitan MC using GeneTitan Hybridization, Wash and Stain Kit for WT Array Plates (Thermo Fisher, Santa Clara, CA). The G4 PhyloChip queries 610,038 different 16S rRNA loci where a locus is defined as a 25-mer nucleotide sequence within a 16S rRNA gene within any of the 10

conserved regions or any of the 9 variable regions. Standard Affymetrix software (GeneChip Microarray Analysis Suite, ThermoFisher Scientific) was used to capture the scans of the array. Only perfect-match probes with fluorescence intensity observed in at least three samples were exported for rank-normalization in Sinfonietta software²⁶ (Second Genome Inc, South San Francisco, CA) and were used as input to empirical probe-set discovery. All probe sets contained three or more probes and the empirical Operational Taxonomic Units (eOTU) tracked by a probe set were taxonomically annotated using StrainSelect version 2016.²⁷ Analyses were conducted on hybridization scores (HybScores), which are the mean normalized rank for all probes within an eOTU. The probes were ranked according to their scaled-background subtracted fluorescence intensities. A total of 502 eOTUs were included in this analysis.

Bioinformatics analysis

Microbiome data

Statistical analysis and visualizations were performed in R v4.0 using the packages *vegan*,²⁸ *zCompositions*,²⁹ *Tidyverse*,³⁰ *rstatix*,³¹ *EnhancedVolcano*³² and *ggplot2*.³³ Due to the complex compositional nature of the microbiome data, we applied a centered log-ratio transformation (CLR) to each sample in our dataset. Datasets, such as 16S rRNA sequencing data, which are generated by next-generation sequencing (NGS) technologies are inherently compositional as the total number of reads produced are limited to the sequencing depth.²⁰ As a result, many standard statistical approaches may not be appropriate as the independence assumption between features does not hold. The CLR transformation has, thus, been suggested as a suitable approach when conducting compositional data analysis (CoDA) as it compares log-ratios rather than raw sequencing counts.²⁰ We first imputed the zeros in the abundance matrices using a count zero multiplicative replacement method (*cmultRepl*, *method* = "CZM") implemented in the *zCompositions* package. Following this, we applied the CLR transformation using the

codaSeq.clr function from the *CoDaSeq* package. The CLR transformation was applied separately to each taxonomic level in the RSV table (from phylum to species level). The Shannon diversity index was used to estimate the species richness and evenness and Wilcoxon test was used to evaluate statistical significance between clinical variables (e.g., disease type, biopsy type and future relapse). Paired testing was performed whenever possible. For beta diversity analysis, we used compositionally coherent Aitchison distance matrix and applied pairwise Permutational Multivariate Analysis of Variance (PERMANOVA) with 9,999 permutations to quantify community level differences.

Differential abundant taxa and genes were identified using ALDEx2 (ANOVA-Like Differential Expression), a compositionally-robust differential abundance analysis approach. ALDEx2 estimates per-feature technical variation within each sample using Monte-Carlo instances ($n = 512$) drawn from the Dirichlet distribution.³⁴ This distribution maintains the proportional nature of the data.²⁰ ALDEx2 uses the Centred Log-Ratio (CLR) transformation that ensures the data are scale invariant and sub-compositionally coherent. ALDEx2 measures the effect size and returns p-value as calculated by Wilcoxon test along with Benjamini-Hochberg (BH) adjusted p-value. Effect sizes are the ratio of the between-group differences to the maximum of within-group differences. For this analysis, we filtered-out rare and low abundant taxa further (recommended as to decrease sparsity in the dataset) by retaining only those taxa present in > 5% of the sample with mean abundance of > 0.01% in at least one of the two groups under comparison. We used a BH adjusted p-value cutoff of 0.1 for microbiome data and a BH adjusted p-value threshold of 0.1 and an effect size threshold of 0.80 for gene expression data.

Host gene expression and pathway enrichment analysis

In the analysis of host gene expression data, we focused only on protein encoding genes, and we filtered out genes expressed in fewer than 25% of

the samples for each disease type separately, retaining 9,925 unique genes for downstream analysis. As this dataset was also generated using NGS technologies, we applied a CLR transformation prior to conducting any downstream analysis.

To identify pathways found to be associated with a disease type, we implemented an enrichment analysis using Fisher's exact test. We used the set of expressed genes input as the background genes and the set of genes associated with a disease type as the genes of interest. We used the KEGG and PID gene sets from the MsigDB canonical pathways collection.³⁵ To avoid pathways that were too large to provide any specific biological insights or too small to provide adequate statistical power, we excluded from our analysis any pathways with more than 85 genes, fewer than 10 genes, or fewer than 5 genes that overlapped between the pathway and the genes of interest. The p-values obtained from Fisher's exact test were adjusted for multiple testing using the Benjamini-Hochberg (FDR) approach.

Host DNA methylation

As the epigenetic data was generated using two different arrays, all analysis conducted on this data type was repeated for each array separately. Prior to initial analysis, the raw beta values were normalized (to $N(0, 1)$) using Quantile Normalization. Principal Component analysis (PCA) was then conducted on this normalized DNA methylation data to identify any differences in methylation between groups under consideration. We also applied pairwise PERMANOVA test on the Euclidean distance matrix using 9,999 permutations to identify any community level differences between groups.

To identify CpG sites associated with disease-type, inflammation status and relapse status, either a linear mixed effect regression model (*lme4* package) or a linear regression model was used, depending on the underlying samples being used. If the underlying samples included paired samples from the same patient, a mixed effect model was used and the patient ID was included as a random effect. In both model types, condition, inflammation status, gender, age, methylation chip, sample position on methylation chip and biopsy location were

included as fixed effects. Reported p-values were then adjusted for multiple testing using the Benjamini-Hochberg correction. To account for cell heterogeneity, this epigenetic association analysis was repeated for each set of significant epigenetic signals but this time incorporating the first 10 principal components (PCs) as covariates in the model. Only those CpG sites that were also significant ($p < 0.05$) based on the PC model were considered significantly associated with the outcome of interest (disease-type, inflammation status, relapse status).

Following this, we used the mCSEA package from Bioconductor to identify differentially methylated regions (DMRs) based on these significant CpG sites.³⁶ In this analysis, we primarily focused on promoter regions and gene bodies. CpG sites were determined to be in a promoter region or gene body using the annotation R packages *IlluminaHumanMethylation450kanno.ilmn12.hg19* and *IlluminaHumanMethylationEPICanno.ilm10b2.hg19*. Briefly, promoter regions were defined as those CpG sites whose *UCSC_RefGene_Group* column was either TSS1500, TSS200, 5'untranslate region [UTR] or 1stExon. CpG sites belonged to a gene body region if the *UCSC_RefGene_Group* was "Body". Prior to identifying DMRs, CpG probes were first ranked using the *rankProbes()* function, with paired analysis conducted where necessary. Raw beta-values were input into this function and converted to M-values prior to calculating the linear models used to rank the CpG sites. In the cases where paired analysis was performed, the patient ID was supplied as the *pairColumn* parameter. Once CpG sites were ranked, the *mCSEATest()* function was used to identify differentially methylated regions (DMRs). Only those regions with an FDR-adjusted p-value < 0.05 were considered differentially methylated.

Integrating methylation and expression data

As both methylation and expression data were available for a subset of the cohort, we used the mCSEA package to integrate the two data types in order to identify significant associations between methylation changes in a DMR and expression alterations in a nearby gene. For

this analysis, samples from methylation and expression datasets were matched by their sample ID, ensuring consistency with the patient ID, disease type, and inflammation status of the sample. Only those DMRs which were significant by *mCSEATest*, using the less stringent cutoff of a p-value less than 0.05, were considered in this analysis. The *mCSEAIIntegrate* function was used to perform a correlation test between the mean DMR methylation and the expression of close genes. Only those regions with a correlation greater than 0.5 and an adjusted p-value less than 0.05 were recorded. By default, the package only reports negative correlations between promoter methylation and gene expression and positive correlations between gene body methylation and gene expression.

Integration of host omics and gut microbiota data

We implemented a lasso penalized regression approach to identify specific associations between individual host features and gut microbial taxa as outlined in *Sambhawa Priya et al., 2022*.¹⁰ Samples from different data types were again matched by their sample ID, as described previously. This was repeated for both host gene expression and host DNA methylation features. Given the large number of CpG sites included in the methylation arrays, we first merged the CpG sites into methylated regions using information from the *mCSEADATA* package. Similar to before, we considered only promoter and gene body regions and used the mean methylation value across the CpG sites associated with these regions to represent the methylation of that region. We implemented this analysis for each disease group (i.e., CD, UC and control) and their tissue biospecimen (i.e., inflamed or noninflamed) separately. For further analysis, we conducted correlation analysis on stability selected host genes-taxa associations. The Spearman correlation coefficient (ρ) was used to depict the strength of association, while a Benjamini-Hochberg test was used to correct multiple testing problem. Pathway enrichment analysis for the

host genes that were associated with specific gut microbes ($q < 0.1$) were carried out using Fisher's exact test as outlined above.

Succinotypes

The premise of succinotypes is that gut microbiomes typically have either *Phascolarctobacterium* or *Dialister* as their dominant succinate-consumer, and thus subjects can be classified as either P-type or D-type, respectively.³⁷ Here, we classified the subjects into P and D types and tested for an association with disease.

To perform this classification, we first identified all RSVs that were classified on the genus level as either *Dialister*, *Phascolarctobacterium*, or *Phascolarctobacterium_A* using the *assignTaxa* function from the *Dada2* package with the *GTDBr95* database and an inclusive bootstrap cutoff of 0.2. This returned 7 RSVs, which we additionally aligned against the SSU references from *GTDB* with *BLAST*. All 7 RSVs had a perfect alignment to at least one of the references, confirming that the taxonomic classification was correct.

To assign succinotypes, we followed the same procedure as in *Anthamatten et al.*³⁷ We then computed the read counts of *Dialister* (D) and *Phascolarctobacterium* (P) in each sample by summing the read counts of the respective RSVs and merging *Phascolarctobacterium* and *Phascolarctobacterium_A*. For each sample, we then computed the relative ratio of *Dialister* as $r_D = n_D / (n_D + n_P)$, where the n are the read counts of D and P, respectively. We assigned a clear D-type to a sample if $r_D > 0.9$ and a clear P-type if $r_D < 0.1$, implying 10× higher abundance of *Phascolarctobacterium* vs. *Dialister*, or *vice versa*. Samples with fewer than 10 combined D and P reads were not considered. If all samples from a subject had the same succinotype assignment, then this assignment was directly given to the subject. For subjects with discordant succinotype assignments between samples, we differentiated between those that had at least one clear assignment and otherwise mixed assignments ($0.1 < r_D < 0.9$) and those that had fully discordant assignments. Those with clear and mixed

assignments retained their clear assignment. One subject had fully discordant assignments, with and $r_D = 1$ in the fecal sample and $r_D = 0$ in all biopsies. For this sample, we assigned the succinotype of the biopsy.

We tested for associations between succinotypes and disease – both combined UC and CD or separate – and between whether patient had a relapse or remained in remission using Fisher's exact tests. To test for an association between the number of relapses experienced by a subject and their succinotype, we initially used a non-parametric Mann-Whitney U-test. Following this, we implemented a zero-inflated Poisson regression approach to better understand the rate at which relapses occur. This model first splits the subjects into two groups: one in which the subjects are in long-term remission with a zero probability of relapses during the observation window (4 years) and a second in which relapses occur at a non-zero rate per year. A maximum likelihood estimation for different models types was performed to identify the models which best fit the data. We defined the (minus) log-likelihood function for the model by choosing a parametrization where we can estimate a joint theta (probability of long-term remission) for all groups or a separate theta value for each group. We always estimate a separate rate lambda (relapse rate) for each group. We compared models that grouped subjects based on succinotype, disease, succinotype and disease, and also CD/Dialister versus others. Once the best model was identified as CD/Dialister vs. others based on the Akaike Information Criterion (AIC), a log ratio test of this model versus one with a single lambda for all subjects was used to obtain a p-value for the differences in relapse rates between groups included in the model.

ML analysis

ML analysis was conducted using the boosted decision tree algorithm, eXtreme Gradient Boosting (XGBoost).³⁸ Six omics data types were considered in this analysis including three microbiome (16S gDNA, 16S cDNA and 16S G4 Phylochip) and three host-omics data types (host genotype, host RNA-Seq and host epigenome). We also considered the case where patient age was included as an

additional feature.³⁹ Models were trained using individual data types as well as multi-omics combinations with a total of 86 combinations assessed for each scenario (disease type and inflammation status).

Multi-omics datasets were combined using a concatenation approach and only samples for which we had full coverage across the data types were considered when training the model. Like previous analyses, samples from different data types were matched based on their sample ID and in the case of host genotype, were matched using patient ID. A full list of the available data types for each sample are provided in Supplementary Table S1. To reduce the dimensionality of the datasets, a number of feature reduction steps were considered. 16S gDNA, 16S cDNA and Host RNA-Seq data types underwent feature selection as described in earlier sections. We further reduced the dimensionality of both 16S sequencing datasets by agglomerating to genus level. For the host genotype, a subset of 264 SNPs which had been previously associated with IBD were considered.⁵ In addition to these feature reduction steps, only features with non-zero values in at least 10% of samples were considered and features with near-zero variance were removed. Similar to previous analyses, a CLR transformation was applied to NGS data types to account for the compositional nature of the data while, in this analysis, beta-values were used to represent the host epigenome dataset.

Given the limited sample size and the lack of an external validation set, a nested cross validation approach was implemented to train and assess the performance of our models (Supplementary Figure 13). The outer loop was a Leave-One-Out (LOO) Cross Validation (CV) and where more than one sample existed for a particular patient, all additional samples were excluded from the training data, to avoid introducing any bias into our pipeline. The held-out sample was used to assess the performance of the trained model and was not used for any other purpose. During the internal model development phase, on the training data an additional feature selection step was applied. A Wilcoxon rank-sum test was used to identify any potential associations between features and the outcome of interest. Those features found to have a significant difference ($p < 0.05$) between

the relapse and remission groups were selected to be included in the training datasets. Model hyperparameters were selected using a randomized search based on a 5-fold CV. That is, we assessed the performance of 250 random combinations of hyperparameters using a 5-fold CV on the training data and those parameters with the highest mean area under the ROC curve (AUC) were chosen as the optimized hyperparameters. The following hyperparameters were tuned in each case *max_depth* (range: 5–80), *colsample_bytree* (range: 0.5–0.8), *subsample* (range: 0.5–0.8) and *alpha* (range: 0–150).

Once optimal hyperparameters were selected, 10 XGBoost models were trained on different subsets of the data by splitting the training set into 10-folds. Nine folds were used to train the model, and the remaining fold was used as model check-point and for early stopping. That is, if performance on this held-out fold did not improve after 20 rounds, training was stopped, and the model was saved. For the inference, the ensemble of 10 models was run and the average prediction from the 10 models was stored and was used to assess the performance on external held-out data. When splitting the training data into folds to create the ensemble of 10 models, predictions on the held-out *validation* fold were calculated for each model trained. These predictions represent the best achievable performance and were used to estimate the validation set performance of our ML approach. The AUC metric was primarily used to evaluate performance in each case (test and validation) as it is a threshold independent metric. Other threshold dependent metrics such as accuracy, sensitivity, specificity and F1-score were also calculated for the test set with the optimal threshold calculated using the geometric mean (G-mean) approach. The G-means is defined as the square root of the product of sensitivity (TPR) and specificity (TNR).

$$G\text{-means} = \sqrt{\text{Sensitivity} \times \text{Specificity}}$$

The threshold for classification is selected such that the g-means value is maximized. As a result, this threshold represents the best trade-off between sensitivity and specificity, ensuring balance performance for both classes.

As part of the ML analysis, we also wanted to assess which features were playing a role in the trained models. This was done using two different techniques (i) feature importance (gain) from XGBoost models and (ii) Shapley additive explanation (SHAP) values.⁴⁰ SHAP values show the contribution of each feature on a prediction of the model. In order to assess the overall impact of each omics type used in the model, we grouped the SHAP values by summing the values of all features in a particular omics dataset. Importance values for each feature, extracted from the trained XGBoost models, were averaged across each iteration of the CV and were normalized by the number of times it was found to have a non-zero value in an ensemble. This ML analysis was conducted using the following python packages: *xgboost*, *pandas*, *scikit-learn*, *numpy*, *shap* and *scipy*.

Results

Cohort demographics

The study cohort comprised 142 subjects with IBD, including patients with CD and UC, and 34 controls, as described previously.⁵ A total of 295 mucosal biopsies were collected from paired inflamed (i) and non-inflamed (ni) colonic sites from patients with IBD, and non-inflamed sites from control subjects (Table 1; Supplementary Table S1). The slightly higher number of samples compared with our previous study⁵ was due to the availability of new data types. We also chose to include a maximum of one inflamed and one non-inflamed biopsy sample per IBD subject (85% paired) and only one non-inflamed sample per control subject to avoid any biases that may arise from including multiple samples of the same type from the same patient. Stool samples were collected from a subset of the cohort ($n = 39$) for additional analysis (Table 1), whereof half of them were collected from patients on the same day as the colonoscopy (prior to bowel preparation) and the remaining within 3 years of biopsy collection. Additional patient information was gathered on disease related outcomes after the time of sampling (max 4 years), including future relapses, treatment with monoclonal

Table 1. Subject characteristics, patient outcomes and data types of study cohort.

	Crohn's Disease	Ulcerative Colitis	Controls	Total
Subjects	54	88	34	176
Biopsy Samples				
Inflamed	52	85	-	137
Non-Inflamed	50	76	32	158
Stool Samples	14	21	4	39
Age (Mean (SD))	41.6 (12.0)	47.9 (13.0)	55.7 (12.5)	47.5 (13.5)
Gender (M/F)	29/25	46/42	18/16	93/83
Patient Outcomes 1–4 years after endoscopy (%)				
Future Relapses (<i>n</i> = 140)	40.7%	43.0%	-	42.1%
Monoclonal antibody treatment (<i>n</i> = 134)	20.4%	11.3%	-	14.9%
Structural GI complications (<i>n</i> = 140)	22.2%	1.2%	-	9.3%
Surgical intervention (<i>n</i> = 142)	16.7%	2.3%	-	7.8%
Number of patients on Medication at time of sampling				
Biologics	6	10	-	16
Corticosteroids	7	11	-	18
Mercaptopurine	13	6	-	19
Mesalazine	5	36	-	41
Nexium	2	2	5	9
Data types: Biopsy				
16S gDNA	85	131	30	246
Inflamed (i)	42	66	-	108
Non-inflamed (ni)	43	65	30	138
16S cDNA	83	133	25	241
Inflamed (i)	45	73	-	118
Non-inflamed (ni)	38	60	25	123
Host RNA-Seq	87	132	26	245
Inflamed (i)	44	71	-	115
Non-inflamed (ni)	43	61	26	130
Host DNA Methylation (Epigenome)	74	76	23	173
Inflamed (i)	37	30	-	67
Non-inflamed (ni)	37	46	23	106

antibodies, surgical intervention, and structural GI complications (Table 1). A patient was reported as having a relapse if there was a recurrence of symptoms and objective evidence of disease as assessed by endoscopy (sigmoidoscopy or colonoscopy) for patients with UC, and by endoscopy or CT scan for patients with CD.

Microbiota composition of IBD subtypes is different to controls

We examined the mucosal microbiota using amplicon sequencing of both 16S rRNA genes (gDNA or standing microbiota; Table 1) and transcripts (cDNA or active microbiota; Table 1). The latter amplicon was added because gDNA cannot differentiate between dead or alive cells⁷ and also because it captures the metabolically active microbial community for a more functional view. 16S gDNA data from stool samples, collected as described previously, was also available for a sub-cohort (14 CD, 21 UC, 4 controls) allowing us to

compare microbial composition between sample types for available patients (Table 1). As these samples were not collected in an RNA preservative, 16S cDNA data was not available.

Microbiota analysis was carried out on a total of 11.4 million error-corrected, non-chimeric ribosomal sequence variant (RSV) reads with a mean count of $21,686 \pm 8,667$ SD using updated methods from previous analysis⁵ (Supplementary Figure S1). In total 12,006 unique RSVs were identified across all 16S rRNA amplicon data types, whereafter filtering for those present in at least 5% of samples in at least one dataset (gDNA biopsy, cDNA biopsy, gDNA stool) resulted in a unique set of 520 RSVs. Any samples with less than 5,000 reads were also removed from further analysis.

Unlike the original study,⁵ beta diversity analysis was performed based on Aitchison distances to better account for the compositional nature of the data. We observed a significant (PERMANOVA, $p < 0.05$) disease-associated shift in microbiome composition along the first two principal components (PCs) for both mucosal 16S rRNA datasets, with CD samples

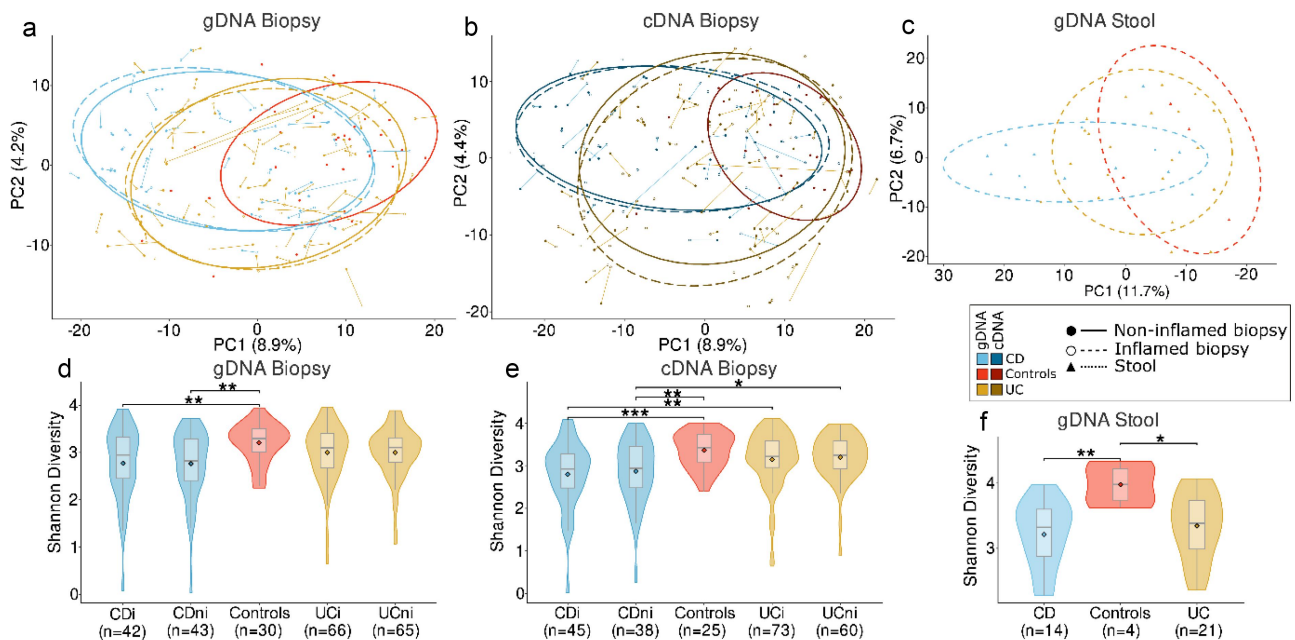


Figure 1. Microbiota composition and diversity of Crohn's disease (CD), ulcerative colitis (UC) and control subjects. Principal component analysis (PCA) based on Aitchison distances of all RSVs present in > 5% of samples in at least one dataset under consideration for (a) gDNA biopsy, (b) cDNA biopsy and (c) gDNA stool datasets, respectively. Samples are grouped by disease type and inflammation status. Points connected by lines highlight samples from the same patient. d-f) Comparison of alpha diversity using the Shannon diversity index for each 16S rRNA dataset. Diversity is compared for each disease type and inflammation status. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

showing the most significant shift from control samples (Figure 1a-b; Supplementary Table S2). The microbiota composition of available stool samples also showed a significant disease-associated shift, with differences observed between CD and controls, and between CD and UC subjects (Figure 1c). Again, CD samples showed the most significant shift away from controls. For a subset of patients with CD ($n = 29$), we had information on ileal involvement in their disease. Beta diversity analysis showed a significant PERMANOVA difference in the standing microbiota of CDi samples from those patients with ileal involvement and those without (Supplementary Figure S2a). This was not observed when examining the active microbiota and no significant difference was observed between these groups when analyzing non-inflamed samples for either data type (Supplementary Figure S2a-b).

In line with previous findings,⁵ CD mucosa exhibited a lower microbial diversity than controls in 16S gDNA data, while inflammatory status of samples did not significantly affect diversity levels (Figure 1d; CDi vs controls $p = 0.02$; CDni vs

controls $p = 0.01$). These differences were also evident for active microbiota (Figure 1e; CDi vs controls $p = 0.001$; CDni vs controls $p = 0.004$), even though CD samples also demonstrated significantly lower diversity than UC samples (UCi vs CDi $p = 0.005$; UCni vs CDi $p = 0.003$). In contrast to the previous study,⁵ there was no significant difference in alpha diversity between UC and controls for either amplicon data type (Supplementary Table S2). In stool samples, both CD and UC patients displayed a significantly lower alpha diversity compared to controls (Figure 1f: CD vs controls $p = 0.008$; UC vs controls $p = 0.049$). However, no difference in diversity was found between UC and CD stool samples, or for the subset of patients with known ileal involvement (Supplementary Figure S2c-d).

Differential abundance analysis of RSVs from mucosal samples was performed using compositional-aware ALDEx2 to compare IBD and controls (Figure 2). While it was possible to detect several differentially abundant taxa, these findings were not always consistent across

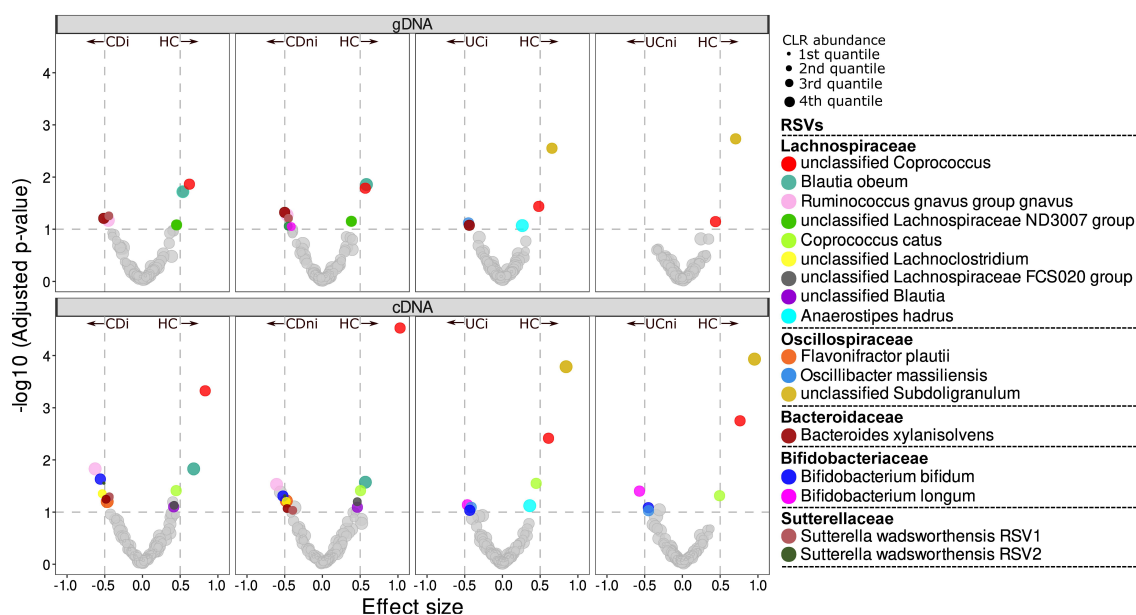


Figure 2. Volcano plots of differential abundance analysis comparing RSV abundances of CD and UC subjects to controls. Analysis was repeated for each inflammation status and 16S rRNA data type. Points above the horizontal line represent those taxa with an adjusted p-value (q) < 0.1 and those outside the vertical lines have an effect size $\geq \pm 0.5$. Species abundance is denoted by point size where 4th quantile denotes that the species is in the least abundant category.

amplicon datasets. For all comparisons of disease type and inflammation status, only an unclassified *Coprococcus* RSV (previously identified as unclassified *Lachnospiraceae*)⁵ was consistently less abundant in both IBD subtypes compared to controls across both 16S data types. *Blautia obeum* and unclassified *Subdoligranulum* RSV were also found to be less abundant in CD and UC, respectively, when examining both active and standing abundances (Figure 2 & Supplementary Figure S3). In contrast to previous analyses,⁵ we did not find a significant difference in *Anaerostipes hadrus* abundance when comparing CD to controls. It was, however, less abundant in UC compared to controls in terms of statistical significance, but the effect size was below the chosen threshold (Figure 2). No significant differences were observed between inflamed and non-inflamed mucosa.

To add an outcome-predictive component, we repeated the above analysis on patients that had at least one relapse within four years of endoscopy sampling and those that remained in remission. Beta diversity was not significantly different between relapse and remission for either 16S amplicon datasets (Figure 3a-b;

Supplementary Table S3). Somewhat counterintuitively, relapsing patients with CD had a higher alpha diversity in their inflamed samples compared to those in remission ($p < 0.05$; Figure 3c; Supplementary Table S3). This was however only observed when considering 16S gDNA abundances and not corroborated in active microbe abundances (Figure 3d). Given the limited number of CD subjects with information on ileal involvement and relapse status, we could not compare these groups. Similarly, we found no bacterial taxa to be differentially abundant in either relapse or remission after adjusting for multiple testing (see Supplementary Figure S4 & S5 for significant RSVs before adjustment).

Integrating host omics data with the gut microbiome

In addition to microbiome data, host omics datasets were also generated from the same mucosal biopsies,⁵ including host transcriptome (245 biopsies from 147 patients; Table 1) and host epigenome data (173 biopsies from 106 subjects; Table 1). As outlined in methods, host epigenome samples were generated on two different Illumina arrays and due to this batch effect, we conducted all

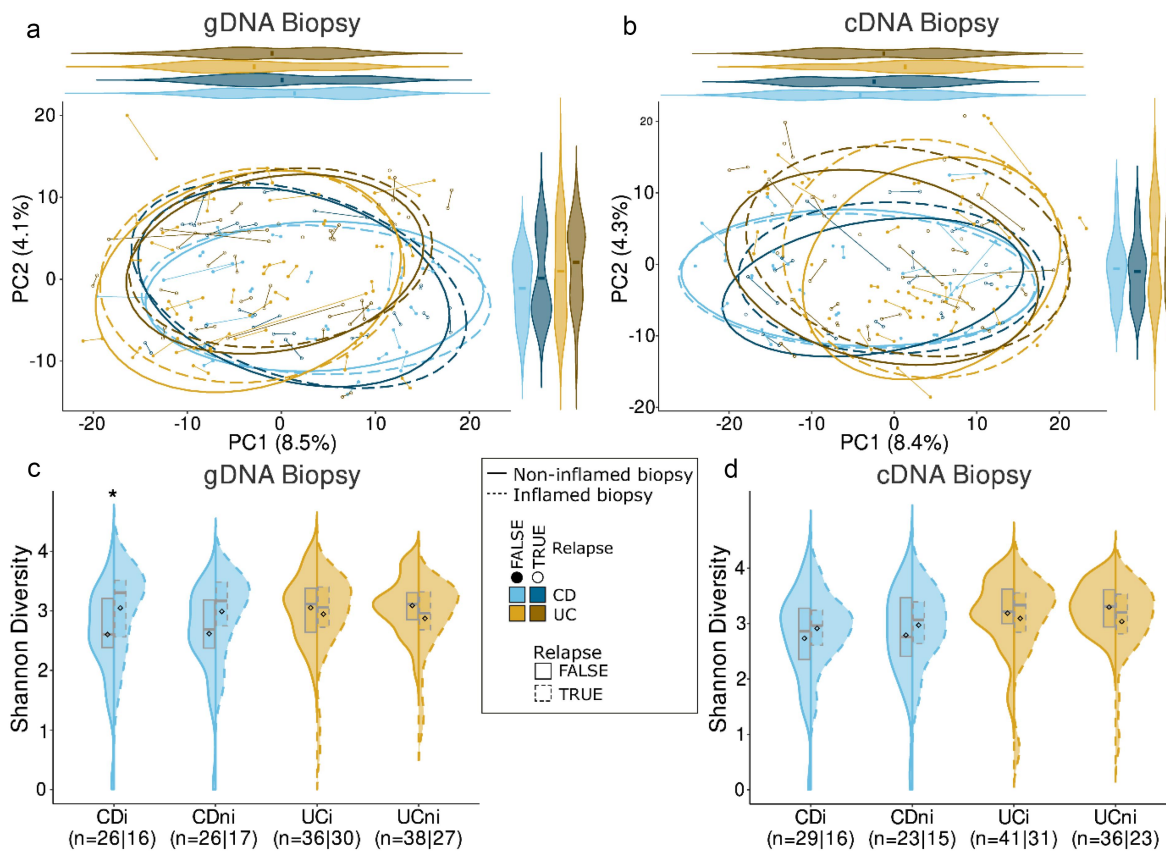


Figure 3. Comparison of microbiota composition and diversity between patients with IBD who experienced a relapse of disease and those who remained in remission within 4 years of sampling. (a-b) Principal component analysis (PCA) based on Aitchison distances grouped by relapse status, inflammation status and disease type with analysis repeated for 16S gDNA and cDNA datasets, respectively. (c-d) Comparison of Shannon alpha diversity between relapse and remission groups for gDNA and cDNA datasets respectively. * $p < 0.05$.

epigenetic analysis separately for each array. We observed significant disease-associated shifts in gene expression and methylation between patients with IBD and controls with CDi samples being the furthest away from controls (Figure 4a-c; PERMANOVA $p < 0.05$; Supplementary Table 4 & 5). Significant inflammation-associated changes in host omics were also identified within both CD and UC samples (Figure 4a-c; $p < 0.05$; Supplementary Table 4 & 5) and this was corroborated by a strong epigenome-transcriptome correlation between the inflammation-associated PC1 values (Figure 4d-e; 450K array $R^2 = 0.8$; EPIC array $R^2 = 0.87$).

Differential expression analysis further highlighted this inflammation-associated change in gene expression, with 704 and 1,134 differentially expressed genes (DEGs) identified in CDi and UCi samples, respectively, compared to controls ($q < 0.05$; effect size $\geq \pm 0.8$; Supplementary Table S6 &

7). These DEGs corresponded to 66 enriched pathways of which 29 were common to both IBD subtypes and included pathways involved in fatty acid, amino acid and carbohydrate metabolism, along with integrin and interleukin pathways (Supplementary Figure S6). The central regulator gene *ETS2*, recently reported as causative of macrophage inflammation in IBD,⁴¹ was also among those DEGs upregulated in IBD. There were, however, no differences in expression across patient groups with the corresponding risk SNP (*rs2836882*).

Examination of individual CpG sites using mixed-effect models also highlighted the difference in methylation by inflammation status within IBD subtypes. We identified 14,601 and 35,322 CpGs that were significantly associated with inflammation in CD and UC samples, respectively. These inflammation-associated CpGs corresponded to three differentially methylated promoter regions

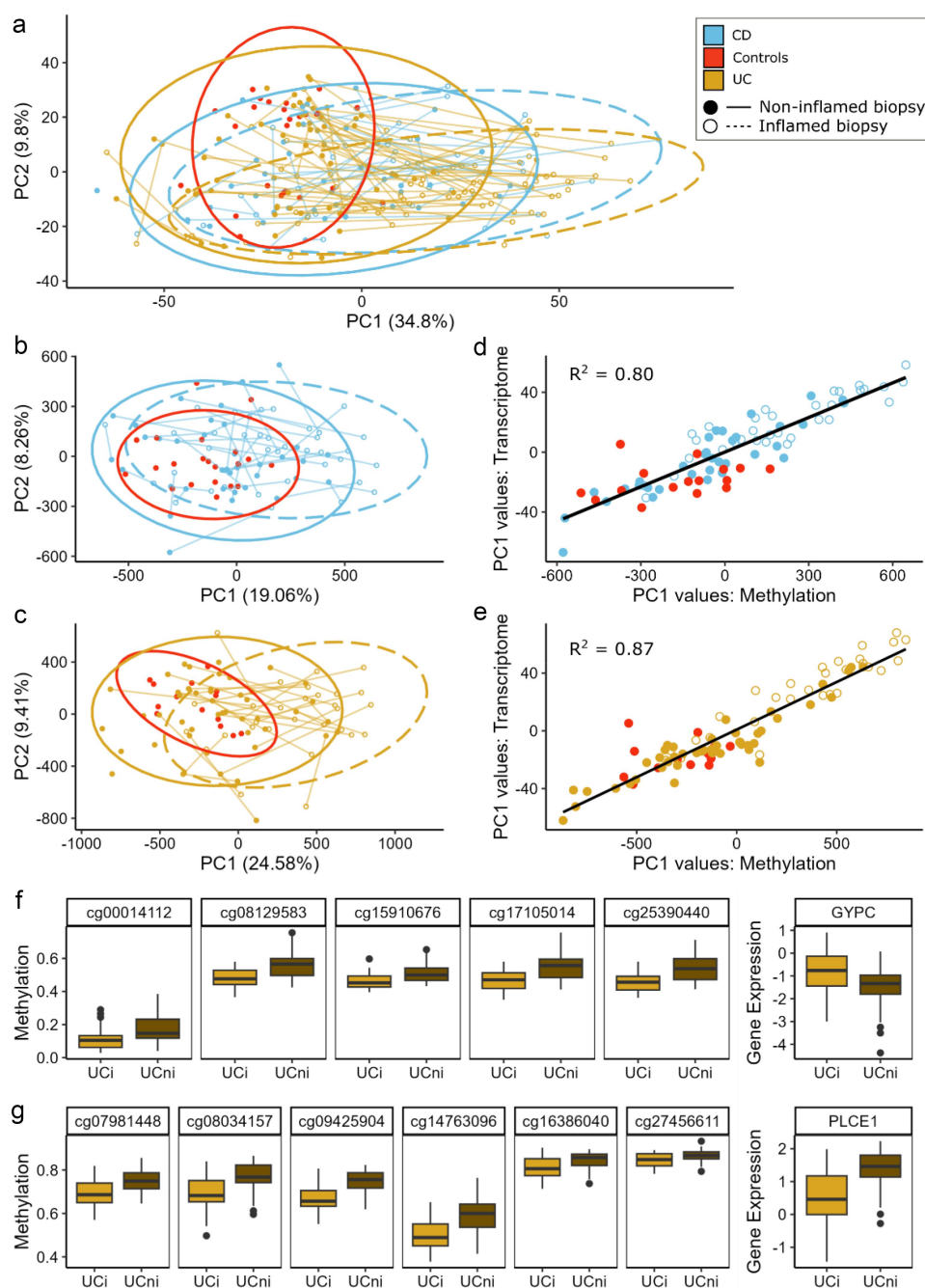


Figure 4. Host gene expression and DNA methylation in Crohn's disease (CD), ulcerative colitis (UC) and control subjects. (a) PCA plot based on Aitchison distances of Host RNA-Seq data grouped by disease type and inflammation status of samples. Points connected by lines highlight those samples from the same patient. (b-c) PCA plots of Host epigenetic data grouped by disease type and inflammation status for those samples generated using the 450K and EPIC methylation arrays, respectively. (d-e) Plot of PC1 values comparing Host methylation and Host transcriptome for each methylation array. (f) Boxplots of methylation of CpG sites associated with the promoter region of the *GYPC* gene and its corresponding gene expression. (g) Boxplots of methylation of CpG sites associated with the gene body of *PLCE1* and its corresponding gene expression.

(*FLJ44606*, *UTS2D*, *HTR2A*, $q < 0.05$) and one differentially methylated gene body (*HLA-DPB1*) in CD and 19 promoter regions and 11 genes bodies in UC (Supplementary Table 8). Using the mCSEA

package³⁶ and a less stringent cutoff for differentially methylated regions (DMRs; $p < 0.05$), it was possible to examine the connection between methylated regions and the corresponding gene

expression. In UCi samples, we found 9 promoter regions and 10 gene bodies whose methylation was significantly correlated with the expression of the corresponding gene (Supplementary Table 9). For example, a strong negative correlation was observed between the methylation of CpG sites in the promoter region of the *GYPC* gene and its corresponding expression (Figure 4f; $\rho = -0.74$) while a strong positive correlation was found between the expression and methylation of the *PLCE1* gene (Figure 4g; $p = 0.8$). In CDi samples, a moderately strong positive correlation was reported between the methylation and expression of genes *AGAP1* ($\rho = 0.66$; Supplementary Figure 7) and *PTPRN2* ($\rho = 0.65$; Supplementary Figure 8).

As matched 16S amplicon data with host RNA-Seq data (215 cDNA and 209 gDNA) and host epigenome data (160 gDNA and 141 cDNA) was available for a subset of samples, we implemented a penalized regression approach to elucidate individual host-microbe associations.¹⁰ Given the large number of CpG sites produced by Illumina arrays and resulting multiple testing issues, we merged these sites into methylated promoter and gene body regions. Enrichment analysis of host gene expression identified relationships between 54 genes and six bacterial taxa across eight pathways

(Table 2; Fisher's exact test; $q < 0.1$). Of all gene-microbe associations identified, only one pathway was significantly enriched based on associations with both the standing and active microbiota of the same taxa. In UC patients, we found the expression of genes from the Integrin beta-1 pathway to be positively correlated with both the standing and active abundance of the *Parasutterella* genus (Supplementary Figure 9).

We identified links between the methylation of 64 promoter regions and two microbial taxa across 27 pathways in CDi samples (Fisher's exact test $q < 0.1$; Supplementary Figure 10). A total of 21 pathways were enriched based on promoter regions associated with the standing abundance of the *Lachnospiraceae* *UCG-004* genus, and a further six pathways were associated with the standing abundance of a *Lachnospiraceae* *CAG-56* RSV. None of these associations were significant when considering the active abundances of these taxa and no enriched pathways were identified based on promoter-microbe associations in UC samples. No pathways were enriched based on associations between methylated gene bodies and microbes for either IBD subtype.

As with the microbiome analysis, we compared relapse and remission groups within both IBD subtypes and omics types but found no significant

Table 2. Pathways enriched for significant gene-microbe associations.

Pathway	DB	Taxa Name (Level)	q	# of Genes	Genes Names
CD Inflamed Samples (gDNA)					
Cardiac Muscle Contraction	KEGG	<i>Ruminococcus torques</i> group <i>torques</i> (RSV)	****	14	<i>COX41I</i> , <i>COX5B</i> , <i>COX6A1</i> , <i>COX6B1</i> , <i>COX7A2</i> , <i>COX7A2L</i> , <i>COX7B</i> , <i>COX7C</i> , <i>COX8A</i> , <i>UQCR10</i> , <i>UQCR11</i> , <i>UQCRB</i> , <i>UQCRH</i> , <i>UQCRQ</i>
Proteasome	KEGG	<i>Ruminococcus torques</i> group <i>torques</i> (RSV)	*	9	<i>PSMA7</i> , <i>PSMB1</i> , <i>PSMB3</i> , <i>PSMB7</i> , <i>PSMC5</i> , <i>PSMD13</i> , <i>PSMD4</i> , <i>PSMD8</i> , <i>SEM1</i>
CD Inflamed Samples (cDNA)					
RNA Degradation	KEGG	<i>Bifidobacterium longum</i> (Species)	***	6	<i>CNOT6L</i> , <i>CNOT7</i> , <i>DCP2</i> , <i>DDX6</i> , <i>EXOSC4</i> , <i>PAPOLA</i>
P53 Signalling Pathway	KEGG	<i>Bifidobacterium longum</i> (Species)	**	5	<i>CCND3</i> , <i>MDM4</i> , <i>PTEN</i> , <i>RRM2B</i> , <i>SESN3</i>
Cardiac Muscle Contraction	KEGG	Unclassified <i>Lachnospiraceae</i> (RSV)	***	5	<i>COX5A</i> , <i>COX5B</i> , <i>COX6B1</i> , <i>COX7A2</i> , <i>COX8A</i>
UC Inflamed Samples (gDNA)					
Integrin 1 Pathway	PID	<i>Parasutterella</i> (Genus)	****	8	<i>COL6A3</i> , <i>COL7A1</i> , <i>ITGA5</i> , <i>LAMA5</i> , <i>LAMC1</i> , <i>LAMC2</i> , <i>TGFBI</i> , <i>TNC</i>
Aurora B Pathway	PID	<i>Parabacteroides distans</i> (RSV)	****	6	<i>BUB1</i> , <i>KIF20A</i> , <i>KIF23</i> , <i>KIF2C</i> , <i>NCAPD2</i> , <i>NCAPH</i>
UC Inflamed Samples (cDNA)					
Inositol Phosphate Metabolism	KEGG	<i>Veillonellaceae</i> (Family)	**	5	<i>INPP5E</i> , <i>INPPL1</i> , <i>PIK3CB</i> , <i>PIP5K1C</i> , <i>PLCG1</i>
Phosphatidylinositol Signalling System	KEGG	<i>Veillonellaceae</i> (Family)	**	5	<i>INPP5E</i> , <i>INPPL1</i> , <i>PIK3CB</i> , <i>PIP5K1C</i> , <i>PLCG1</i>
Integrin 1 Pathway	PID	<i>Parasutterella</i> (Genus)	****	5	<i>COL7A1</i> , <i>ITGA5</i> , <i>LAMA5</i> , <i>LAMC2</i> , <i>TNC</i>

**** $q < 0.0001$; *** $q < 0.001$; ** $q < 0.01$; * $q < 0.1$.

differences in overall gene expression or methylation between these groups in both CD and UC (Supplementary Table 4 & 5). No DEGs met our significant threshold once we adjusted for multiple testing. When considering methylation data for inflamed and non-inflamed samples together, 3,584 CpGs were significantly associated with relapse status in CD and 590 CpGs in UC subjects ($q < 0.05$; Supplementary Table 10). However, once samples were split by inflammation status, no CpGs remained significantly associated with relapse, after adjusting for multiple testing.

Succinotypes of patients with IBD are associated with number of future relapses

Recent work has showed that individuals can be partitioned based on their gastrointestinal succinotype, i.e. the taxonomic identity of their dominant succinate-consuming bacterium, either *Dialister* (D) or *Phascolarctobacterium* (P).³⁷ Succinate can act as a pro-inflammatory signaling molecule, which when produced above a certain threshold in the colon can contribute to starting and/or maintaining an inflammatory signaling cascade.^{42,43} In patients with IBD, the slower succinate-consuming D-succinotypes have been reported as overrepresented compared to healthy controls, suggesting a potential contribution of succinate removal to pathophysiology.³⁷ We therefore wanted to assess the distribution of these dominant succinotypes in our patients with IBD and how they may relate to future relapse.

Here, we identified 7 RSVs in both 16S rRNA datasets that taxonomically classified as either P or D and verified that these RSVs mapped perfectly to known representatives of the respective genera (see Methods). The grouped relative abundances for both standing and active abundances P and D were indeed bimodally distributed with strong mutually exclusivity between the two (Figure 5a-b; $\rho_D = 0.892$; $\rho_P = 0.812$), consistent with the concept of succinotypes. Following Anthamatten et al. (2024),³⁷ we subsequently assigned succinotypes to samples based on the relative proportion of *Dialister*, r_D , defined as D counts divided by the

sum of D and P counts. A sample was assigned either D- or P-succinotype if $r_D > 0.9$ or $r_D < 0.1$, respectively, and there were at least 10 reads assigned to D and P. A substantial number of samples did not have any reads assigned to either D or P, though this proportion was notably higher for the biopsies (42%; 207/487) than for the stool samples (21%; 8/39) (Fisher's exact test, $p = 0.007$). We checked for the consistency in succinotype assignment for an individual by comparing across all respective samples (gDNA/cDNA, inflamed/non-inflamed). Only one single individual had discordant succinotype assignments across samples, where the fecal sample was a D-succinotype and the three biopsy samples were P-types. In nine other individuals, samples had non-zero counts of both D and P, but the remaining 103 individuals had consistent succinotype assignments across all samples (Figure 5c). We thus concluded that succinotypes were also well-defined for the samples used here and were able to assign succinotypes for 113 of the 175 subjects.

The distribution of succinotypes across disease types was marginally significantly different between IBD (CD+UC) and controls (Figure 5c; Fisher's exact test, $p = 0.062$). Once split into CD and UC, we found no significant difference in succinotypes between CD and controls ($p = 0.170$), but did notice a significant difference in distribution between UC and controls ($p = 0.046$). There was no difference in the proportion of subjects with and without future relapses between succinotypes, both in UC ($p = 0.784$) and CD ($p = 1$). We did, however, observe a trend for a higher number of relapses in CD patients with the D compared to the P succinotype (patients with at least one relapse, Mann-Whitney U test, $p = 0.080$), but not in UC ($p = 0.82$). To more carefully evaluate this trend, we fitted a zero-inflated Poisson model to the number of relapses (Figure 5d). The zero-inflation accounts for a certain probability of relapse during the observation window (4 years), and the Poisson distribution models that when patients do have a relapse, these relapses occur with a certain rate. We did not find any significant succinotype differences in terms of the probability of having a relapse. However, we did observe that

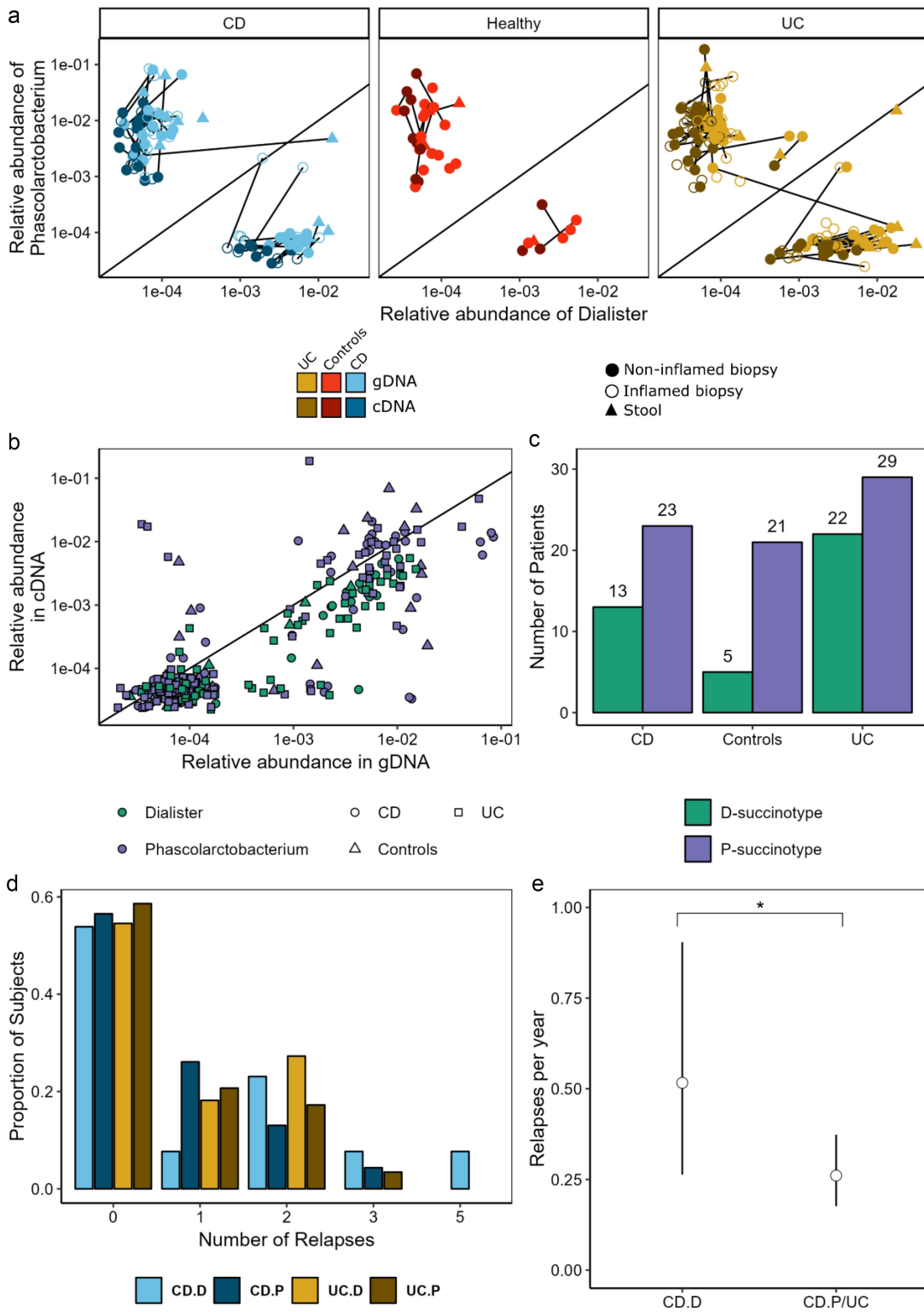


Figure 5. Succinotypes can be defined in CD, UC and control subjects. (a) Relative abundance of *Dialister* vs *Phascolarctobacterium* within samples for both standing and active datasets. (b) Comparison of the relative abundance of genera in gDNA and cDNA. (c) Bar plot of succinotypes grouped by disease-type. (d) Number of relapses by disease and succinotype. (e) Relapses per year with 95% confidence intervals for CD.D vs all other groups estimated using zero-inflated poisson model. * $p < 0.1$.

patients with CD who had the D-succinotype had a significantly higher relapse rate (mean 0.51 relapses/year; $p = 0.059$), while P-types and UC patients had lower relapse rates (mean 0.26 relapses/year; [Figure 5e](#)). This suggests that the D-succinotype is potentially associated with a higher frequency of relapses.

Prediction of relapse using machine learning

As it proved difficult to distinguish between future relapse and remission using traditional statistical analysis of single omics datasets (with the exception of associating succinotypes with number of relapses), we employed the ML method Extreme Gradient Boosting (XGBoost) to see if we could better predict these two groups using combinations of the omics datasets available. This powerful ML algorithm has already shown promising results in different areas of omics research due to its ability to handle different data types and missing data as well as its easy interpretation.^{5,11,44–46} In addition to the biopsy datasets used above, we included two datasets generated from the same samples that did not previously have sufficient power for single omics analysis.⁵ This resulted in three host omics data types (transcriptome, genotype, epigenome) and three microbiome data types (16S gDNA and cDNA genera and 16S gDNA G4 Phylochip (eOTUs)). The XGBoost models were trained on data from inflamed, non-inflamed and paired samples from UC and CD subjects. In each case, the analysis was performed as part of a cross-validation (CV) performance assessment routine, where an ensemble of 10 XGBoost models was used as a predictor (see Methods).

Models trained on inflamed CD samples had, in general, better performance than those trained on CD non-inflamed or paired samples ([Figure 6a](#); [Supplementary Table 11](#)). The highest Area Under the ROC Curve (AUC) was achieved when predicting relapse using both host RNA-Seq and 16S gDNA G4 Phylochip features (AUC = 0.84). We also observed promising performance from a model trained on host epigenome features combined with patient age (AUC = 0.81). In both cases, the datasets were generated from the inflamed mucosal samples of patients with CD. For models trained on either non-inflamed or paired data, the

highest AUCs achieved were 0.68 (Host RNA-Seq + 16S cDNA + 16S G4 Phylochip) and 0.72 (16S G4 Phylochip), respectively. Additional performance metrics for these models, including accuracy, sensitivity, specificity and F1-score, are provided in [Supplementary table 11](#).

Models also performed well in predicting relapse in UC when using either inflamed or paired data ([Figure 6b](#); [Supplementary Table 12](#)). The highest AUC for UC was achieved by combining features from host genotype and 16S gDNA G4 Phylochip datasets in conjunction with age when trained on paired inflamed and non-inflamed samples (AUC = 0.85). A model trained using multi-omics data from inflamed mucosal samples also showed high performance, achieving an AUC of 0.81 (host genotype, host epigenome and 16S cDNA datasets). When considering UC and CD patients together (IBD), we saw lower performing models than when each subtype was considered separately ([Supplementary Figure 11](#); [Supplementary Table 13](#)). Given the complexity of multi-omics datasets and lack of an external validation dataset, we assessed the generalizability and stability of our models by comparing the performance of our model on the validation set (inner loop of CV) and the test set (outer loop of CV), see [Supplementary Information 1](#).

To add an interpretative component, we elucidated which features were having the biggest role in the higher performing models. We therefore examined both feature importance values extracted from XGBoost models and SHapley Additive exPlanation (SHAP) values, which highlight the contribution of each feature on a prediction of the model ([Figure 6c-d](#)). Based on these results, the expression of *MTF1* and *RCAN1* genes were the most important features when predicting relapse in CD subjects based on inflamed sample data. For UC, the CpG site cg03256584 and the SNP rs11805303 of the pro-inflammatory gene *IL23R* were the most important when models were trained on inflamed and paired data, respectively. Also among the top 10 features for UC subjects were microbial features such as the active genera *Gordonibacter* and *Sellimonas* (inflamed model) and species such as *Bacteroides plebeius* and *Bacteroides*

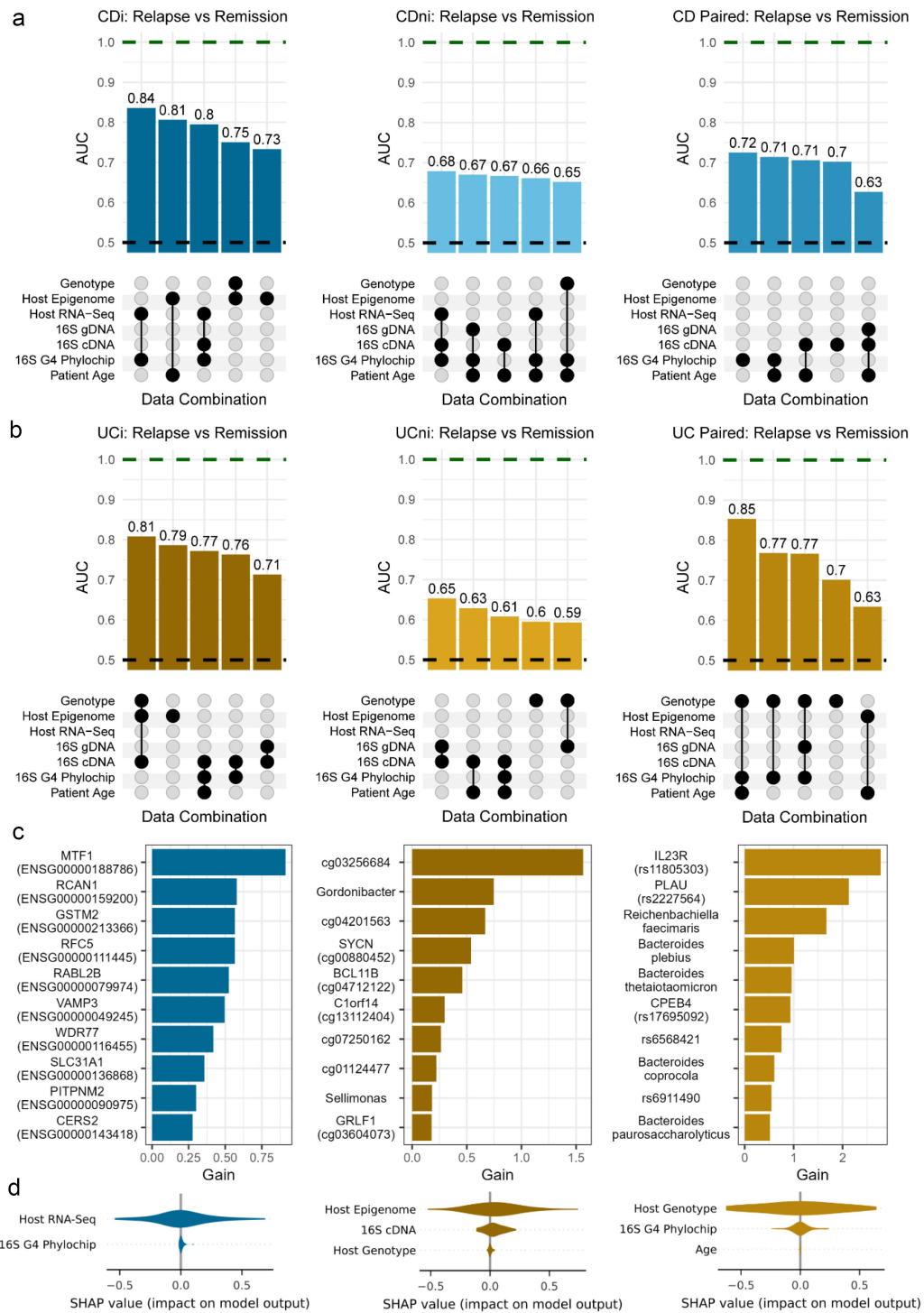


Figure 6. Outline of XGBoost model performance to predict relapse in (a) Crohn's disease (CD) and (b) ulcerative colitis (UC) subjects. UpSet plots outline AUCs and the combination of features used to achieve model performance. Top row shows top 5 models when predicting future relapse in patients with CD, where models were trained on inflamed, non-inflamed and paired data, respectively (left-right). Second row shows top 5 models when predicting future relapse in patients with UC. Green dashed line indicates perfect performance. Black dashed line is equivalent to a random model. (c) Top 10 important features based on gain importance metric from XGBoost. Features presented are those from the highest performing model when models were trained on CDi, UCi and UC paired samples, respectively (left-right). (d) SHAP values extracted from same models as c) but values were grouped (summed) by those omics types used to train the model.

thetaitaomicron (paired model), which have all been previously associated with IBD.^{47–51} In addition to considering feature importance, we calculated SHAP values for each model which we grouped by omics types. Here, grouped SHAP values indicated that host omics data had the largest impact on predictions for each of the top models outlined previously (Figure 6d).

Discussion

The potential causes of immune-mediated diseases such as CD and UC are widely considered multifactorial implicating both human genetics and the gut microbiota, thus requiring corresponding omics types to investigate relevant host-microbe interactions.^{1,52} In this study, we examined a comprehensive multi-omics dataset from mucosal biopsies to get further insight into IBD subtypes and the future disease states of these patients. We considered both single omics and integrative approaches to highlight differences in the disease types and their relapse states.

Our analysis showed differences in microbial composition between IBD subtypes and controls, which was consistent with existing literature.^{5,7,53} This was further expanded upon to include, not only the standing microbiota abundances, but also the metabolically active microbes. We observed similar trends for both 16S rRNA data types in terms of alpha and beta diversity, with CD patients having the greatest difference in composition compared to controls. These trends were more pronounced in the active microbiota, emphasizing their potential importance. Taxa such as a *Coprococcus* species, *Blautia obeum*, a *Subdoligranulum* species were consistently differentially abundant in disease relative to controls across both data types, and all of these taxa have been previously associated with IBD.^{5,54–56}

Simultaneously considering both 16S rRNA data types has only been done once previously in the context of IBD.⁷ The authors in that study ($n = 89$) reported a significant reduction in the ‘active’ alpha diversity of CD subjects compared to healthy, consistent with our findings. In another IBD multi-omics study, the authors compared 78 paired fecal metagenomes and metatranscriptomes,⁵⁷

highlighting more pronounced results in their active (metatranscriptomics) data, broadly consistent with our findings.

Our analysis of host omics data highlighted differences between IBD and control subjects with inflamed IBD samples showing the largest change in both expression and methylation. Enrichment analysis of DEGs showed many significant pathways many of which were common to both IBD disease-types, including fatty acid, amino acid, and carbohydrate metabolism, along with integrin and interleukin pathways (IL-23 and IL-12), all of which have well established associations with IBD.^{58–63} Our examination of methylation regions between inflamed and non-inflamed samples highlighted several DMRs significantly correlated with the corresponding gene expression. Methylation of the promoter region of the *GYPC* gene had a strong negative correlation with its gene expression in UC patients. This gene has previously been associated with response to corticosteroid therapy in pediatric UC patients.⁶⁴ Similarly, a strong positive correlation was found between the methylation and expression of *PLCE1* in UC, which was previously linked to this disease.⁶⁵ A subgroup of patients with UC was defined by the genes *SLC44A4*, *EPB41LAB* and *PLCE1* with patients in this subgroup reported to have milder clinical condition, but more likely to progress to colorectal cancer.⁶⁵ We also observed moderate correlations between methylation and gene expression for genes such as *AGAP1* and *PTPRN2* in CDi samples. To our knowledge, no direct associations between *AGAP1* and CD have previously been made, however, several studies have observed links between *PTPRN2* and CD.^{66,67} For example, studies have found *PTPRN2*, a gene which encodes the protein tyrosine phosphatase, to be differentially methylated in both adipose stem cells⁶⁶ and peripheral blood cells of patients with CD.⁶⁷

As both microbiome and host omics data was available for most of our cohort, we supplemented the single omics analysis by examining the interplay between individual host features and different taxonomic levels of gut microbes. We were able to identify associations between several microbes and various metabolic pathways for both CDi and UCi samples. Interestingly, the integrin beta-1 pathway was significantly enriched based on genes

associated with the abundance of the *Parasutterella* genus in UC patients for both the standing and active abundances. Genes from the integrin beta-1 pathway had previously been associated with taxa such as *Dialister*, *Phascolarctobacterium*, and *Intestinibacter* in subjects with IBD.¹⁰ *Dialister*, *Phascolarctobacterium* and *Parasutterella* all play a role in controlling the level of succinate present in the gut with the two former being known succinate consumers and the latter a succinate producer.^{68,69} Interestingly, when we examined the distribution of the dominant succinate consuming bacteria in our cohort, UC patients had a different distribution of succinotypes compared to controls while no difference was observed between CD and control subjects.

A common finding across our single omics analyses was that while we could identify differences between disease types, it proved more difficult to distinguish future relapse and remission. Only CDi samples (gDNA only) of those subjects that reported a relapse had a significantly different alpha diversity compared to CDi samples of those that remained in remission. No other differences in terms of alpha/beta diversity, taxa abundances or host gene expression were found when comparing the two outcome groups. Similarly, other investigators did not find significant microbial diversity differences between relapse and remission (stool metagenomics) in UC or CD subjects.⁵⁵

While most of our single omics analyses fell short of significantly distinguishing between future relapse and remission, we did observe that CD patients with a D-succinotype had a significantly higher number of relapses per year. Succinate has been implicated in IBD pathology, for example by perpetuating a pro-inflammatory state in macrophages⁷⁰ or contributing to the formation of fistulas.⁷¹ However, having a D-succinotype alone is not sufficient to cause disease as both D and P types are evenly distributed in healthy individuals.³⁷ *Dialister* consumes succinate more slowly compared to *Phascolarctobacterium*, leading to higher intestinal succinate concentrations in D-succinotype individuals compared to P ones. Thus, it is conceivable that other factors contribute to the onset of disease activity, which in turn is exacerbated by higher intestinal succinate concentrations resulting in flares. So not only is the slower

succinate-removing D-succinotype more common in IBD,³⁷ as is the increased abundance of *Dialister invisus* in general in IBD compared to non-IBD patients,⁸ but our findings also indicate predictive potential for future relapse frequency.

Encouragingly, it was possible to achieve good performance in predicting relapse by applying a ML approach to multiple omics datasets, in particular from inflamed samples. The highest performance was achieved when combining both host and microbial data, highlighting the host-microbial importance of any predictive profiles. While previous studies have attempted to predict relapse in IBD, the definition of relapse often differs and very few studies include more than one omics types in their analysis. Most studies applying ML to multi-omics data did so in order to classify disease.^{5,11,72,73} Sarrabayrouse and colleagues combined baseline microbiota and fungal loads from qPCR measurements of stool samples, inflammatory markers and flare history to predict relapse one year later for both CD and UC.¹³ It is however not surprising that clinical meta-data like previous flare history can significantly improve prediction of future relapse. Protein and metabolomics biomarkers in serum in another IBD study were associated with relapse within two years by using logistic regression.¹² Based on stool microbiota composition alone, an AUC of 0.67 was obtained when predicting onset of CD within 5 years in healthy first-degree relatives of patients with CD.⁶ None of these studies based their predictive models on integrated host-microbial molecular data from intestinal mucosa. However, the predictive performances achieved in our study appear comparable to those reported in existing literature with the performance of many models matching or surpassing those already published.

Although, we are seeing promising relapse-predicting results in terms of our succinotype and ML analyses, there are some limitations to our study. Firstly, our cohort consisted of only adult patients with IBD who were not newly diagnosed and therefore not treatment naive. As a result, our findings may not fully account for potential biases introduced by long-term illness and exposure to various treatments. In future studies it may be beneficial to examine treatment-naïve patients to assess if baseline features may be predictive of

future relapse or else collect extensive meta data on patient clinical features such as past and present medication. Secondly, our dataset consisted of single time-point data which was used to predict future relapse. A longitudinal approach may be more informative as it would be possible to follow the trends in each omics dataset across multiple time points and disease states.

While we had a large cohort of patients, not all data types were available for all samples and in many parts of our analysis only those samples with full coverage were considered. Consequently, our ML analysis was conducted using a nested cross validation approach and could not be validated as no suitable external dataset was available. While finding an equally comprehensive multi-omics study with future outcome data will be challenging, our findings should ideally be externally validated on new subjects when a prospectively recruited suitable validation cohort becomes available. Additionally, while we recognize that generating this type of multi-omics dataset may not be feasible for some researchers or in clinical settings due to cost and sample constraints, we hope our analysis will be helpful in guiding future studies by highlighting potentially more informative data types. Once validated, these models could then progress the development of prognostic tools in a clinical setting.

In conclusion, in contrast to a single omics approach, multi-omics analysis incorporating both host and microbiome data was predictive of clinical relapses with IBD. Future validation studies could next be designed with the best performing omics data in mind, which could eventually progress the development of prognostic tools.

Acknowledgments

We thank research nurses Margot Hurley, Mark Loughrey, Catherine O’Riordan and Mary Looby for their assistance in the clinic, and Silvia Melgar for critically reading the manuscript.

Disclosure statement

G.E.L. is an employee and shareholder of PharmaBiome AG and inventor on the patent application WO 2023/118460 A1 entitled “New biomarker for disorders and

diseases associated with intestinal dysbiosis”. M.J.C. is co-founder and Head of Bioinformatics of SeqBiome. T.Z.D. is a co-founder and was Vice-president of Second Genome; S.I., K.D., and E.R. were employees of Second Genome at the time of the analysis. F.S. is a co-founder of three campus companies: Alimentary Health Ltd, Tuscan Health Ltd (now names 4D Pharma Cork) and Alantia Food Clinical Trials.

Funding

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (grant numbers 11/SIRG/B2162, 18/CRT/6214 and SFI/12/RC/2273), the Irish Health Research Board (grant number PD/2009/30), the Irish Research Council (grant Number GOIPG/2017/1573), and Second Genome, Inc., South San Francisco, California, USA.

ORCID

Jill O’Sullivan  <http://orcid.org/0009-0003-7921-3155>

Shriram Patel  <http://orcid.org/0000-0002-5062-0289>

Gabriel E. Leventhal  <http://orcid.org/0000-0002-4463-166X>

Rachel S. Fitzgerald  <http://orcid.org/0000-0002-5569-6311>


Emilio J. Laserna-Mendieta  <http://orcid.org/0000-0002-9039-7667>

Aonghus Lavelle  <http://orcid.org/0000-0002-4461-8596>

Fergus Shanahan  <http://orcid.org/0000-0003-0467-0936>

Andriy Temko  <http://orcid.org/0000-0001-6548-0971>

Shoko Iwai  <http://orcid.org/0000-0002-6943-1912>

Marcus J. Claesson  <http://orcid.org/0000-0002-5712-0623>

Author contributions

M.J.C., F.S., T.Z.D., and S.I. designed and managed the project; J.O.S., S.P., G.E.L., R.S.F., E.J.L.M., C.E.H., N.K., E.R., K.D., A.T., S.I. performed the analyses; F.S., and A.L. provided clinical expertise; J.O.S., S.P., G.E.L., F.S. and M.J.C. wrote the paper; T.Z.D., S.I., A.T., A.L. reviewed and edited the manuscript. M.J.C. and F.S. secured funding.

Data availability statement

Sequence data and array data generated as part of the original study are available at NCBI BioProject (<https://www.ncbi.nlm.nih.gov/bioproject/>) PRJNA398187 and NCBI GEO (<https://www.ncbi.nlm.nih.gov/geo/>) GSE103027 and GSE105120. The corresponding metadata is available in Supplementary Table 1. All other datasets are available from the corresponding author upon reasonable request.

Ethics approval and consent to participate

The study was approved by the Cork hospital ethics committee and written informed consent was provided by all patients.

References

- Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, Andrews E, Ajami NJ, Bonham KS, Brislawn CJ, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*. 2019;569(7758):655–662. doi:10.1038/s41586-019-1237-9.
- Franzosa EA, Sirota-Madi A, Avila-Pacheco J, Fornelos N, Haiser HJ, Reinker S, Vatanen T, Hall AB, Mallick H, McIver LJ, et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat Microbiol*. 2019;4(2):293–305. doi:10.1038/s41564-018-0306-4.
- Lavelle A, Sokol H. Gut microbiota-derived metabolites as key actors in inflammatory bowel disease. *Nat Rev Gastroenterol Hepatol*. 2020;17(4):223–237. doi:10.1038/s41575-019-0258-z.
- Clooney AG, Eckenberger J, Laserna-Mendieta E, Sexton KA, Bernstein MT, Vagianos K, Sargent M, Ryan FJ, Moran C, Sheehan D, et al. Ranking microbiome variance in inflammatory bowel disease: a large longitudinal intercontinental study. *Gut*. 2021;70(3):499–510. doi:10.1136/gutjnl-2020-321106.
- Ryan FJ, Ahern AM, Fitzgerald RS, Laserna-Mendieta EJ, Power EM, Clooney AG, O'Donoghue KW, McMurdie PJ, Iwai S, Crits-Christoph A, et al. Colonic microbiota is associated with inflammation and host epigenomic alterations in inflammatory bowel disease. *Nat Commun*. 2020;11(1):1–12. doi:10.1038/s41467-020-15342-5.
- Raygoza Garay JA, Turpin W, Lee S-H, Smith MI, Goethel A, Griffiths AM, Moayyedi P, Espin-Garcia O, Abreu M, Aumais GL, et al. Gut microbiome composition is associated with future onset of Crohn's disease in healthy first-degree relatives. *Gastroenterology*. 2023;165(3):670–681. doi:10.1053/j.gastro.2023.05.032.
- Rehman A, Rausch P, Wang J, Skieceviciene J, Kiudelis G, Bhagalia K, Amarapurkar D, Kupcinskis L, Schreiber S, Rosenstiel P, et al. Geographical patterns of the standing and active human gut microbiome in health and IBD. *Gut*. 2016;65(2):238–248. doi:10.1136/gutjnl-2014-308341.
- Ravichandar JD, Rutherford E, Chow C-E, Han A, Yamamoto ML, Narayan N, Kaplan GG, Beck PL, Claesson MJ, Dabbagh K, et al. Strain level and comprehensive microbiome analysis in inflammatory bowel disease via multi-technology meta-analysis identifies key bacterial influencers of disease. *Front Microbiol*. 2022;13:961020. doi:10.3389/fmicb.2022.961020.
- Ananthakrishnan AN. Epidemiology and risk factors for IBD. *Nat Rev Gastroenterol Hepatol*. 2015;12(4):205–217. doi:10.1038/nrgastro.2015.34.
- Priya S, Burns MB, Ward T, Mars RAT, Adamowicz B, Lock EF, Kashyap PC, Knights D, Blekhman R. Identification of shared and disease-specific host gene–microbiome associations across human diseases using multi-omic integration. *Nat Microbiol*. 2022;7(6):780–795. doi:10.1038/s41564-022-01121-z.
- Hu S, Bourgonje AR, Gacesa R, Jansen BH, Björk JR, Bangma A, Hidding IJ, van Dullemen HM, Visschedijk MC, Faber KN, et al. Mucosal host-microbe interactions associate with clinical phenotypes in inflammatory bowel disease. *Nat Commun*. 2024;15(1):1470. doi:10.1038/s41467-024-45855-2.
- Borren NZ, Plichta D, Joshi AD, Bonilla G, Sadreyev R, Vlamakis H, Xavier RJ, Ananthakrishnan AN. Multi-“omics” profiling in patients with quiescent inflammatory bowel disease identifies biomarkers predicting relapse. *Inflamm Bowel Dis*. 2020;26(10):1524–1532. doi:10.1093/ibd/izaa183.
- Sarrabayrouse G, Elias A, Yáñez F, Mayorga L, Varela E, Bartoli C, Casellas F, Borrueal N, Herrera de Guise C, Machiels K, et al. Fungal and bacterial loads: noninvasive inflammatory bowel disease biomarkers for the clinical setting. *MSystems*. 2021;6(2). doi:10.1128/mSystems.01277-20.
- Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glockner FO. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res*. 2013;41(1):e1. doi:10.1093/nar/gks808.
- Andrews S. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinf. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
- Callahan BJ, Pj M, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: high-resolution sample inference from illumina amplicon data. *Nat Methods*. 2016;13(7):581–583. doi:10.1038/nmeth.3869.
- Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 2010;26(19):2460–2461. doi:10.1093/bioinformatics/btq461.
- Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*. 2013;41(database issue):d590–D596. doi:10.1093/nar/gks1219.
- Allard G, Ryan FJ, Jeffery IB, Claesson MJ. SPINGO: a rapid species-classifier for microbial amplicon sequences. *BMC Bioinf*. 2015;16(1):324. doi:10.1186/s12859-015-0747-1.
- Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. *Front Microbiol*. 2017;8(nov):2224. doi:10.3389/fmicb.2017.02224.

21. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*. 2014;30(15):2114–2120. doi:10.1093/bioinformatics/btu170.
22. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and hisat-genotype. *Nat Biotechnol*. 2019;37(8):907–915. doi:10.1038/s41587-019-0201-4.
23. Liao Y, Smyth GK, Shi W. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014;30(7):923–930. doi:10.1093/bioinformatics/btt656.
24. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, Beck S. A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450 k DNA methylation data. *Bioinformatics*. 2013;29(2):189–196. doi:10.1093/bioinformatics/bts680.
25. Mizejewski M, Schnauffer T, Muravsky M, Wang S, Caro-Aguilar I, Secore S, Thiriot DS, Hsu C, Rogers I, DeSantis T, et al. An in vitro culture model to study the dynamics of colonic microbiota in Syrian golden hamsters and their susceptibility to infection with *Clostridium difficile*. *ISME J*. 2015;9(2):321–332. doi:10.1038/ismej.2014.127.
26. West KA, Yin X, Rutherford EM, Wee B, Choi J, Chrisman BS, Dunlap KL, Hannibal RL, Hartono W, Lin M, et al. Multi-angle meta-analysis of the gut microbiome in autism spectrum disorder: a step toward understanding patient subgroups. *Sci Rep*. 2022;12(1):17034. doi:10.1038/s41598-022-21327-9.
27. DeSantis TZ, Cardona C, Narayan NR, Viswanatham S, Ravichandar D, Wee B, Chow C-E, Iwai S. StrainSelect: a novel microbiome reference database that disambiguates all bacterial strains, genome assemblies and extant cultures worldwide. *Heliyon*. 2023;9(2):e13314. doi:10.1016/j.heliyon.2023.e13314.
28. Dixon P. VEGAN, a package of R functions for community ecology. *J Veg Sci*. 2003;14(6):927–930. doi:10.1111/j.1654-1103.2003.tb02228.x.
29. Palarea-Albaladejo J, Martín-Fernández JA. Zcompositions — R package for multivariate imputation of left-censored data under a compositional approach. *Chemom And Intell Lab Syst*. 2015;143:85–96. doi:10.1016/j.chemolab.2015.02.019.
30. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, et al. Welcome to the Tidyverse. *J Open Source Softw*. 2019;4(43):1686. doi:10.21105/joss.01686.
31. Kassambara A. Rstatix: pipe-friendly framework for basic statistical tests. 2023. <https://cran.r-project.org/package=rstatix>.
32. Blighe K, Sharmila Rana ML. 2018 EnhancedVolcano: publication-ready volcano plots with enhanced colouring and labeling. <https://bioconductor.org/packages/devel/bioc/vignettes/EnhancedVolcano/inst/doc/EnhancedVolcano.html>.
33. Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer-Verlag; 2016. ISBN: 978-3-319-24277-4.
34. Fernandes AD, Reid JNS, Macklaim JM, Ta M, Edgell DR, Gloor GB. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*. 2014;2(1):15. doi:10.1186/2049-2618-2-15.
35. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. 2011;27(12):1739–1740. doi:10.1093/bioinformatics/btr260.
36. Martorell-Marugán J, González-Rumayor V, Carmona-Sáez P, Valencia A. MCSEA: detecting subtly differentially methylated regions. *Bioinformatics*. 2019;35(18):3257–3262. doi:10.1093/bioinformatics/btz096.
37. Anthamatten L, von Bieberstein Pr, Menzi C, Zünd JN, Lacroix C, de Wouters T, Leventhal GE, von Bieberstein PR. Stratification of human gut microbiomes by succinotype is associated with inflammatory bowel disease status. *Microbiome*. 2024;12(1):186. doi:10.1186/s40168-024-01897-8.
38. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery. New York, NY, USA; 2016. p. 785–794.
39. Ghosh TS, Das M, Jeffery IB, O'Toole PW. Adjusting for age improves identification of gut microbiome alterations in multiple diseases. *Elife*. 2020;9. doi:10.7554/eLife.50240.
40. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*; 2017. Red Hook (NY), USA: Curran Associates Inc.; p. 4768–4777.
41. Stankey CT, Bourges C, Haag LM, Turner-Stokes T, Piedade AP, Palmer-Jones C, Papa I, Silva dos Santos M, Zhang Q, Cameron AJ, et al. A disease-associated gene desert directs macrophage inflammation through ETS2. *Nature*. 2024;630(8016):447–456. doi:10.1038/s41586-024-07501-1.
42. Macias-Ceja DC, Ortiz-Masiá D, Salvador P, Gisbert-Ferrándiz L, Hernández C, Hausmann M, Rogler G, V EJ, Hinojosa J, Alós R, et al. Succinate receptor mediates intestinal inflammation and fibrosis. *Mucosal Immunol*. 2019;12(1):178–187. doi:10.1038/s41385-018-0087-3.
43. Connors J, Dawe N, Van Limbergen J. The role of succinate in the regulation of intestinal inflammation. *Nutrients*. 2018;11(1):25. doi:10.3390/nu11010025.

44. Ma B, Meng F, Yan G, Yan H, Chai B, Song F. Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. *Comput Biol Med.* 2020;121:103761. doi:10.1016/j.compbimed.2020.103761.
45. Hou X, Ma B, Liu M, Zhao Y, Chai B, Pan J, Wang P, Li D, Liu S, Song F. The transcriptional risk scores for kidney renal clear cell carcinoma using XGBoost and multiple omics data. *Math Biosci Eng.* 2023;20(7):11676–11687. doi:10.3934/mbe.2023519.
46. Volkova A, Ruggles KV. Predictive metagenomic analysis of autoimmune disease identifies robust autoimmunity and disease specific microbial signatures. *Front Microbiol.* 2021;12:12. doi:10.3389/fmicb.2021.621310.
47. Dahal RH, Kim S, Kim YK, Kim ES, Kim J. Insight into gut dysbiosis of patients with inflammatory bowel disease and ischemic colitis. *Front Microbiol.* 2023;14:14. doi:10.3389/fmicb.2023.1174832.
48. Mondot S, Lepage P, Seksik P, Allez M, Tréton X, Bouhnik Y, Colombel JF, Leclerc M, Pochart P, Doré J, et al. Structural robustness of the gut mucosal microbiota is associated with Crohn's disease remission after surgery. *Gut.* 2016;65(6):954–962. doi:10.1136/gutjnl-2015-309184.
49. Durant L, Stentz R, Noble A, Brooks J, Gicheva N, Reddi D, O'Connor MJ, Hoyles L, McCartney AL, Man R, et al. Bacteroides thetaiotaomicron-derived outer membrane vesicles promote regulatory dendritic cell responses in health but not in inflammatory bowel disease. *Microbiome.* 2020;8(1):88. doi:10.1186/s40168-020-00868-z.
50. García-Villalba R, Giménez-Bastida JA, Cortés-Martín A, MÁ Á-G, Tomás-Barberán FA, Selma MV, Espín JC, González-Sarrías A. Urolithins: a comprehensive update on their metabolism, bioactivity, and associated gut microbiota. *Mol Nutr Food Res.* 2022;66(21):2101019. doi:10.1002/mnfr.202101019.
51. Quagliariello A, Del Chierico F, Reddel S, Russo A, Onetti Muda A, D'Argenio P, Angelino G, Romeo EF, Dall'oglio L, De Angelis P, et al. Fecal microbiota transplant in two ulcerative colitis pediatric cases: Gut microbiota and clinical course correlations. *Microorganisms.* 2020;8(10):1486. doi:10.3390/microorganisms8101486.
52. Agrawal M, Allin KH, Petralia F, Colombel J-F, Jess T. Multiomics to elucidate inflammatory bowel disease risk factors and pathways. *Nat Rev Gastroenterol Hepatol.* 2022;19(6):399–409. doi:10.1038/s41575-022-00593-y.
53. Halfvarson J, Brislawn CJ, Lamendella R, Vázquez-Baeza Y, Walters WA, Bramer LM, D'Amato M, Bonfiglio F, McDonald D, Gonzalez A, et al. Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat Microbiol.* 2017;2(5). doi:10.1038/nmicrobiol.2017.4.
54. Zhu S, Han M, Liu S, Fan L, Shi H, Li P. Composition and diverse differences of intestinal microbiota in ulcerative colitis patients. *Front Cell Infect Microbiol.* 2022;12:12. doi:10.3389/fcimb.2022.953962.
55. Serrano-Gómez G, Mayorga L, Oyarzun I, Roca J, Borrueal N, Casellas F, Varela E, Pozuelo M, Machiels K, Guarner F, et al. Dysbiosis and relapse-related microbiome in inflammatory bowel disease: a shotgun metagenomic approach. *Comput Struct Biotechnol J.* 2021;19:6481–6489. doi:10.1016/j.csbj.2021.11.037.
56. Qiu Z, Yang H, Rong L, Ding W, Chen J, Zhong L. Targeted metagenome based analyses show gut microbial diversity of inflammatory bowel disease patients. *Indian J Microbiol.* 2017;57(3):307–315. doi:10.1007/s12088-017-0652-6.
57. Schirmer M, Franzosa EA, Lloyd-Price J, McIver LJ, Schwager R, Poon TW, Ananthakrishnan AN, Andrews E, Barron G, Lake K, et al. Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. *Nat Microbiol.* 2018;3(3):337–346. doi:10.1038/s41564-017-0089-z.
58. Heimerl S, Moehle C, Zahn A, Boettcher A, Stremmel W, Langmann T, Schmitz G. Alterations in intestinal fatty acid metabolism in inflammatory bowel disease. *Biochim Et Biophys Acta (BBA) - Mol Basis Of Disease.* 2006;1762(3):341–350. doi:10.1016/j.bbdis.2005.12.006.
59. Dotan I, Allez M, Danese S, Keir M, Tole S, McBride J. The role of integrins in the pathogenesis of inflammatory bowel disease: approved and investigational anti-integrin therapies. *Med Res Rev.* 2020;40(1):245–262. doi:10.1002/med.21601.
60. Gubatan J, Keyashian K, Rubin SJS, Wang J, Buckman CA, Sinha S. Anti-integrins for the treatment of inflammatory bowel disease: current evidence and perspectives. *Clin Exp Gastroenterol.* 2021;14:333–342. doi:10.2147/CEG.S293272.
61. Verstockt B, Salas A, Sands BE, Abraham C, Leibovitch H, Neurath MF, Vande Casteele N, Danese S, D'Haens G, Eckmann L, et al. IL-12 and IL-23 pathway inhibition in inflammatory bowel disease. *Nat Rev Gastroenterol Hepatol.* 2023;20(7):433–446. doi:10.1038/s41575-023-00768-1.
62. Scoville EA, Allaman MM, Brown CT, Motley AK, Horst SN, Williams CS, Koyama T, Zhao Z, Adams DW, Beaulieu DB, et al. Alterations in lipid, amino acid, and energy metabolism distinguish Crohn's disease from ulcerative colitis and control subjects by serum metabolomic profiling. *Metabolomics.* 2018;14(1):17. doi:10.1007/s11306-017-1311-y.

63. Schmitt H, Neurath MF, Atreya R. Role of the IL23/IL17 pathway in Crohn's disease. *Front Immunol.* 2021;12:622934. doi:10.3389/fimmu.2021.622934.
64. Kabakchiev B, Turner D, Hyams J, Mack D, Leleiko N, Crandall W, Markowitz J, Otley AR, Xu W, Hu P, et al. Gene expression changes associated with resistance to intravenous corticosteroid therapy in children with severe ulcerative colitis. *PLoS One.* 2010;5(9):e13085. doi:10.1371/journal.pone.0013085.
65. Ma J-L, Zhang H-J, Zhang C-F, Zhang Y-Y, Wang G-M. Construction of molecular subgroups of ulcerative colitis. *Eur Rev Med Pharmacol Sci.* 2023;27(19):9333–9345. doi:10.26355/eurev_202310_33961.
66. Serena C, Millan M, Ejarque M, Saera-Vila A, Maymó-Masip E, Núñez-Roa C, Monfort-Ferré D, Terrón-Puig M, Bautista M, Menacho M, et al. Adipose stem cells from patients with Crohn's disease show a distinctive DNA methylation pattern. *Clin Epigenet.* 2020;12(1):53. doi:10.1186/s13148-020-00843-3.
67. Li Yim A YF, Duijvis NW, Zhao J, de Jonge WJ, D'Haens GRAM, Gram D, Mmam M, Te Velde Aa, Anpm P, de Jonge WJ, et al. Peripheral blood methylation profiling of female Crohn's disease patients. *Clin Epigenet.* 2016;8(1):65. doi:10.1186/s13148-016-0230-5.
68. Fernández-Veledo S, Vendrell J. Gut microbiota-derived succinate: friend or foe in human metabolic diseases? *Rev Endocr Metab Disord.* 2019;20(4):439–447. doi:10.1007/s11154-019-09513-z.
69. Ju T, Kong JY, Stothard P, Willing BP. Defining the role of *Parasutterella*, a previously uncharacterized member of the core gut microbiota. *Isme J.* 2019;13(6):1520–1534. doi:10.1038/s41396-019-0364-5.
70. Fremder M, Kim SW, Khamaysi A, Shimshilashvili L, Eini-Rider H, Park IS, Hadad U, Cheon JH, Ohana E. A transepithelial pathway delivers succinate to macrophages, thus perpetuating their pro-inflammatory metabolic state. *Cell Rep.* 2021;36(6):109521. doi:10.1016/j.celrep.2021.109521.
71. Ortiz-Masiá D, Gisbert-Ferrándiz L, Bauset C, Coll S, Mamie C, Scharl M, V EJ, Alós R, Navarro F, Cosín-Roger J, et al. Succinate activates EMT in intestinal epithelial cells through SUCNR1: a novel protagonist in fistula development. *Cells.* 2020;9(5):1104. doi:10.3390/cells9051104.
72. Jiang Z, Li J, Kong N, Kim J-H, Kim B-S, Lee M-J, Park YM, Lee S-Y, Hong S-J, Sul JH. Accurate diagnosis of atopic dermatitis by combining transcriptome and microbiota data with supervised machine learning. *Sci Rep.* 2022;12(1):290. doi:10.1038/s41598-021-04373-7.
73. Jacobs JP, Lagishetty V, Hauer MC, Labus JS, Dong TS, Toma R, Vuyisich M, Naliboff BD, Lackner JM, Gupta A, et al. Multi-omics profiles of the intestinal microbiome in irritable bowel syndrome and its bowel habit subtypes. *Microbiome.* 2023;11(1):5. doi:10.1186/s40168-022-01450-5.