













OPEN ACCESS

Original research

# Faecal metabolome and its determinants in inflammatory bowel disease

Arnau Vich Vila <sup>1,2</sup> Shixian Hu <sup>1,2</sup> Sergio Andreu-Sánchez <sup>2,3</sup>  
Valerie Collij <sup>1,2</sup> Bernadien H Jansen,<sup>1</sup> Hannah E Augustijn <sup>2</sup> Laura A Bolte <sup>1</sup>  
Renate A A Ruigrok,<sup>1,2</sup> Galeb Abu-Ali,<sup>4</sup> Cosmas Giallourakis,<sup>4</sup> Jessica Schneider,<sup>4</sup>  
John Parkinson,<sup>4</sup> Amal Al-Garawi,<sup>4</sup> Alexandra Zhernakova <sup>2</sup> Ranko Gacesa <sup>1,2</sup>  
Jingyuan Fu <sup>2,3</sup> Rinse K Weersma <sup>1</sup>

► Additional supplemental material is published online only. To view, please visit the journal online (<http://dx.doi.org/10.1136/gutjnl-2022-328048>).

<sup>1</sup>Department of Genetics, University Medical Centre, Groningen, The Netherlands

<sup>2</sup>Department of Pediatrics, University Medical Centre, Groningen, The Netherlands

<sup>3</sup>Department of Gastroenterology and Hepatology, University Medical Centre, Groningen, The Netherlands

<sup>4</sup>Gastroenterology Drug Discovery Unit, Takeda Pharmaceutical, Cambridge, Massachusetts, USA

## Correspondence to

Prof. Dr. Rinse K Weersma, Department of Gastroenterology and Hepatology, University Medical Centre Groningen, Groningen, 9713 GZ, The Netherlands; [r.k.weersma@umcg.nl](mailto:r.k.weersma@umcg.nl)  
Dr Arnau Vich Vila; [arnauvich@gmail.com](mailto:arnauvich@gmail.com)

SH, SA-S and VC contributed equally.

Received 14 June 2022

Accepted 5 March 2023

Published Online First

23 March 2023



© Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

**To cite:** Vich Vila A, Hu S, Andreu-Sánchez S, et al. *Gut* 2023;**72**:1472–1485.

## ABSTRACT

**Objective** Inflammatory bowel disease (IBD) is a multifactorial immune-mediated inflammatory disease of the intestine, comprising Crohn's disease and ulcerative colitis. By characterising metabolites in faeces, combined with faecal metagenomics, host genetics and clinical characteristics, we aimed to unravel metabolic alterations in IBD.

**Design** We measured 1684 different faecal metabolites and 8 short-chain and branched-chain fatty acids in stool samples of 424 patients with IBD and 255 non-IBD controls. Regression analyses were used to compare concentrations of metabolites between cases and controls and determine the relationship between metabolites and each participant's lifestyle, clinical characteristics and gut microbiota composition. Moreover, genome-wide association analysis was conducted on faecal metabolite levels.

**Results** We identified over 300 molecules that were differentially abundant in the faeces of patients with IBD. The ratio between a sphingolipid and L-urobilin could discriminate between IBD and non-IBD samples (AUC=0.85). We found changes in the bile acid pool in patients with dysbiotic microbial communities and a strong association between faecal metabolome and gut microbiota. For example, the abundance of *Ruminococcus gnavus* was positively associated with tryptamine levels. In addition, we found 158 associations between metabolites and dietary patterns, and polymorphisms near *NAT2* strongly associated with coffee metabolism.

**Conclusion** In this large-scale analysis, we identified alterations in the metabolome of patients with IBD that are independent of commonly overlooked confounders such as diet and surgical history. Considering the influence of the microbiome on faecal metabolites, our results pave the way for future interventions targeting intestinal inflammation.

## INTRODUCTION

Characterisation of the host–microbiota symbiosis is crucial in the context of intestinal disorders such as inflammatory bowel disease (IBD) in which the gut environment is severely perturbed, yet the disease-causing mechanisms are still largely unknown. IBD is a chronic inflammatory

## WHAT IS ALREADY KNOWN ON THIS TOPIC

- ⇒ The gut microbiome is increasingly recognised as a metabolic organ that affects host health.
- ⇒ Accumulating evidence suggests that gut microbiota-derived metabolites are critical mediators between microbiota and immune response.
- ⇒ While dysbiosis in gut communities has been extensively explored in inflammatory bowel disease (IBD), unravelling the mechanisms behind host–microbiota interactions remains complex and requires insights into the metabolic activity in the gut.

## WHAT THIS STUDY ADDS

- ⇒ Gut metabolism in patients with IBD is characterised by lower levels of molecules derived from saccharolytic fermentation and increased metabolites from proteolytic fermentation.
- ⇒ Gut microbiota composition is the main determinant of faecal metabolite content, as compared with host lifestyle, genetic and clinical phenotypes.
- ⇒ In the gut of patients with IBD, the expansion of pathobionts co-occurs with an increased concentration of sphingolipids, ethanolamine and primary bile acids.
- ⇒ Intestinal resections have a long-lasting effect on intestinal bile acid and lipid metabolism.

## HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

- ⇒ Future studies investigating host metabolism should account for microbial composition and history of intestinal surgeries.
- ⇒ The faecal metabolome offers new avenues for identifying biomarkers for IBD and other immune-mediated inflammatory diseases.
- ⇒ Understanding the gut microbiota's contribution to human metabolism in health and diseases is essential for designing future dietary interventions.

disorder of the gastrointestinal tract that consists of two main subtypes: ulcerative colitis (UC) and Crohn's disease (CD).<sup>1,2</sup> In IBD, periods of active

disease are characterised by loss of strictly anaerobic bacteria, blooming of facultative anaerobes and alterations in the chemical environment in the gut.<sup>3</sup> For example, reductions of gut barrier-protecting short-chain fatty acids (SCFA) and alterations in bile acids, sphingolipids and tryptophan-derived metabolites have been consistently reported in faeces of patients with IBD.<sup>4,5</sup> However, a large number of molecules in the human body remain uncharacterised, and thus their implications for human health remain unknown. Considering that a subset of small molecules, including microbiome-derived metabolites, have been shown to regulate the immune response, it is crucial to characterise these metabolites and understand which factors determine their concentrations in the gut.

Recent technological advances in mass spectrometry techniques have enabled high-throughput characterisation and quantification of a wide range of known and chemically unannotated molecules.<sup>6</sup> In this context, the characterisation of faecal metabolites holds great potential for discovering non-invasive biomarkers and therapeutic targets. To date, however, studies performing untargeted metabolomics on the faeces of patients with IBD have been scarce, limited in sample size and lacking in-depth information on host genetics, lifestyle, diet and clinical characteristics.<sup>4,7</sup>

In this study, we aimed to determine alterations in the gut metabolism of patients with IBD to pinpoint factors influencing faecal metabolite levels. Our findings highlight the potential of faecal metabolites as biomarkers for IBD and show that, despite the influence of lifestyle, genetics and disease, faecal microbes are a strong predictor of the levels and composition of metabolites in the gut.

## METHODS

### Cohort and metadata description

Samples were obtained from two established cohorts: LifeLines,<sup>8</sup> a population biobank from the north of the Netherlands, and 1000IBD,<sup>9</sup> a cohort of patients with IBD from the University Medical Centre of Groningen. In this study, we included 255 non-IBD controls, 238 patients with CD and 174 patients with UC. Sample collection and storage are described in online supplemental methods and cohort characteristics are summarised in online supplemental table 1 and table 1.<sup>10,11</sup>

### Metabolite quantification

Metabolomics measurements performed by Metabolon (North Carolina, USA) detected 1684 faecal metabolites (online supplemental table 2). The concentrations of eight short-chain and branched-chain fatty acids were measured using liquid chromatography with tandem mass spectrometry (online supplemental table 3).

### Metabolic data processing

Metabolomic data were handled as a compositional dataset and transformed using centered log-ratios. Metabolites were split into three categories based on their prevalence. The first group consisted of metabolites present in more than 70% of the samples in both the cases and controls (x=854). Missing values were imputed using k-Nearest Neighbour Imputation.<sup>12</sup> We set the number of nearest neighbours to 10 (k=10) for the imputation and Euclidean distance as a metric. The second group of metabolites (prevalence <70% and >20%, x=514) were transformed into binary traits (metabolite presence/absence). Rare metabolites (prevalence <20%, x=316) were excluded from analyses.

**Table 1** Cohort description

Phenotypes	Control	IBD	P value
	<b>255</b>	<b>424</b>	
Age (mean (SD))	46.83 (12.08)	42.61 (14.69)	<0.001
Body mass index (mean (SD))	25.17 (3.84)	25.44 (5.02)	0.462
Sex=male (%)	114 (44.7)	170 (40.1)	0.272
Diagnosis (%)			
CD	NA	238 (56.1)	
IBDU	NA	12 (2.8)	
UC	NA	174 (41.0)	
Disease location (%)			
Both	NA	97 (22.8)	
Colon	NA	228 (53.7)	
Ileum	NA	69 (16.2)	
Disease activity (%)			
Active	NA	107 (25.2)	
Not active	NA	311 (73.3)	
Estimated bowel movement a day (mean (SD))	1.33 (0.66)	2.73 (2.37)	<0.001
Smoker=yes (%)	38 (14.9)	87 (20.5)	0.084
Faecal calprotectin >200 = yes (%)	8 (3.2)	168 (45.0)	<0.001
Montreal B (%)			
B1	NA	129 (54.2)	
B2	NA	76 (31.9)	
B3	NA	28 (11.7)	
Montreal B perianal (%)			
No	NA	326 (76.8)	
Yes	NA	68 (16.0)	
Montreal S (%)			
S0	NA	7 (4.0)	
S1	NA	56 (32.1)	
S2	NA	69 (39.6)	
S3	NA	27 (15.5)	
Montreal A (%)			
A1	NA	54 (12.7)	
A2	NA	245 (57.7)	
A3	NA	85 (20.0)	
Montreal L (%)			
L1	NA	63 (14.8)	
L2	NA	203 (47.8)	
L3	NA	94 (22.1)	
L4	NA	21 (4.9)	
Montreal E (%)			
E1	NA	21 (12.0)	
E2	NA	57 (32.7)	
E3	NA	80 (45.9)	
Microbiome Shannon Index (mean (SD))	3.05 (0.33)	2.89 (0.41)	<0.001

CD, Crohn's disease; IBD, inflammatory bowel disease; IBDU, inflammatory bowel disease unclassified; NA, not available; UC, ulcerative colitis.

### Identification of metabolites associated with IBD

We performed linear regression analysis using the *lm* function in R. The abundance of each metabolite was compared between disease groups (IBD/CD/UC) and controls. Technical factors (storage time, input grams of faeces and sample batch), host characteristics (age, sex, body mass index (BMI) and bowel movements per day), intestinal integrity (any resection: yes/no) and 24 dietary patterns that were significantly different between cases and controls were included as covariates in the regression

models (online supplemental table 4). Less prevalent metabolites (prevalence <70% and >20% of samples) were evaluated using logistic regressions. All p values were adjusted for multiple testing using Benjamini-Hochberg. A false discovery rate (FDR) <0.05 was used as the threshold for statistical significance.

### Prediction of IBD based on metabolomics profiles

We used CoDaCoRe<sup>13</sup> (V.0.0.1) to identify ratios of metabolites and bacterial abundances that could predict IBD and its subphenotypes (see online supplemental methods).

### Association between metabolites and phenotypes

An association analysis between phenotypes and metabolites was performed within each cohort (controls, CD and UC). Each phenotype–metabolite combination was tested using linear regression, including age, sex, BMI, bowel movements per day and technical factors as covariates.

The results of the metabolite–phenotype analyses were combined in a meta-analysis using random-effects models implemented in R package *meta* (V.4.8). Results were considered statistically significant when the meta-FDR <0.05 (online supplemental methods).

### Genome-wide association analysis on faecal metabolites

Exome sequencing and genomic array data were available for both cohorts (see online supplemental methods). Linear regression was used for metabolites present in >70% of the samples and logistic regression for those present in between 70% and 20% of the samples. Analyses were performed per cohort, and results were combined in a meta-analysis, as previously described.<sup>14</sup> In addition to accounting for the confounders described above (see the Identification of metabolites associated with IBD section), we included population genetic structure as a covariate in the analysis. To determine the statistical significance of our findings, we adopted two thresholds: a genome-wide significance ( $p < 5e^{-08}$ ), and a more conservative cut-off, a study-wide significance ( $p < 2.97e^{-11}$ ). The study-wide significance threshold was determined by dividing the genome-wide threshold by the total number of metabolites ( $5.0e^{-08}/1684$ ).

### Co-occurrence patterns between bacteria and metabolites

The *QIIME*<sup>15</sup> implementation of *mmvec* V.1.0.6<sup>16</sup> was used to estimate the co-occurrence probabilities between highly prevalent metabolites and bacteria. Furthermore, we assessed the associations between individual microbiome features (taxa, gene clusters and metabolic pathways) and metabolites using regression models considering the interaction between bacteria and dysbiosis (online supplemental methods).

### Differential abundance analyses of faecal microbiome features

Linear regression analysis was used to identify microbiome features that differed between controls and IBD. Age, sex, BMI, average bowel movements per day, history of intestinal resections (yes/no) and raw sequencing read depth were included as covariates in the regression models (online supplemental methods).

### Association between dysbiosis and faecal metabolites

Phenotypic differences between patients with dysbiotic and eubiotic microbiota were established using  $\chi^2$  or Wilcoxon-rank test for categorical and continuous variables, respectively. Differences in the abundance of faecal metabolites between the two groups of patients were tested using linear regression. Age, sex,

BMI, intestinal resection (yes/no), ileocecal valve in situ (yes/no), average bowel movements per day and differences in 12 dietary patterns (online supplemental table 5) were added as covariates in the regression models.

### Metabolite-level prediction

To predict the levels of metabolites in faeces, we performed regression models with L1 regularisation (lasso) using the *glmnet* R package<sup>17</sup> (see online supplemental methods).

## RESULTS

### Patients with IBD have a distinct faecal metabolite profile

Metabolites were assessed in the faecal samples from 238 patients with CD, 174 patients with UC and 255 non-IBD controls (table 1). On average, 1011 metabolites were detected per sample, ranging from 784 to 1241 molecules.

Principal coordinate analysis (PCoA) based on metabolites levels showed that IBD samples are dispersed across a cluster that partially overlaps with controls (figure 1A). The first component of the PCoA captured 18% of the variation and was driven by the levels of carnitine and bile and fatty acids, while the second component, representing 8% of cohort variation, was driven by the abundance of dipeptides and several unclassified metabolites (online supplemental table 6, figure 1B–D).

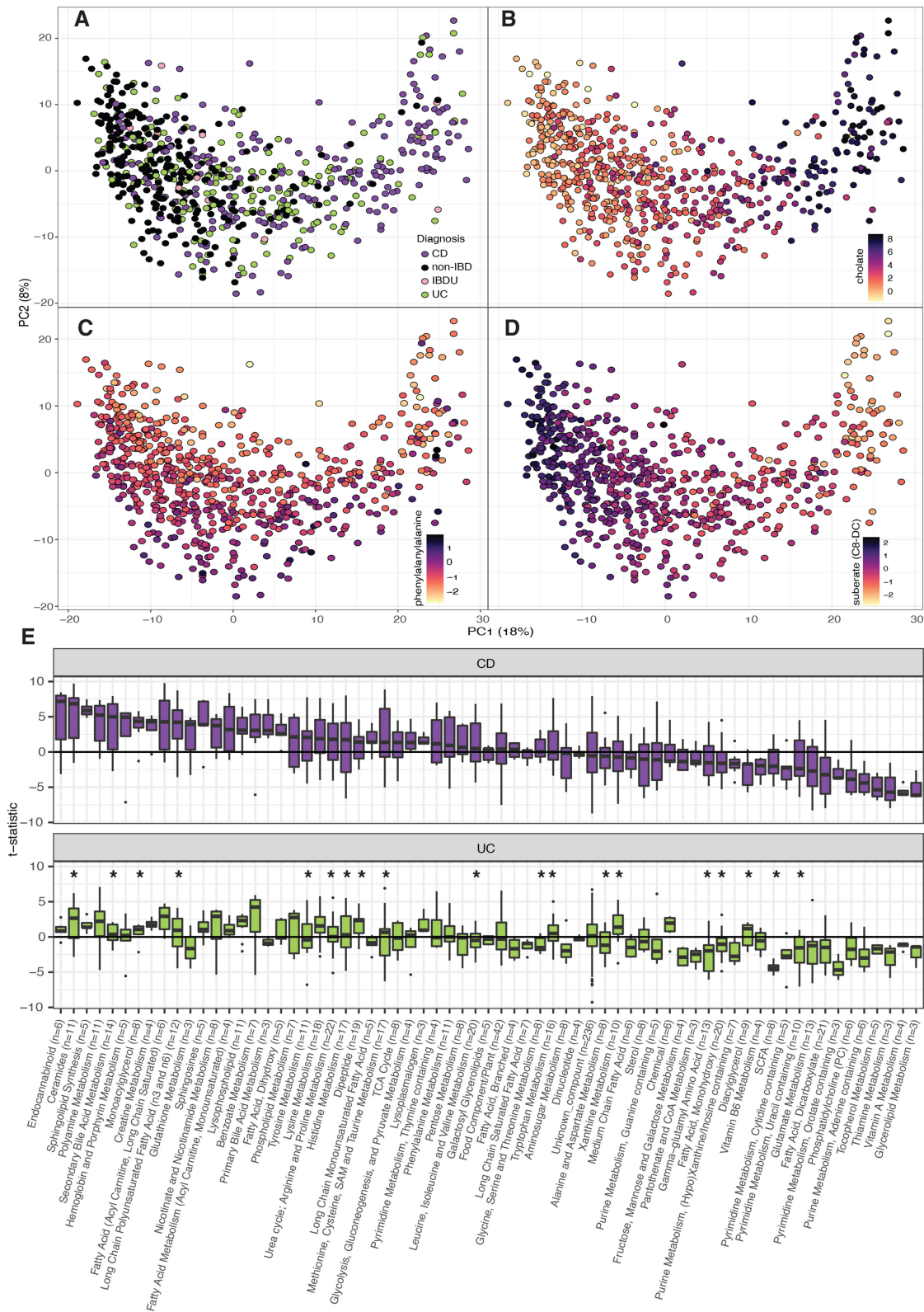
Differential abundance analysis revealed 324 associations when comparing patients with CD to controls and 308 associations when comparing patients with UC to controls (FDR <0.05) (online supplemental table 7, online supplemental figure 1A–E). Moreover, when looking into lower prevalence metabolites, we found that products of the metabolism of bile acids, ceramides and steroids were more prevalent in faeces of patients with IBD than in controls (182 and 119 molecules associated with CD and UC, respectively, online supplemental table 7).

A prominent signal in both disease groups was the depletion of vitamins and fatty acid-related molecules compared with controls (figure 1E). Patients with IBD presented higher levels of the phenolic compound p-cresol sulphate. The level of indole-propionic acid was decreased in UC, while tryptamine and kynurenine were increased in both CD and UC (FDR <0.05, online supplemental figure 1D). Patients with IBD also showed higher levels of arachidonic acid (20:4n6) and a lower ratio of omega-6/omega-3 fatty acids (online supplemental table 8, online supplemental figure 1E).

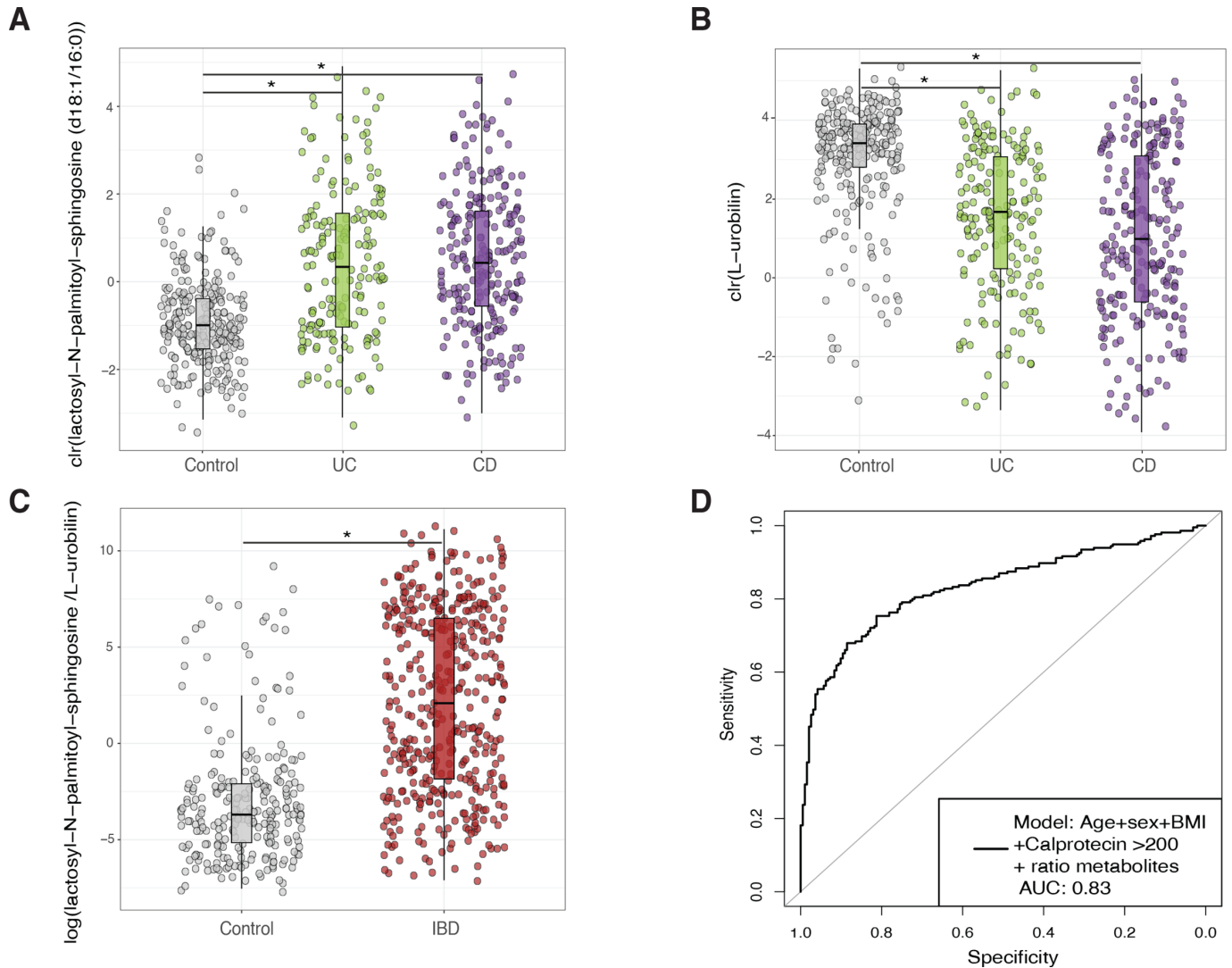
We also found that 106 metabolites were differentially abundant between CD and UC. For example, patients with UC presented higher levels of diaminopimelate (DAPA), an alpha-amino acid present in the cell membrane of gram-negative bacteria. Interestingly, DAPA-containing peptidoglycans can trigger the immune response mediated by *NOD1*<sup>18</sup> (online supplemental figure 1C, online supplemental table 9).

### Patients with UC show the lowest concentrations of SCFAs in faeces

The concentrations of SCFAs are essential for immune modulation, and their synthesis is dependent on colonic bacterial fermentation of polysaccharides.<sup>19</sup> After correcting for potential confounding effects, including anthropometric measurements, batch and sample storage time (see online supplemental methods, online supplemental table 4), acetate, propionate and butyrate were found in lower concentrations in patients with UC when compared to controls (FDR<sub>UC</sub> <0.05). No significant differences in these metabolites were observed between CD and controls. In



**Figure 1** Faecal metabolite alterations in patients with Crohn's disease and ulcerative colitis. (A–D). Principal coordinate analyses depicting the clustering of 255 non-IBD (black), 238 CD (purple), 174 UC (green) and 12 IBDU (pink) samples according to their metabolomic composition. The first principal component is mainly driven by the levels of cholic acid and suberate (B, D) and the second component by the concentrations of phenylalanylalanine (panel C). Light–dark colour gradient represents low–high metabolite values. Metabolite concentrations are expressed as centred log-ratio (clr) of the AUC raw values. (E). Metabolite differences between cases and controls grouped into metabolomic pathways. For clarity, only categories with three or more metabolites are shown (number of metabolites per categories are indicated on the x-axis). The y-axis represents the t-statistic value from the linear regression model (see online supplemental methods). Asterisks indicate significant differences between CD and UC (FDR<0.05, online supplemental tables 7–9). AUC, area under the curve; CD, Crohn's disease; FDR, false discovery rate; IBD, inflammatory bowel disease; IBDU, inflammatory bowel disease unclassified; UC, ulcerative colitis.



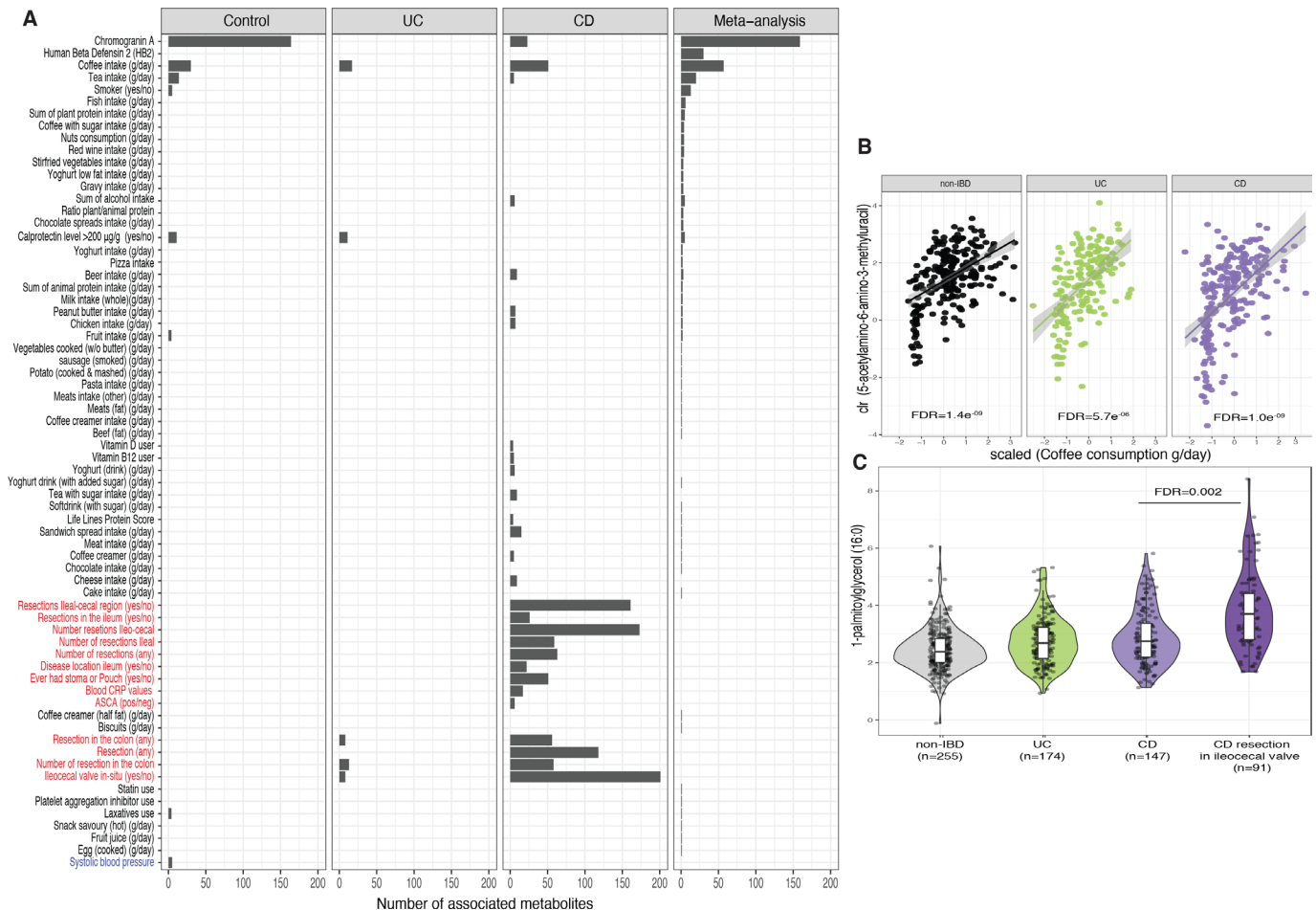
**Figure 2** Biomarker discovery for the diagnosis of IBD. (A, B) Show the abundance of the metabolites with the highest potential to discriminate between samples from non-IBD (grey) and IBD (UC in green and CD in purple). (C). Boxplots depict the value of a potential biomarker for IBD. The y-axis is the log-transformed value of the ratio constructed from the levels of lactosyl-N-palmitoyl-sphingosine (d18:1/16:0) and L-urobilin. Boxplot in grey depicts values in non-IBD controls. Boxplot in red depicts values in patients with IBD. (D). Receiver operating characteristic curve (ROC curve) of the prediction model based on patient characteristics (age, sex and BMI), the levels of faecal calprotectin (expressed as a binary trait (yes/no) if levels of this marker were >200 µg/g of faeces) and the ratio between metabolites. The prediction value, expressed as the area under the curve (AUC), reached a value of 0.83 in the test dataset. Metabolite values are clr-transformed. Boxplot shows the median and interquartile range (25th and 75th). Whiskers show the 1.5\*IQR range. Asterisks indicate significant differences between groups (FDR<0.05). BMI, body mass index; CD, Crohn’s disease; FDR, false discovery rate; UC, ulcerative colitis.

contrast, levels of hexanoic and valeric acids were significantly lower in both groups of patients (online supplemental figure 2, online supplemental table 7).

**Faecal metabolomic profiles correctly classify IBD samples**

Given the substantial variations observed in the metabolite levels between patients with IBD and non-IBD controls, we investigated the possibility of enhancing the accuracy of the faecal calprotectin test by combining multiple metabolites. To identify potential biomarkers, we employed a machine learning approach to predict disease phenotypes (see online supplemental methods). Including the ratio between the sphingolipid lactosyl-N-palmitoyl-sphingosine (d18:1/16:0) and L-urobilin improved the accuracy of age, sex, BMI and faecal calprotectin levels as disease predictors (AUC<sub>cv</sub>=0.85, AUC<sub>test</sub>=0.83, p=9.89e<sup>-13</sup>, figure 2, online supplemental

table 10). In addition, the ratio between these two metabolites was higher in patients with a long-term remission compared with controls (no flare-ups registered 1 year before and after sample collection, n=61, Wilcoxon test, p=0.0036) although significantly lower when compared with samples from other individuals with IBD in our cohort (Wilcoxon test, p=5.05e<sup>-5</sup>, online supplemental figure 3A). A similar performance was achieved with bacteria abundances (AUC<sub>cv</sub>=0.86, AUC<sub>test</sub>=0.84, p=6.04e<sup>-14</sup>, online supplemental figure 3B,C). Combining metabolite and microbiome ratios led to a modest but significant increase in model performance (AUC<sub>test</sub>=0.85, p=4.34 e<sup>-09</sup>). Within patients with IBD, metabolites showed a limited power to correctly classify CD or UC samples (AUC<sub>cv</sub>=0.78, AUC<sub>test</sub>=0.67) and active disease versus remission (AUC<sub>cv</sub>=0.72, AUC<sub>test</sub>=0.60) (online supplemental table 10).



**Figure 3** Potential determinants of faecal metabolite levels. (A) Bar plot showing the number of significant associations between phenotypes and metabolites in each of the cohorts and in the meta-analysis (online supplemental table 11). Only phenotypes with more than three associations are shown. Red labels indicate phenotypes exclusively available for cases and blue labels for controls. (B). Correlation plot showing the relation between AAMU (expressed as clr-transformed AUC values) and coffee consumption (x-axis) per cohort. Coffee consumption is represented as the estimated consumption per day (grams/day) adjusted by overall individual calorie intake (see online supplemental methods). (C). Boxplots showing the levels of 1-palmitoylglycerol (16:0). Boxplot shows the median and IQR (25th and 75th). Whiskers show the 1.5\*IQR range. Data distribution is represented by background violin-plot. Lines in the correlation plot show linear regression and shadows indicate the 95% CI. AAMU, 5-acetyl-amino-6-amino-3-methyluracil; AUC, area under the curve.

### Intestinal resections are associated with long-term metabolic alterations

After identifying alterations in the faecal metabolome in individuals with IBD, we aimed to describe which lifestyle, dietary and clinical factors contributed to the levels of faecal metabolites. We assessed the association between faecal metabolites and 229 host characteristics, including dietary habits, medication use and clinical features. We carried out association analysis per condition (i.e., CD only, UC only and controls only) and combined the results of overlapping metadata in a meta-analysis.

In patients with CD, resection of the ileocecal valve was associated with changes in the abundance of 212 metabolites, including cholic acid and several monoacylglycerols. Colonic resection was associated with modifications in the levels 56 molecules in CD and 8 molecules in UC (online supplemental table 11, figure 3 A-C). For example, colonic resection negatively correlated with the faecal levels of pyridoxamine (vitamin B<sub>6</sub>).

There were no significant differences in metabolites between different groups of disease behaviour or disease severity after statistically adjusting for gut surgery (resected vs non-resected).

However, we did observe several interesting trends ( $p < 0.05$ ,  $FDR > 0.05$ , online supplemental table 12). Patients with CD and penetrating diseases had lower butyrate levels (B1 vs B3). Disease severity (Montreal S score) positively correlated with tyramine faecal abundance. In patients with UC, participants with proctitis (E1 classification) had lower levels of 2R-3R-hydroxybutyrate and higher levels of cytidine compared with patients with extensive inflammation in the colon (E3 classification) (online supplemental figure 4A,B).

The levels of chromogranin A showed the largest number of associations with faecal metabolites in non-IBD controls, including positive associations with N-formylmethionine, cholesterol and secondary bile acids. Chromogranin A has been reported as a potential biomarker of gut health, showing a strong correlation with the microbiota composition in the gut.<sup>10</sup> Furthermore, participants with calprotectin  $> 200 \mu\text{g/g}$  showed lower levels of cytidine in faeces and increased sphingosines and ceramides in UC but not CD (figure 3A, online supplemental table 11).

Furthermore, the detection of several metabolites reflected aspects of the lifestyle of participants in our cohort. We found

158 associations between metabolites and dietary patterns ( $FDR_{meta} < 0.05$ , online supplemental table 11), however, approximately one-third of these were related to the consumption of coffee ( $n=57$ ), including positive correlations between coffee intake and the levels of picolinate and 5-acetylamino-6-amino-3-methyluracil (AAMU), a major caffeine metabolite (figure 3B) ( $FDR_{Meta} < 0.05$ , online supplemental table 11). Cotinine, an alkaloid found in tobacco plants, was found in faeces of self-reported smokers ( $FDR_{meta} = 1.31e^{-06}$ , online supplemental table 13) and O-desmethyltramadol, the primary metabolite of the opioid tramadol, was detected in several patients with CD using opioids ( $FDR=0.009$ , online supplemental table 13).

To investigate if the observed associations between lifestyle, clinical factors and faecal metabolites were driven by alterations in the gut microbiota we conducted a mediation analysis. We observed evidence for 119, 38 and 695 mediated effects in controls, UC and CD, respectively. Specifically, we found that *Lawsonibacter asaccharolyticus* mediated the relationship between coffee intake and several caffeine derivatives, such as 1,3-dimethyluric acid. In patients with CD, we observed that the resection of the ileocecal valve resulted in a decline in the abundance of *Faecalibacterium prausnitzii*, which negatively impacted the levels of anti-inflammatory metabolites, including butyrate (online supplemental table 14, online supplemental figure 5).

### NAT2 genotype strongly associated with coffee metabolism

Next, we carried out a faecal metabolome genome-wide association analysis to examine the correlation between host genetics and levels of faecal metabolites. Overall, genetics showed a relatively small impact on the faecal metabolite levels compared to the impact of genetics on blood metabolite levels reported in other studies.<sup>20–22</sup> At a study-wide significance level, we found an association between a genetic polymorphism located closely to NAT2 (rs4921913) and AAMU ( $p_{meta} = 1.79e^{-11}$ ). This genetic variant is in linkage disequilibrium with a SNP reported to be associated with the ratio between 1,3-dimethylurate and AAMU (rs35246381,  $r^2 > 0.8$ ).<sup>23</sup> As expected, we could also replicate this finding in our cohort ( $p_{IBD} = 8.46e^{-09}$ ,  $p_{controls} = 4.17e^{-09}$ ,  $p_{meta} = 3.57e^{-13}$ , figure 4). AAMU is a metabolite derived from coffee, and its levels in faeces correlate with coffee consumption. Nonetheless, this gene-metabolite association remained significant even after adjusting for coffee intake ( $p_{IBD} = 2.2e^{-16}$ ,  $p_{controls} = 2.0e^{-09}$ , online supplemental table 15).

### Gut microbiota composition is linked to metabolomic profiles

The gut microbiota of individuals with IBD often undergoes transitions from a healthy state (eubiosis) to an unhealthy state (dysbiosis).<sup>3,24</sup> Understanding the metabolic changes that accompany dysbiosis may provide crucial insights into the pathomechanisms of IBD.

In our cohort, participants with dysbiosis were more likely to have CD ( $n=130$ ,  $\chi^2$  test  $FDR=2.45e^{-04}$ ) and ileocecal valve resections ( $n=76$ ,  $\chi^2$  test,  $FDR=2.93e^{-10}$ ), but no significant differences were found in faecal calprotectin levels (proportion of individuals with calprotectin  $>200 \mu\text{g/g}$ ,  $\chi^2$  test  $FDR=0.65$ ). A significant increase in the abundance of pathobionts such as *Clostridium boltae*, *Erysipelatoclostridium ramosum* and *Ruminococcus gnavus* was observed, as well as a decreased abundance of 52 bacterial species in dysbiotic communities (online supplemental table 5, figure 5A,B).

Comparing the metabolite composition of IBD samples from patients with dysbiosis to those with eubiotic microbial

communities revealed the enrichment of 202 metabolites and the depletion of 258 metabolites. In dysbiotic samples, we observed reduced levels of indolin-2-one and 3-phenylpropionate and increased levels of imidazole propionate, long-chain polyunsaturated fatty acids (PUFAs) and primary bile acids ( $FDR < 0.05$ , online supplemental table 5, figure 5C). Alterations in the bile acids pools were also reflected in a higher prevalence of taurine-conjugated and sulphated bile acids in dysbiotic samples ( $FDR < 0.05$ , online supplemental table 5).

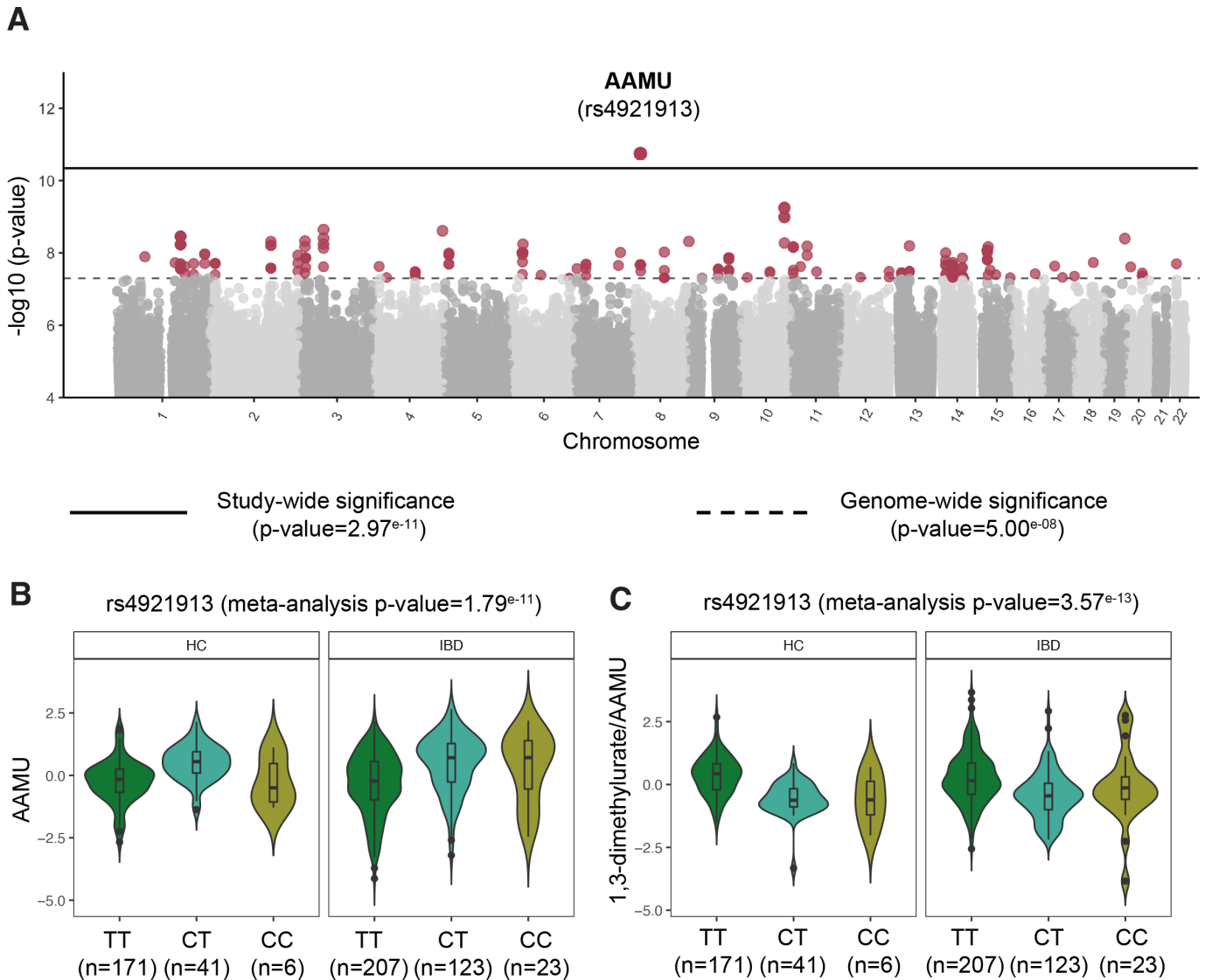
Next, we investigated the correlation between gut microbiota and metabolites while correcting for disease phenotypes (non-IBD, CD or UC) and dysbiotic status (yes/no) (see online supplemental methods). We found a total of 13 761 significant associations between bacteria presence/absence and metabolites levels, and 5942 significant associations between bacterial abundances and metabolites (online supplemental tables 16 and 17, figure 6A, online supplemental figure 6,  $FDR < 0.05$ ).

Of these associations, 1137 showed a significant interaction effect with dysbiosis status, with only 56 associations exhibiting different directionality between dysbiotic and eubiotic samples. For instance, the detection of *Ruthenibacterium lactatiformans* in eubiotic samples showed a negative association with butyrate levels, while in dysbiotic samples, this correlation was positive. Regardless of dysbiosis, the presence of *Akkermansia muciphila* and *Oscillibacter* spp (CAG 241) in faeces were associated with higher levels of dicarboxylic acids, sebacate (C10-DC) and dodecanedioate (C12), and the presence of *Bilophila wadsworthia* was associated with lower levels of taurine and N,N,N-trimethylalanyl-proline betaine (figure 6B,  $FDR < 0.05$ ). Furthermore, our results revealed strong positive correlations between *F. prausnitzii* and hypoxanthine abundances, ( $FDR = 1.46e^{-11}$ ), *R. gnavus* and tryptamine ( $FDR = 1.46e^{-11}$ ) (figure 6C), as well as imidazole propionate and *Streptococcus parasanguinis* ( $FDR = 0.007$ ).

Additionally, the abundance of specific microbial metabolic pathways and gene clusters were found to be linked with the metabolic profiles (figure 6D–F). Positive correlations were observed between the abundance of bile acid-inducible operons (*bai* operon) and levels of lithocholic acid ( $FDR = 9.76e^{-19}$ ), as well as negative correlations with cholic acid ( $FDR = 8.90e^{-06}$ , online supplemental table 18, figure 6F). However, these effects were more pronounced in dysbiotic samples ( $FDR_{interaction\ dysbiosis} < 0.05$ ). On the other hand, reductions in the levels of palmitoyl-ethanolamide and oleoyl-ethanolamide were associated with an increase in the abundance of ethanolamine utilisation operons (*eut* operon,  $FDR < 0.05$ , figure 6E). The *eut* operon is known to be carried by several gut pathobionts, allowing the use of ethanolamine as a source of carbon and nitrogen.<sup>25</sup> Moreover, a higher abundance of genes involved in the L-histidine degradation pathway I (MetaCyc ID: HISDEG-PWY) was associated with lower levels of histidine, a metabolite found to be increased in samples from patients with IBD ( $FDR = 3.64e^{-06}$ , online supplemental table S19, figure 6D).

### Microbiome composition predicts metabolite levels in faeces

Finally, the predictability of each metabolite was estimated using a combination of host information, dietary habits and the faecal microbiome. Dietary intake predicted the levels of 37 metabolites ( $>20\%$  of explained variation), with the top 10 dietary-predicted metabolites being 7 unclassified molecules and 3 coffee-related metabolites. Meanwhile, bacterial abundances were a strong predictor of 82 metabolites ( $>40\%$  of the variation), including the levels of molecules such as lithocholate (41%, s.d. 18%) and dimethylarginine (ADMA/SDMA, 53%,



**Figure 4** Genome-wide association between genetic polymorphisms and faecal metabolites. (A) Manhattan plot shows the strong association between a single nucleotide polymorphism located near the *NAT2* gene and AAMU, a metabolite derived from caffeine. Solid horizontal line signifies the significance threshold corrected by multiple hypothesis testing. Dashed line indicates the classic genome-wide significance threshold. Metabolites passing this threshold (in red) are considered suggestive associations (online supplemental table 15). (B) Boxplot depicting the levels of AAMU in non-IBD controls and IBD, stratified by SNP rs4921913 genotype. (C) Boxplot showing the relation between SNP rs4921913 and the ratio of 1,3-dimethylurate to AAMU. This association was previously described in the TwinsUK cohort.<sup>23</sup> Metabolite values are presented as the residuals of the model regressing the covariates age, sex, BMI and technical confounders. Boxplot shows the median and IQR (25th and 75th). Whiskers show the 1.5\*IQR range. Data distribution is represented by background violin-plot. AAMU, 5-acetylamino-6-amino-3-methyluracil; BMI, body mass index; IBD, inflammatory bowel disease.

s.d. 4%). Adding diet and participants' characteristics slightly improved microbiome-based models (paired Wilcoxon-test,  $p < 2.2 \times 10^{-6}$ ) (figure 7, online supplemental table 20).

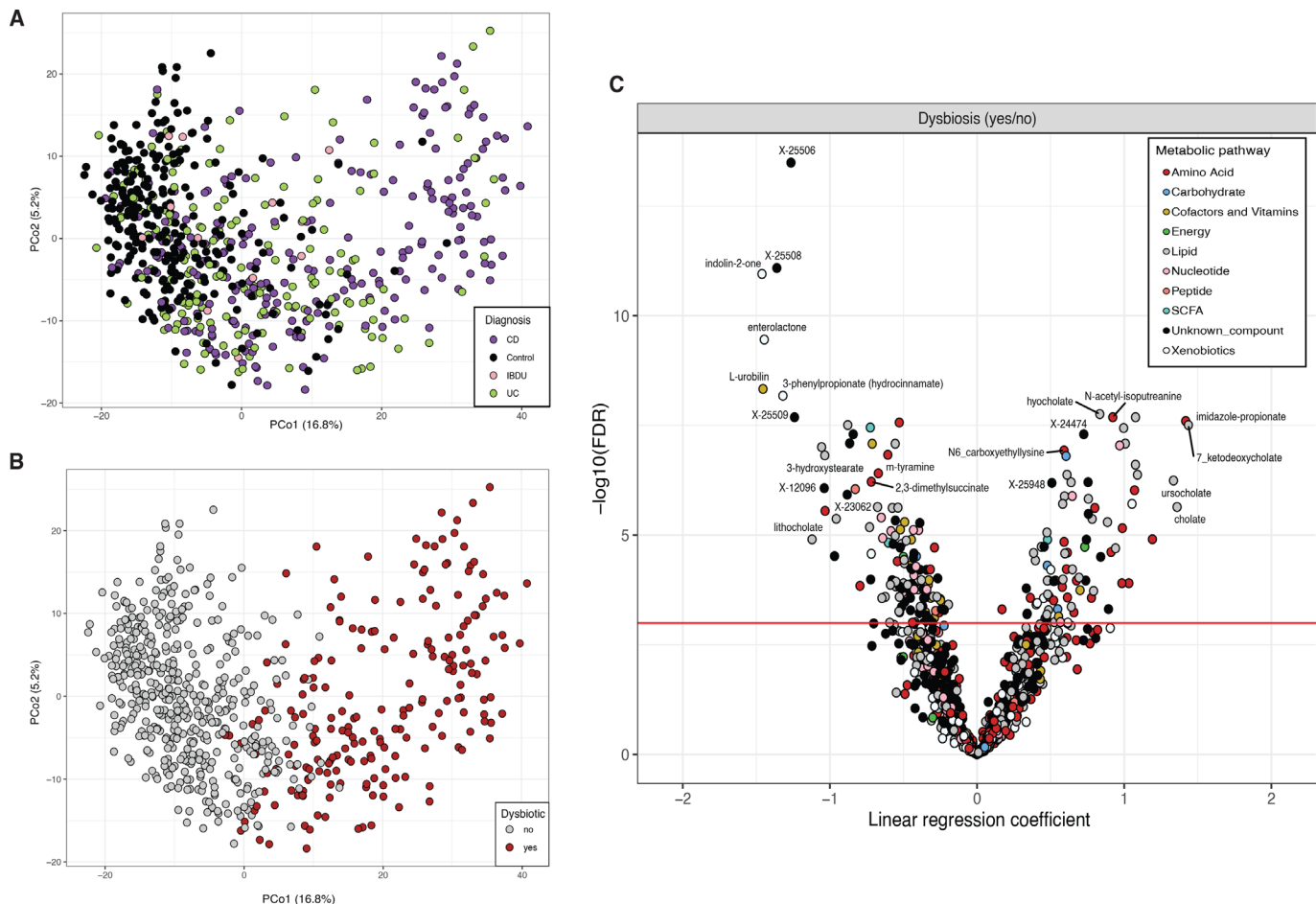
## DISCUSSION

We comprehensively characterised faecal metabolites in samples from patients with IBD and representatives of the Dutch population. Our results revealed alterations in the levels of more than 300 highly prevalent faecal metabolites in patients with IBD. Additionally, we described potential determinants of faecal metabolome composition by integrating untargeted metabolomics with extensive information on dietary habits, host genetics, clinical characteristics and gut microbiota composition.

The drastic alteration in faecal metabolite composition in patients with IBD suggests a shift from saccharolytic to proteolytic fermentation metabolism,<sup>26</sup> as evidenced by increased levels of metabolites derived from the metabolism of aromatic amino acids, such as p-cresol sulphate ( $FDR_{IBD} = 8.29 \times 10^{-6}$ ) and 3-indoxyl sulphate<sup>27-29</sup> ( $FDR_{IBD} = 0.04$ ). (online supplemental table 7). The accumulation of these compounds has been linked to various health conditions, such as chronic kidney disease<sup>30</sup> and colorectal cancer<sup>31,32</sup>, suggesting that higher presence of these molecules and lower levels of saccharolytic products, like SCFAs, may indicate an unhealthy gut milieu.

The overlap in the faecal metabolite signatures between patients with CD and UC suggests a common underlying alteration in gut metabolism. In total, 58% of the metabolites





**Figure 5** Metabolic signature of patients with intestinal dysbiosis. (A) Principal coordinate analysis on microbiome composition per sample (dots). Colours indicate disease phenotypes: CD (purple), UC (green), IBD-undetermined (pink), non-IBD (black). (B) Red dots depict samples considered to be dysbiotic based on the median distance to non-IBD samples. (C) Volcano plot showing the p value (y-axis) and regression coefficients (x-axis, positive values indicate enrichment in dysbiosis) of the association analyses between dysbiotic and non-dysbiotic IBD samples (online supplemental table 5). Dot colour indicates pathway annotations provided by Metabolon (online supplemental table 2). CD, Crohn's disease; IBD, inflammatory bowel disease; UC, ulcerative colitis.

significantly associated with UC were also found to be associated with CD. When comparing the faecal metabolite profiles of patients with CD and UC, we observed significant differences in the levels of 106 metabolites. For instance, alterations in the bile acid pool were a distinctive feature of CD, while a reduction in the concentrations of SCFAs was a common characteristic of UC.

In patients with CD, we observed a marked increase in the faecal levels of sphingolipids, including several sphingomyelins and ceramides. Sphingolipids are components of the intestinal cell membrane and are produced either by the de novo condensation of serine to palmitoyl-CoA or the uptake of endogenous and dietary sphingolipids. In addition to their structural role, sphingolipids can act as signalling molecules, mediating cell differentiation, apoptosis, and inflammation.<sup>33</sup> Previous studies have shown an accumulation of sphingolipids in colitis mouse models and in faeces of patients with IBD<sup>34</sup>; however, the mechanisms underlying this dysregulation and whether it precedes the development of inflammation are still unclear.

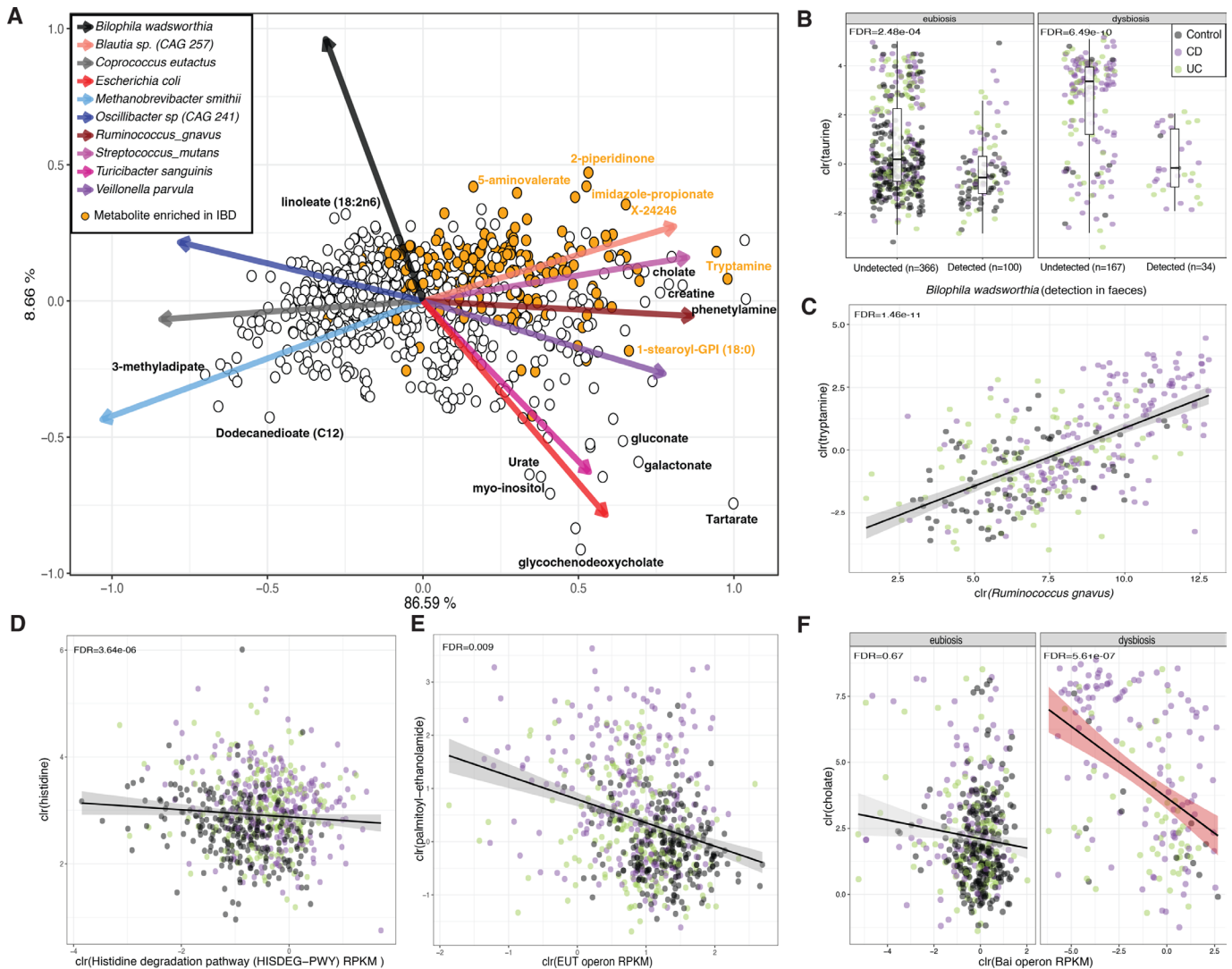
Experimental evidence suggests that an increase in ceramides levels, either due to the activation of ceramide synthetase or the increased breakdown of sphingomyelins into ceramides, can activate the proinflammatory transcription factor NF- $\kappa$ B, leading to the production of prostaglandin E<sub>2</sub> via the induction of COX-2 gene expression.<sup>35</sup> Ceramides can also be converted

into ceramide 1-phosphate or further degraded into sphingosine, which can also be phosphorylated to form sphingosine 1-phosphate (S1P). These molecules play a key role in regulating inflammatory processes, making S1P a promising target for IBD due to its role in modulating lymphocyte migration from lymph nodes.<sup>36</sup>

Contrary to the proinflammatory effects of host-produced sphingolipids, it has been shown that sphingolipids produced by bacteria like *Bacteroides* can exert anti-inflammatory effects,<sup>31</sup> emphasising the importance of microbial-produced molecules in maintaining intestinal health and the delicate balance between pro-inflammatory and anti-inflammatory molecules.

In addition to the increased levels of sphingolipids, we also report higher levels of N-acyl ethanolamines (palmitoylethanolamide, linoleoyl-ethanolamide, oleoyl-ethanolamide and stearoyl-ethanolamide) in the faeces of patients with CD compared with non-IBD controls. Although the mechanism behind the accumulation of these atypical endocannabinoids still needs to be elucidated, current evidence suggests that ethanolamides might shape the gut microbiota during inflammation.<sup>37</sup>

Our study found that patients with IBD also have elevated levels of long-chain fatty acids (LCFAs) and PUFAs, such as acylcarnitines and arachidonic acid, in their faeces. Previous research by Smith *et al* has identified palmitoylcarnitine (C16) as a faecal



**Figure 6** Metabolite co-occurrence with faecal microbes. (A) Biplot representing conditional probabilities of co-occurrence between metabolites (dots) and microbes (arrows). Distances between dots and arrow tips represent the probability of co-occurrence of each metabolite and microbe (online supplemental table 21). Orange dots highlight metabolites enriched in samples from patients with IBD in the linear regression analysis (online supplemental table 7). Arrow direction indicates the probability of microbes co-occurring with the levels of metabolites To enhance interpretability, names of only a few metabolites are shown and only the top-10 species explaining the largest amount of variation are visualised. (B) Taurine levels stratified by the presence or absence of *Bilophila wadsworthia* in faecal metagenomes. (C) Correlation between levels of tryptamine and abundance of *Ruminococcus gnavus*. Only samples in which the bacterium had a non-zero relative abundance are shown (n=339). (D–F) The relation between histidine and MetaCyc Histidine degradation pathway (D), between oleoyl-ethanolamide and the *eut* operon (E) and between cholic acid and the *bai* operon (F) are shown as examples of the correlation between microbiota metabolic potential and metabolite levels. Metabolite, bacteria and pathway values are clr-transformed. Boxplot shows the median and IQR (25th and 75th). Whiskers show the 1.5\*IQR range. Correlation plot lines show linear regression. Shadows indicate the 95% CI. IBD, inflammatory bowel disease.

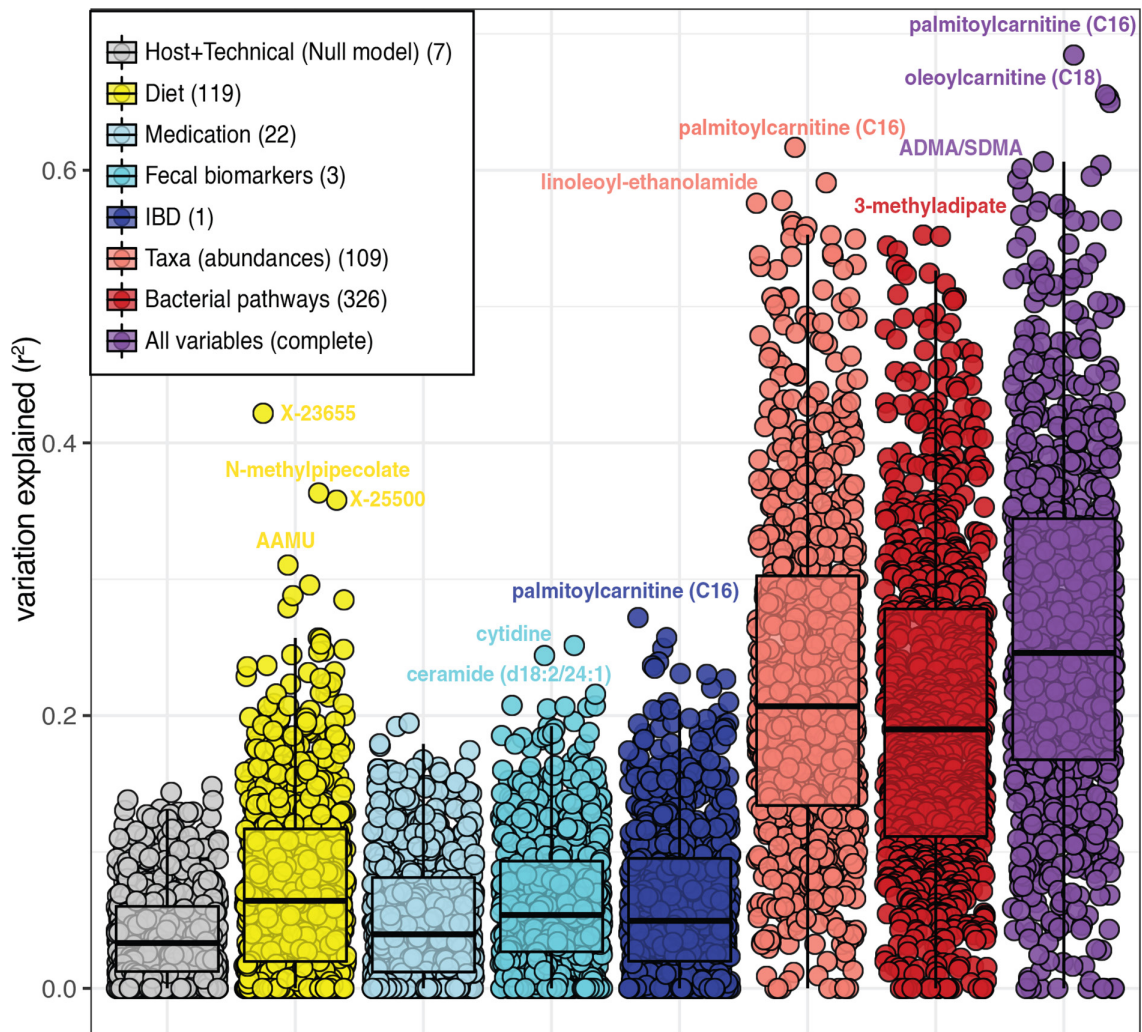
biomarker for IBD, linking acylcarnitine accumulation in the intestinal lumen to a reduced LCFA uptake in colonocytes during inflammation.<sup>38</sup> Furthermore, increasing evidence suggests that diets high in PUFAs can contribute to intestinal inflammation. Exposure to omega-3 and omega-6 PUFA can trigger an inflammatory response in intestinal organoids from patients with CD and in mice models with an impaired glutathione peroxidase 4 (GPX4) gene expression.

We also observed a significant increase in the levels of amino acids and their derivatives in patients with IBD. These findings align with previous research conducted on a cohort of newly diagnosed patients with IBD (n=78), where the levels of several amino acids could differentiate IBD samples from controls with

high accuracy.<sup>39</sup> In particular, we found a strong increase in taurine levels in IBD samples.

It has been shown that bacteria in the colon can use taurine as a substrate, releasing sulfite, which can be further converted to hydrogen sulfide.<sup>40</sup> This accumulation of hydrogen sulfide has been linked to epithelial damage and colitis. *B. wadsworthia* is a sulfate-reducing bacterium that has been shown to have the capability to metabolise taurine, which provides a potential explanation for the inverse relationship observed between the detection of *B. wadsworthia* and taurine levels in our cohort.

Furthermore, faeces from patients with IBD exhibited depletion of nucleotides, enterolactone (a bacterial product produced from the breakdown of dietary lignans) and biotin (vitamin B<sub>7</sub>).



**Figure 7** Metabolite prediction. Microbial abundances (light red) and bacterial pathways (dark red) show the largest potential to predict the levels of metabolites. Boxplots show the ability to predict metabolites levels of eight different models using seven types of data. Dots represent metabolites, and values in the y-axis represent the percentage of variation explained from cross-validated penalised regression methods using different sets of predictors (see online supplemental methods). The number of features in each model are indicated in parentheses in the legend (online supplemental table 20).

These findings suggest that the loss of bacterial diversity and biomass in the gut of patients with IBD<sup>41</sup> could drive the reduction in essential functions such as fibre digestion and vitamin production. The restoration of microbial production of these metabolites through dietary administration of their precursors could serve as a potential strategy to prevent flare-ups and address the dysregulation of the gut microbiome in IBD.

### The impact of genetics and lifestyle on the faecal metabolite levels

Along with the influence of IBD, diet and lifestyle are determinants of the abundance of small molecules in the human body. By correlating metabolites to dietary data, medication use and lifestyle factors, we found that daily habits such as smoking or coffee and tea consumption strongly correlated with their derivative molecules (online supplemental tables 11–13). Despite these associations, long-term dietary habits were moderately associated with faecal metabolome composition. Our prediction model revealed that only a few faecal metabolites could be predicted using dietary information (15 metabolites, explained variance >25%, online supplemental table 20), including

several unclassified metabolites, coffee and derivatives (AAMU, N-methyl pipercolate, theophylline) and enterolactone (a lignan derivative). Conversely, recent studies have reported more substantial impact of dietary intake on the levels of circulating metabolites.<sup>21</sup> We hypothesise that our dietary data underestimates the contribution of food intake to levels of faecal metabolites since it is based on food frequency questionnaires. Future studies should consider the use of 24-hour dietary recalls to capture daily dietary variations when aiming to explore relations between food intake and faecal metabolites and food–microbiome interactions. Furthermore, the impact of the host’s absorption rates, metabolism and biotransformations in the gastrointestinal tract should be considered when studying the relationship between dietary habits and faecal metabolites.

Our mediation analysis provided statistical evidence for the role of the gut microbiota as a mediator between faecal metabolites and clinical and lifestyle factors associations. For example, the levels of several coffee-related metabolites in faeces partially depended on the abundance of *L. asaccharolyticus*. Although this specie has been associated with coffee intake before,<sup>42</sup> its capacity to metabolise molecules found in coffee, such as

AAMU and 1,3-dimethylruic acid, remains unknown. However, relations between exposures, microbiota and metabolites are complex, for example, metabolites can shape the gut microbiota composition and bacteria can establish cross-feeding metabolic networks; therefore, functional validations are needed to better estimate the directionality of these interactions.

Host genetics showed a small impact on metabolite levels in faeces. The only association that passed our significance threshold was between a single nucleotide polymorphism near the *NAT2* gene and AAMU, a caffeine metabolism product (online supplemental table 15). *NAT2* encodes an N-acetyltransferase enzyme that detoxifies several xenobiotics, including coffee and certain types of medication. A study in the TwinsUK biobank also reported this association and estimated that host genetics has a moderate effect on faecal metabolites, with an average heritability of ~18%.<sup>23</sup> This relatively low heritability contrasts with the impact of host genetics on the levels of circulating metabolites<sup>43 44</sup> and might be explained by the fact that faecal metabolites are primarily influenced by microbial transformations occurring in the colon, which can potentially mask genetic effects. Moreover, sample sizes are still a limiting factor for discovering metabolite–genome associations. In fact, when using a looser significance cut-off ( $p < 5 \times 10^{-8}$ ), we found >200 suggestive associations pointing to the metabolism of cholesterol and serotonin. For example, *LRP5L* was associated with serotonin and *PNLIPRP2* with 1-palmitoyl-2-linoleoyl-digalactosyl glycerol (16:0/18:2) (online supplemental table 15). *LRP5L* belongs to the low-density lipoprotein (LDL) receptor family found to be involved in controlling serotonin levels in the duodenum.<sup>45</sup> Both *PNLIPRP2* and 1-palmitoyl-2-linoleoyl-digalactosyl glycerol (a choline derivative) are linked to cholesterol metabolism, supporting that choline supplements maintain blood cholesterol homeostasis,<sup>46</sup> and *PNLIPRP2* has been associated with LDL levels in blood.<sup>47</sup>

### The relation between gut microbiota composition and gut metabolism

The strong relationship between the microbiome and metabolites enabled us to estimate the levels of faecal metabolites using metagenomic sequencing data (online supplemental table 20). In line with previous studies<sup>23 43 48 49</sup>, well-predicted molecules included putrescine, urobilin, bile salts and fatty acids. However, further functional evidence is necessary to verify that all these well-predicted molecules can indeed be products of microbial metabolism. Notably, models trained on controls and tested on IBD samples showed lower prediction accuracy compared with models trained on both IBD and non-IBD datasets. This low cross-predictability between cases and controls has also been reported by Muller *et al*<sup>49</sup> and implies that some microbiota–metabolite associations may be context-specific or become more evident when microbial communities are perturbed. For instance, patients with IBD often exhibit alterations in their gut microbiota composition, leading to dysbiosis. In line with this, our analysis revealed 1137 metabolite–microbiota associations significantly influenced by dysbiosis.

Patients with IBD and dysbiosis displayed an enrichment of 202 highly prevalent metabolites, including a significant increase in primary and conjugated bile acid levels compared with eubiotic IBD samples (online supplemental table 5). The accumulation of cholic acid in the colon has been shown to exert selective pressure on the gut ecosystem due to its antimicrobial properties,<sup>50</sup> which could explain the expansion of bile-resistant bacteria, such as *C. boltae* and *R. gnavus* in dysbiosis. Additionally, the

loss of certain bacteria may also contribute to the accumulation of primary bile acids, as evidenced by the decreased abundance of *bai* operons and the increased ratio of primary to secondary bile acids in dysbiotic samples compared with eubiotic samples. Furthermore, bile acids play a pivotal role in regulating metabolism, exerting signalling effects in preserving the intestinal barrier and regulating the host's immune system,<sup>51</sup> thereby making them a highly attractive target for therapeutic intervention in IBD.

In our cohort, dysbiosis was associated with ileum disease involvement and an ileocecal valve resection (online supplemental table 5). Accumulating literature demonstrates that disruptions in the small intestine due to inflammation or surgery can significantly impact faeces' metabolite and microbial composition.<sup>52 53</sup> For example, Halfvarson *et al*<sup>24</sup> showed that patients with intestinal surgery in the ileum had a less stable microbiota and more frequently transitioned between eubiotic and dysbiotic states. Given the critical role of the small intestine in nutrient absorption, it is plausible that disruptions in this section of the gut lead to persistent alterations in the concentrations of bile acids, amino acids and lipids in the colon, which might reshape the microbial composition towards a dysbiotic state. The stratification of patients based on their disease location and microbiome composition should be considered in future metabolomic studies and clinical interventions, as it can potentially uncover more targeted and effective treatments for the disease.

Furthermore, the coabundance analysis performed in this study provides insight into the relationship between bacteria and their associated metabolic products. This information can serve as a basis for identifying potential therapeutic targets for treating IBD. For instance, *F. prausnitzii*, a specie which is depleted in the faeces of patients with IBD, was positively correlated with SCFAs and hypoxanthine levels (online supplemental table 17). Hypoxanthine can be produced by *F. prausnitzii* through the metabolism of adenine<sup>54</sup> and plays a role in maintaining the intestinal epithelium.<sup>55</sup> Similarly, *R. gnavus*, which is highly abundant in dysbiotic samples from patients with IBD,<sup>56</sup> was positively correlated with the levels of tryptamine (online supplemental table 17). *R. gnavus* is capable of producing tryptamine via the decarboxylation of tryptophan.<sup>57</sup> Accumulation of tryptamine may increase gut motility via activation of serotonin receptor-4, which could explain why some patients experience decreased intestinal transit times during flares.<sup>58</sup> In line with our findings, higher levels of tryptamine have been reported in individuals with irritable bowel syndrome and diarrhoea and have been associated with the metabolic activity of *R. gnavus*.<sup>59</sup> Additionally, the positive correlation we observed between *S. parasanguinis* and imidazole propionate could be explained by the capacity of this bacterium to degrade histidine.<sup>60</sup> Imidazole propionate has been linked to the risk of developing type 2 diabetes and regulates activation of the mTORC1 signalling pathway,<sup>61 62</sup> which is implicated in IBD.<sup>63</sup>

Overall, the substantial variations in metabolite composition between eubiotic and dysbiotic microbial communities, and the strong co-occurrence between metabolites and bacterial species, support the notion that faecal metabolomics partially reflects the metabolic activity of the gut microbiota. Further functional validation and longitudinal monitoring of microbial–metabolite associations are necessary to determine the direction of these relationships and assess their impact on disease progression.

### Faecal metabolites as novel biomarkers for IBDs

As opposed to colonoscopies, the current invasive gold standard for diagnosing IBD, we demonstrated the potential of faecal

metabolites as a non-invasive method for disease diagnosis. The ratio between the levels of two metabolites, lactosyl-N-palmitoyl-sphingosine (d18:1/16:0) and L-urobilin was identified as a biomarker for IBD in our cohort (online supplemental table 10). Other studies have also observed reduced levels of L-urobilin and increased sphingolipids in patients with IBD faeces.<sup>3 5</sup> In a North American longitudinal cohort,<sup>3</sup> we observed that the ratio between a sphingolipid (ceramide (18:1/16:0)), and L-urobilin were consistently higher in patients with IBD compared with non-IBD controls underscoring our findings (online supplemental figure 7). Faecal measurements targeting these two molecules could be relatively easy to implement in combination with the faecal calprotectin test.

It is important to note that our study cohort primarily consisted of subjects with a prolonged history of IBD. Therefore, the validity of the identified biomarker must be confirmed in newly diagnosed patients with IBD, as well as in individuals with other gastrointestinal disorders.

Finally, it is also important to acknowledge the limitations of untargeted metabolomics approaches. This study focused on annotated molecules with relatively high prevalence, but a substantial number of metabolites remained unidentified, and their physiological significance is unknown. Approximately one-third of the metabolites detected in our dataset (492 out of 1684 metabolites) could not be linked to a previously characterised compound, emphasising the need for further efforts to fully characterise the molecular diversity in the human body. Additionally, the semiquantitative nature of untargeted metabolomics limits the ability to establish the normal concentration range of each metabolite in faeces.

In conclusion, this study provides a detailed characterisation of the faecal metabolites in the context of health and intestinal inflammation, replicating known disease-relevant molecules and expanding our knowledge of disease heterogeneity. In addition, we pinpoint multiple associations between microbiota, diet and faecal metabolite levels, which we believe provide valuable resources for further investigation of metabolite-based or microbiota-based interventions and treatment in IBD.

**Correction notice** This article has been corrected since it published Online First. The supplementary figures have been replaced.

**Twitter** Arnau Vich Vila @arnauvich, Valerie Collij @CollijValerie, Hannah E Augustijn @HannahAugustijn, Alexandra Zhernakova @SashaZhernakova, Ranko Gacesa @RGacesa and Jingyuan Fu @jingyuan\_fu

**Acknowledgements** The authors thank Kate Mc Intyre for substantive English editing.

**Contributors** AVV and RKW designed the study. AVV, SH, SA-S, RG, HEA contributed to the data analysis. AVV and SH wrote the manuscript. BHDJ handled the samples in the laboratory. LAB, SH, SA-S, VC, HEA, JF, AZ, AAG, JP, JS, CG, GA-A, RAAAR and RKW critically reviewed the manuscript. AVV and RKW are responsible for the overall content as guarantor.

**Funding** This study was funded by Takeda Development Center Americas. JF is supported by the Dutch Heart Foundation IN-CONTROL (CVON2018-27), the ERC Consolidator grant (grant agreement No. 101001678), NWO-VICI grant VI.C.202.022, and the Netherlands Organ-on-Chip Initiative, an NWO Gravitation project (024.003.001) funded by the Ministry of Education, Culture and Science of the government of The Netherlands. AZ is supported by the Dutch Heart Foundation IN-CONTROL (CVON2018-27), the ERC Starting Grant 715772, NWO-VIDI grant 016.178.056, ZONMW MEMORABEL grant 733050814 and the NWO Gravitation grant Exposome-NL (024.004.017). RKW is supported by HORIZON-HLTH-2022-STAYHLTH-02-0 and the Seerave Foundation.

**Competing interests** This study was funded by Takeda Development Center Americas. RKW acted as a consultant for Takeda and received unrestricted research grants from Takeda and Johnson and Johnson pharmaceuticals and speaker fees from AbbVie, MSD, Olympus and AstraZeneca. GA-A, CG, JS, JP and AAG are or were employees of Takeda Pharmaceuticals at the time this study was conducted. No disclosures: All other authors have nothing to disclose.

**Patient and public involvement** Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

**Patient consent for publication** Not applicable.

**Ethics approval** This study involves human participants and was approved by M12.1139652008.338. Participants gave informed consent to participate in the study before taking part.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data are available on reasonable request. Tables containing the levels of faecal metabolites and bacterial taxa abundances are provided with the manuscript. The raw metagenomics, host genomics and phenotypic data used in this study are available from the European Genome-Phenome Archive data repository: 1000 Inflammatory bowel disease (IBD) cohort (<https://www.ebi.ac.uk/ega/datasets/EGAD00001004194>), Lifelines DEEP cohort (<https://www.ebi.ac.uk/ega/datasets/EGAD00001001991>). This includes submitting a letter of intent to the corresponding data access committees. Codes are publicly available at: [https://github.com/GRONINGEN-MICROBIOME-CENTRE/Fecal\\_Metabolites\\_IBD](https://github.com/GRONINGEN-MICROBIOME-CENTRE/Fecal_Metabolites_IBD)

**Supplemental material** This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

**Open access** This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

#### ORCID iDs

Arnau Vich Vila <http://orcid.org/0000-0003-4691-5583>  
 Shixian Hu <http://orcid.org/0000-0002-1190-0325>  
 Sergio Andreu-Sánchez <http://orcid.org/0000-0002-3503-9971>  
 Valerie Collij <http://orcid.org/0000-0003-3743-1544>  
 Hannah E Augustijn <http://orcid.org/0000-0002-1862-6699>  
 Laura A Bolte <http://orcid.org/0000-0001-6036-0831>  
 Alexandra Zhernakova <http://orcid.org/0000-0002-4574-0841>  
 Ranko Gacesa <http://orcid.org/0000-0003-2119-0539>  
 Jingyuan Fu <http://orcid.org/0000-0001-5578-1236>  
 Rinse K Weersma <http://orcid.org/0000-0001-7928-7371>

#### REFERENCES

- Kaplan GG, Ng SC. Understanding and preventing the global increase of inflammatory bowel disease. *Gastroenterology* 2017;152:313–21.
- Kostic AD, Xavier RJ, Gevers D. The microbiome in inflammatory bowel disease: current status and the future ahead. *Gastroenterology* 2014;146:1489–99.
- Lloyd-Price J, Arze C, Ananthakrishnan AN, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 2019;569:655–62.
- Lavelle A, Sokol H. Gut microbiota-derived metabolites as key actors in inflammatory bowel disease. *Nat Rev Gastroenterol Hepatol* 2020;17:223–37.
- Franzosa EA, Sirota-Madi A, Avila-Pacheco J, et al. Author correction: gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat Microbiol* 2019;4:898.
- Johnson CH, Ivanisevic J, Szuздak G. Metabolomics: beyond biomarkers and towards mechanisms. *Nat Rev Mol Cell Biol* 2016;17:451–9.
- Bauset C, Gisbert-Ferrández L, Cosin-Roger J. Metabolomics as a promising resource identifying potential biomarkers for inflammatory bowel disease. *J Clin Med* 2021;10:622.
- Tigchelaar EF, Zhernakova A, Dekens JAM, et al. Cohort profile: lifelines deep, a prospective, general population cohort study in the Northern Netherlands: study design and baseline characteristics. *BMJ Open* 2015;5:e006772.
- Imhann F, Van der Velde KJ, Barbieri R, et al. The 1000IBD project: multi-omics data of 1000 inflammatory bowel disease patients; data release 1. *BMC Gastroenterol* 2019;19:5.
- Zhernakova A, Kurilshikov A, Bonder MJ, et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* 2016;352:565–9.
- Vich Vila A, Imhann F, Collij V, et al. Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. *Sci Transl Med* 2018;10:eap8914.
- Kowarik A, Templ M. Imputation with the R package VIM. *J Stat Softw* 2016;74:1–16.

- 13 Gordon-Rodriguez E, Quinn TP, Cunningham JP. Learning sparse log-ratios for high-throughput sequencing data. *Bioinformatics* 2021;38:157–63.
- 14 Kurilshikov A, Medina-Gomez C, Bacigalupe R, et al. Large-scale association analyses identify host factors influencing human gut microbiome composition. *Nat Genet* 2021;53:156–65.
- 15 Bolyen E, Rideout JR, Dillon MR, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* 2019;37:852–7.
- 16 Morton JT, Aksenov AA, Nothias LF, et al. Learning representations of microbe-metabolite interactions. *Nat Methods* 2019;16:1306–14.
- 17 Friedman JH, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010;33:1–22.
- 18 Chamailard M, Hashimoto M, Horie Y, et al. An essential role for NOD1 in host recognition of bacterial peptidoglycan containing diaminopimelic acid. *Nat Immunol* 2003;4:702–7.
- 19 Parada Venegas D, De la Fuente MK, Landskron G, et al. Corrigendum: short chain fatty acids (SCFAs)-mediated gut epithelial and immune regulation and its relevance for inflammatory bowel diseases. *Front Immunol* 2019;10:1486.
- 20 Diener C, Dai CL, Wilmski T, et al. Genome-microbiome interplay provides insight into the determinants of the human blood metabolome. *Nat Metab* 2022;4:1560–72.
- 21 Chen L, Zernakova DV, Kurilshikov A, et al. Influence of the microbiome, diet and genetics on inter-individual variation in the human plasma metabolome. *Nat Med* 2022;28:2333–43.
- 22 Yin X, Chan LS, Bose D, et al. Genome-wide association studies of metabolites in finnish men identify disease-relevant loci. *Nat Commun* 2022;13:1644.
- 23 Zierer J, Jackson MA, Kastenmüller G, et al. The fecal metabolome as a functional readout of the gut microbiome. *Nat Genet* 2018;50:790–5.
- 24 Halfvarson J, Brislawn CJ, Lamendella R, et al. Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat Microbiol* 2017;2:17004.
- 25 Garsin DA. Ethanolamine utilization in bacterial pathogens: roles and regulation. *Nat Rev Microbiol* 2010;8:290–5.
- 26 Korpela KD. Diet, microbiota, and metabolic health: trade-off between saccharolytic and proteolytic fermentation. *Annu Rev Food Sci Technol* 2018;9:65–84.
- 27 Gryp T, Vanholder R, Vaneechoutte M, et al. P-cresyl sulfate. *Toxins (Basel)* 2017;9:52.
- 28 Kikuchi K, Saigusa D, Kanemitsu Y, et al. Gut microbiome-derived phenyl sulfate contributes to albuminuria in diabetic kidney disease. *Nat Commun* 2019;10:1835.
- 29 Edamatsu T, Fujieda A, Itoh Y. Phenyl sulfate, indoxyl sulfate and p-cresyl sulfate decrease glutathione level to render cells vulnerable to oxidative stress in renal tubular cells. *PLoS One* 2018;13:e0193342.
- 30 Liu W-C, Tomino Y, Lu K-C. Impacts of indoxyl sulfate and p-cresol sulfate on chronic kidney disease and mitigating effects of AST-120. *Toxins (Basel)* 2018;10:367.
- 31 Brown DG, Rao S, Weir TL, et al. Metabolomics and metabolic pathway networks from human colorectal cancers, adjacent mucosa, and stool. *Cancer Metab* 2016;4:11.
- 32 Al Hinai EA, Kullamethee P, Rowland IR, et al. Modelling the role of microbial p-cresol in colorectal genotoxicity. *Gut Microbes* 2019;10:398–411.
- 33 Abdel Hadi L, Di Vito C, Riboni L. Fostering inflammatory bowel disease: sphingolipid strategies to join forces. *Mediators Inflamm* 2016;2016:3827684.
- 34 Maceyka M, Spiegel S. Sphingolipid metabolites in inflammatory disease. *Nature* 2014;510:58–67.
- 35 Nixon GF. Sphingolipids in inflammation: pathological implications and potential therapeutic targets. *Br J Pharmacol* 2009;158:982–93.
- 36 Danese S, Furfaro F, Vetrano S. Targeting S1P in inflammatory bowel disease: new avenues for modulating intestinal leukocyte migration. *J Crohns Colitis* 2018;12:S678–86.
- 37 Fornelos N, Franzosa EA, Bishai J, et al. Growth effects of N-acyl ethanolamines on gut bacteria reflect altered bacterial abundances in inflammatory bowel disease. *Nat Microbiol* 2020;5:486–97.
- 38 Smith SA, Ogawa SA, Chau L, et al. Mitochondrial dysfunction in inflammatory bowel disease alters intestinal epithelial metabolism of hepatic acylcarnitines. *J Clin Invest* 2021;131:e133371.
- 39 Jagt JZ, Struys EA, Ayada I, et al. Fecal amino acid analysis in newly diagnosed pediatric inflammatory bowel disease: a multicenter case-control study. *Inflamm Bowel Dis* 2022;28:755–63.
- 40 Walker A, Schmitt-Kopplin P. The role of fecal sulfur metabolome in inflammatory bowel diseases. *Int J Med Microbiol* 2021;311:151513.
- 41 Vieira-Silva S, Sabino J, Valles-Colomer M, et al. Quantitative microbiome profiling disentangles inflammation- and bile duct obstruction-associated microbiota alterations across PSC/IBD diagnoses. *Nat Microbiol* 2019;4:1826–31.
- 42 Asnicar F, Berry SE, Valdes AM, et al. Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. *Nat Med* 2021;27:321–32.
- 43 Bar N, Korem T, Weissbrod O, et al. A reference map of potential determinants for the human serum metabolome. *Nature* 2020;588:135–40.
- 44 Shin S-Y, Fauman EB, Petersen A-K, et al. An atlas of genetic influences on human blood metabolites. *Nat Genet* 2014;46:543–50.
- 45 Yadav VK, Ryu J-H, Suda N, et al. LRP5 controls bone formation by inhibiting serotonin synthesis in the duodenum. *Cell* 2008;135:825–37.
- 46 Al Rajabi A, Castro GSF, da Silva RP, et al. Choline supplementation protects against liver damage by normalizing cholesterol metabolism in Pempt/Ldlr knockout mice fed a high-fat diet. *J Nutr* 2014;144:252–7.
- 47 Liu DJ, Peloso GM, Yu H, et al. Exome-wide association study of plasma lipids in > 300,000 individuals. *Nat Genet* 2017;49:1758–66.
- 48 Mallick H, Franzosa EA, McIver LJ, et al. Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. *Nat Commun* 2019;10:3136.
- 49 Muller E, Algavi YM, Borenstein E. A meta-analysis study of the robustness and universality of gut microbiome-metabolome associations. *Microbiome* 2021;9:203.
- 50 Wahlström A, Sayin SI, Marschall H-U, et al. Intestinal crosstalk between bile acids and microbiota and its impact on host metabolism. *Cell Metab* 2016;24:41–50.
- 51 Thomas JP, Modos D, Rushbrook SM, et al. The emerging role of bile acids in the pathogenesis of inflammatory bowel disease. *Front Immunol* 2022;13:829525.
- 52 Fang X, Vázquez-Baeza Y, Elijah E, et al. Gastrointestinal surgery for inflammatory bowel disease persistently lowers microbiome and metabolome diversity. *Inflamm Bowel Dis* 2021;27:603–16.
- 53 Jansson J, Willing B, Lucio M, et al. Metabolomics reveals metabolic biomarkers of Crohn's disease. *PLoS ONE* 2009;4:e6386.
- 54 Heinken A, Khan MT, Paglia G, et al. Functional metabolic map of Faecalibacterium prausnitzii, a beneficial human gut microbe. *J Bacteriol* 2014;196:3289–302.
- 55 Lee JS, Wang RX, Alexeev EE, et al. Hypoxanthine is a checkpoint stress metabolite in colonic epithelial energy modulation and barrier function. *J Biol Chem* 2018;293:6039–51.
- 56 Hall AB, Yassour M, Sauk J, et al. A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients. *Genome Med* 2017;9:103.
- 57 Williams BB, Van Benschoten AH, Cimermanic P, et al. Discovery and characterization of gut microbiota decarboxylases that can produce the neurotransmitter tryptamine. *Cell Host Microbe* 2014;16:495–503.
- 58 Bhattarai Y, Williams BB, Battaglioli EJ, et al. Gut microbiota-produced tryptamine activates an epithelial G-protein-coupled receptor to increase colonic secretion. *Cell Host Microbe* 2018;23:775–85.
- 59 Zhai L, Huang C, Ning Z, et al. Ruminococcus gnavus plays a pathogenic role in diarrhea-predominant irritable bowel syndrome by increasing serotonin biosynthesis. *Cell Host Microbe* 2023;31:33–44.
- 60 Molinaro A, Bel Lassen P, Henricsson M, et al. Imidazole propionate is increased in diabetes and associated with dietary patterns and altered microbial ecology. *Nat Commun* 2020;11:5881.
- 61 Koh A, Molinaro A, Ståhlman M, et al. Microbially produced imidazole propionate impairs insulin signaling through mTORC1. *Cell* 2018;175:947–61.
- 62 Koh A, Mannerås-Holm L, Yunn N-O, et al. Microbial imidazole propionate affects responses to metformin through p38γ-dependent inhibitory AMPK phosphorylation. *Cell Metab* 2020;32:643–53.
- 63 Lin X, Sun Q, Zhou L, et al. Colonic epithelial mTORC1 promotes ulcerative colitis through COX-2-mediated Th17 responses. *Mucosal Immunol* 2018;11:1663–73.

## SUPPLEMENTARY METHODS

### Sample collection

Sample collection took place between 2014-2017. Participants were asked to collect and freeze the faecal samples at home and were picked up and transported on dry ice and stored at  $-80^{\circ}\text{C}$  within 24 h after collection. For this study, fresh frozen samples were drilled on dry ice until obtaining on average 0.5 mg of faecal material, transferred into a 2ml tube and finally shipped to Metabolon facilities for metabolomic measurements. Samples from the LifeLines and 1000IBD cohorts were collected using the same protocol and processed by the same lab technician.

### Metabolite's quantification

Metabolomics measurements were performed by Metabolon Inc. (North Carolina, U.S.A.). In short, proteins and organic solvent were removed from each sample. Next, each sample was divided into four fractions for analysis: two for analysis by two separate reverse phases (RP)/UPLC-MS/MS methods with positive ion mode electrospray ionisation (ESI), one for analysis by RP/UPLC-MS/MS with negative ion mode ESI, one for analysis by HILIC/UPLC-MS/MS with negative ion mode ESI. Raw data processing and quality control were performed according to Metabolon's standards.

In addition to untargeted metabolomics, the concentration of eight short-chain and branched-chain fatty acids. i.e., acetic acid (C2), propionic acid (C3), isobutyric acid (C4), butyric acid (C4), 2-methyl-butyric acid (C5), isovaleric acid (C5), valeric acid (C5) and caproic acid (hexanoic acid, C6), were measured using LC-MS/MS methods. Acetic acid was the most abundant SCFA in the faecal samples (mean: 2339  $\mu\text{g/g}$ , s.d. 1,131  $\mu\text{g/g}$ ), followed by butyrate (mean: 1,072.5  $\mu\text{g/g}$ , s.d. 678  $\mu\text{g/g}$ ) and propionate (mean: 955.58  $\mu\text{g/g}$ , s.d. 515.8  $\mu\text{g/g}$ ), while hexanoic acid presented the lowest concentrations (mean 74  $\mu\text{g/g}$ , s.d. 110  $\mu\text{g/g}$ ).

### Phenotypes selection

We retrieved the metadata including 180 entries consisting of dietary habits, medication and anthropomorphic measurements overlapping in both cohorts and 33

and 13 phenotypes specific to the IBD and the control cohorts, respectively. This information was included in further analyses if each category had at least 10 entries. A complete list of phenotypes is provided in **Supplementary Table 1**. Samples from patients with colectomy or stomas at the time of sample collection were removed from our analysis since their faecal samples are not representative of the content of the whole intestinal tract (n=68).

To adjust for differences in intestinal transit time, we combined the information about “bowel movements per day” present in the control cohort with questionnaires on the type of stools and frequencies a day in the IBD cohort. In the group of patients with IBD, disease remission or flare was defined using a combination of biomarkers, i.e., faecal calprotectin above 200 µg/g and Harvey-Bradshaw index >4 or Simple Clinical Colitis Activity Index (SCCAI) >2.5, and colonoscopy reports when available<sup>1</sup>.

Dietary intake was assessed through a validated food frequency questionnaire (FFQ) collected concurrently with faecal samples as described before<sup>2,3</sup>. Estimated food and nutrient intakes were adjusted by total caloric intake using regression analysis described in<sup>4</sup>. In addition, nutrient ratios and dietary patterns were calculated using pre-defined scoring systems:

- **Lifelines Protein score**, reflecting a higher protein energy percentage within the acceptable macronutrient distribution range for protein and a higher plant to animal protein ratio.
- **Lifelines Diet score**, expressing relative dietary quality with a higher score reflecting a high intake of vegetables, fruits, nuts, legumes and fish and lower intakes of red and processed meats and high sugar snacks and beverages.
- **Plant-to-Animal protein ratio**, reflecting a higher intake of plant protein relative to animal protein

### **Metabolite ratios calculation**

In addition to individual metabolites, we calculated the ratios between molecules of interest. Ratios were calculated by dividing the raw metabolite's levels, log transforming and scaling the resulting value.

In total, we evaluated 12 different ratios. The ratio between primary and secondary bile acids (deoxycholate/choleate, lithocholate/chenodeoxycholate,



ursodeoxycholate/chenodeoxycholate), the ratio between conjugated and unconjugated bile acids (glycol + tauro bile acids / unconjugated bile acid), the ratios between kynurenine, tryptamine, serotonin and tryptophan and the ratios between omega-3 PUFA and omega-6 PUFA. In our dataset, we could quantify the levels of docosahexanoate (DHA), docopentaenoate (DPA), eicosapentaenoate (EPA), hexadecatrienoate and stearidonate as omega-3 PUFAs, and arachidonate, dihomolinoleate, dihomolinolenate, docosadienoate, hexadecadienoate and linoleate as omega-6 PUFAs.

### **Prediction of microbial abundance**

Metagenomic reads mapping to the human genome were removed and reads containing Illumina adapters were trimmed using *KneadData* (v0.4.6.1)<sup>5</sup>. Other potential contaminants were also filtered out using *Kraken2*<sup>6</sup> and the NCBI UniVec database, with the confidence parameter set to 0.5. After quality control of the sequenced reads, the microbial taxonomic and functional profiles were determined using *MetaPhlAn* (v3.0)<sup>5</sup>. Moreover, *HUMANN 3.0* pipeline was used to estimate the metabolic potential of each microbial community<sup>5</sup>.

Three samples from patients with IBD were removed due to failure in the identification of bacterial species in their faecal sample. Previous to statistical tests, bacterial and pathway abundances were transformed using a centred-log ratio approach (CLR). Bacterial species and pathways present in more than 20% of the samples were kept for further analysis.

### **Estimation of bacterial metabolic gene clusters in metagenomic samples**

Metagenomic reads were aligned to a collection of predicted metabolic gene clusters (MGC) predicted using GutSMASH<sup>7</sup>. BiG-MAP<sup>8</sup> pipeline, with its default parameters, was used for read mapping and coverage calculations. In total, 6083 MGC were found in our dataset, for which, 1102 were kept after filtering for minimum coverage of 5% in the core genes of each cluster. To summarise the overall metabolic capacity of the microbial community, MGC were collapsed according to their predicted function by summing RPKM values. For example, the 5 different bai operons found in *Dorea sp. D27*, *Dorea sp AF36-15-AT*, *Clostridium scidens* (ATC 35704), *Clostridium hylemonae*

(*DSM 15053*) and *Clostridium hiranonis* (*DSM 13275*) genomes, were merged into one *bai* operon category.

Centred log-ratio transformation was applied before data analysis. In total, 136 pathways were identified and 134 were kept for analysis after removing pathways that were present in less than 20% of the samples: “lysine degradation acetate to butyrate” and a “nitrate reductase”.

### **Definition of dysbiosis**

Samples were defined as “dysbiotic” based on the microbiota composition in a similar way as described in Lloyd-Price et al.<sup>9</sup>. Euclidean distances between samples were computed on a clr-transformed bacterial abundances matrix. Non-IBD samples were used as a reference of eubiosis. Then, we computed the median distance between each sample and this reference group. A threshold for dysbiosis was defined at the 95<sup>th</sup> quantile of the median distances between non-IBD samples. Samples exceeding this threshold were considered dysbiotic.

### **Genome-wide association analysis power analysis**

Power estimations were conducted as described here<sup>10</sup>. First the relation between sample size and detection power was calculated while taking a grid search in the variance explained by the SNP (0.0~0.1). We then calculated the effects of metabolite detection rates (10%~100%) on the statistical power. The sample size in this study allowed us to have 80% power to detect genetic associations with 8% trait variation. The genetic effect of variants located in the NAT2 gene can explain 8% of the 5-acetylamino-6-amino-3-methyluracil variation (a metabolite with ~99% of prevalence in both IBD and controls) (**Suppl. Figure 8**).

### **Defining host genetics combining whole-exome sequencing (WES) and global screening array (GSA)**

Library preparation, sequencing and variant calling were done at the Broad Institute of the Massachusetts Institute of Technology (MIT) and Harvard University. On average, 86.06 million high-quality reads were generated per sample and 98.85% of reads were aligned to a human reference genome (hg19). Moreover, 81% of the exonic regions were covered with a read depth >30x. Next, the Genome Analysis Toolkit (GATK) was

used for variant calling<sup>11</sup>. Variants with a call rate <0.99 or Hardy-Weinberg equilibrium  $\chi^2$  test with p-value<0.0001 were excluded using PLINK 1.9<sup>12</sup>

GSA data was generated using the Infinium GSA-24 v1.0 BeadChip combined with the optional multi-Disease drop-in panel. Genotypes were called using OptiCall, QC steps were performed using PLINK 1.9 (variants with minor allele frequency (MAF) < 5%, call rate < 0.99 or Hardy-Weinberg equilibrium  $\chi^2$  test p-value<0.0001). Genotype data were phased using the Eagle<sup>13</sup> and imputed to the Haplotype Reference Consortium reference panel using the Michigan Imputation Server<sup>14</sup>. After imputation, genetic variants were filtered for imputation quality  $R^2 > 0.4$ . GSA genotype data was combined with WES data using PLINK 1.9. Variants with a MAF < 5% were removed. In total, the combination of GSA and exome data covered 7,798,353 variants for 397 patients with IBD (CD =234 and UC=166) and 218 Lifelines Deep individuals.

### **Prediction of IBD based on metabolomics profiles**

We used CoDaCoRe<sup>15</sup> (v 0.0.1) to identify ratios of metabolites and bacterial abundances that could predict IBD and its sub-phenotypes. Patients with a history of intestinal surgeries (n = 136) were excluded, and only highly prevalent metabolites (>70% of the samples) were considered in this analysis. Here, we first split the data into a training and a test set for each prediction, using 75% of the samples in the training process. Next, we estimated the added predictive value of using ratios of metabolites compared to a model including only host age, sex, BMI and faecal calprotectin levels (calprotectin levels >200  $\mu\text{g/g}$ , yes/no). Furthermore, we tested if the ratio of metabolite identified to discriminate between the samples from IBD and non-IBD participants had a predictive value in a group of less severe patients. Patients with a less severe IBD were defined as participants with calprotectin <200 $\mu\text{g/g}$  and SSCAI <2.5 or Harvey Bradshaw<5 at the time of sampling and no records of active disease periods 1 year prior and 1 year post sample collection.

Next, we explored the levels of the predictive metabolites in a separate cohort of samples from the Human Microbiome Project 2 (HMP2)<sup>9</sup>. Data was obtained through the Metabolomics Workbench portal (<https://www.metabolomicsworkbench.org>). Due

to the differences in metabolomic platforms and metabolite annotation libraries, we encountered some challenges in aligning the metabolites found in our study with those reported in the HMP2 cohort. For example, lactosyl-N-palmitoyl-sphingosine was not annotated in their data, but we identified a structurally similar molecule, N-palmitoyl-sphingosine or Cer(d18:1/16:0), which showed a strong correlation with lactosyl-N-palmitoyl-sphingosine levels in our dataset (Spearman correlation,  $\rho = 0.70$ ). To further validate our findings, we compared the ratio Cer(d18:1/16:0) / L-urobilin between IBD and non-IBD samples at each time point where at least 5 non-IBD samples were available.

### **Co-occurrence patterns between bacteria and metabolites**

The *QIIME*<sup>6</sup> implementation of *mmvec* v.1.0.6<sup>17</sup> was used to estimate the co-occurrence probabilities between highly prevalent metabolites and bacteria (**Suppl. Table 21**).

Furthermore, we assessed the associations between individual microbiome features (taxa, gene clusters and metabolic pathways) and metabolites using regression models. The association between metabolites levels and bacterial taxa were assessed using two different models: firstly, recoding bacteria as detected or undetected (1 and 0) and secondly, considering only non-zero abundance value. For bacterial pathways and gene clusters only the second approach was used. In addition to the previously mentioned confounders (age, sex, BMI, sample storage time, batch, amount of faecal material, estimate bowel movements a day and intestinal resections), dysbiosis (yes/no) and disease phenotype (CD, UC, non-IBD) were also included as covariates in the model. Finally, we additionally tested context-specific effects by adding an interaction factor between microbial features and dysbiosis as predictor in the model.

### **Association between metabolites and phenotypes**

An association analysis between phenotypes and metabolites was performed within each cohort (controls, CD and UC). We included information about lifestyle, including use of 31 different types of medication, dietary patterns represented by 144 food frequency-related scores and the levels of 3 faecal biomarkers (faecal calprotectin, chromogranin A and human beta-defensin) (see *Phenotypes selection* section). Each

phenotype–metabolite combination was tested using linear regression, including age, sex, BMI, bowel movements per day and technical factors as covariates.

### Mediation analysis

To establish if associations between phenotypes and metabolites could be related to the intestinal microbiota, we performed a mediation analysis in each cohort (CD, UC and controls). Phenotypes were considered exposures and metabolites outcomes. For each phenotype with at least one significant association with a metabolite ( $FDR < 0.05$ ) we first selected the potential mediators by correlating the phenotype with bacterial abundances. Exposures, mediators, and outcomes were standardized prior to analysis and the impact of confounders (age, sex, BMI, estimate bowel movements a day, sample storage time (month), batch, sequencing read depth) was regressed in both mediators and outcomes. Because multiple bacteria can mediate in the same phenotype-metabolite relation, we used the *regmed* R's package (v. 2.0.5) to perform a regularized mediation analysis. This approach allows the input of multiple features as mediators, selecting the most relevant factors in the exposure-outcome relation. Additionally, for each mediated association, we also estimated the proportion of mediated effects using the *mediation* (v. 4.5) package in R.

### Differential abundance analyses of faecal microbiome features

Linear regression analysis was used to identify microbiome features (taxa, pathways and metabolic gene clusters, **Suppl. Table 22**) that differed between controls and IBD. Age, sex, BMI, average bowel movements per day, history of intestinal resections (yes/no) and sequencing read depth were included as covariates in the regression models.

### Metabolite levels prediction

For each of the metabolites and in each of the 8 defined models, we performed a 5-fold cross-validation (CV) procedure to select the best set of predictors based on the mean of squared errors. A 10-fold CV step was used in each of the CV-training sets to tune the lasso penalty parameter ( $\lambda$ ) in the lasso regression. Using the estimates of the model minimising the mean of squared errors, we computed the  $R^2$

coefficient in the whole dataset. We defined 8 different models representing different data categories available in both cohorts (IBD and non-IBD samples).

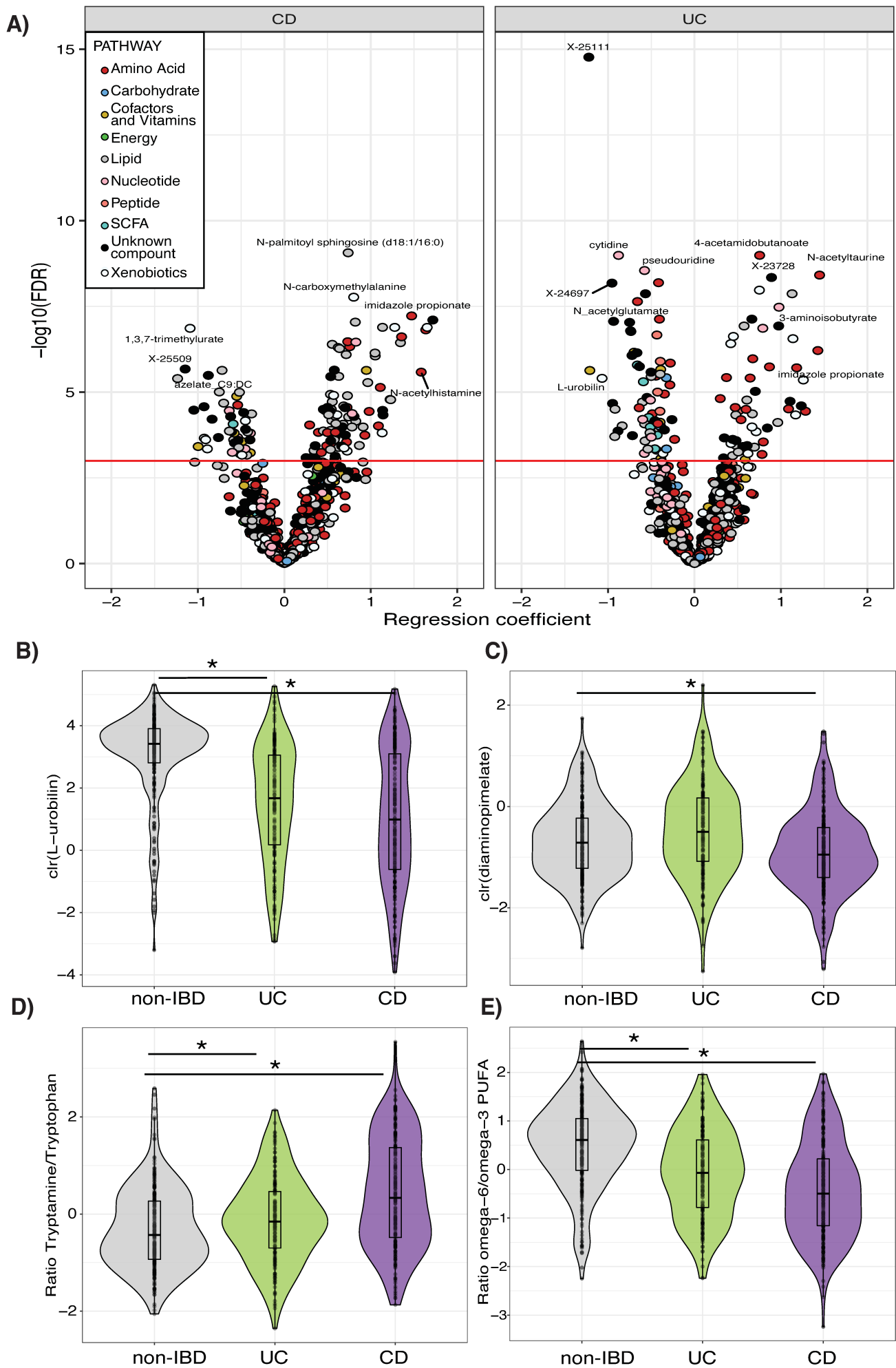
- 1) Host and technical factors: Which included information about the sex, age, BMI, average bowel movements per day, storage time at -80°C, batch and amount in grams of sample used for measuring metabolomics. All other models also included these variables to consider confounders' effects.
- 2) Diet: 119 dietary food patterns adjusted by total caloric intake.
- 3) Biomarkers: The levels of chromogranin A, human-beta defensin 2 and faecal calprotectin levels above 200 (yes/no).
- 4) Medication: The use of 22 medication categories (yes/no).
- 5) Disease: IBD (yes/no)
- 6) Taxa abundance: Relative abundance of 109 microbial species.
- 7) Bacterial pathways: 326 MetaCyc pathways
- 8) All: A model containing all variables described in the previous model.

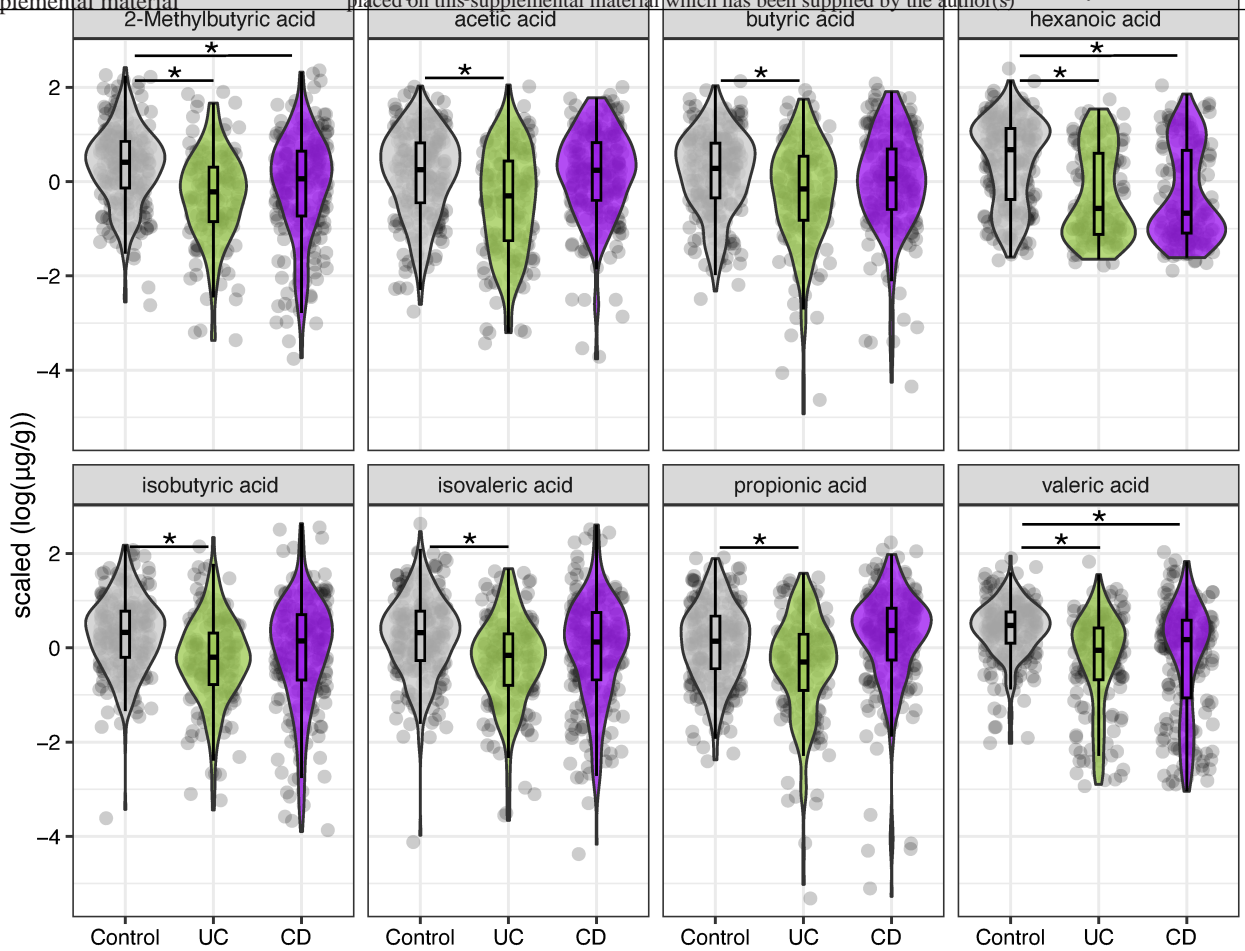
**REFERENCES**

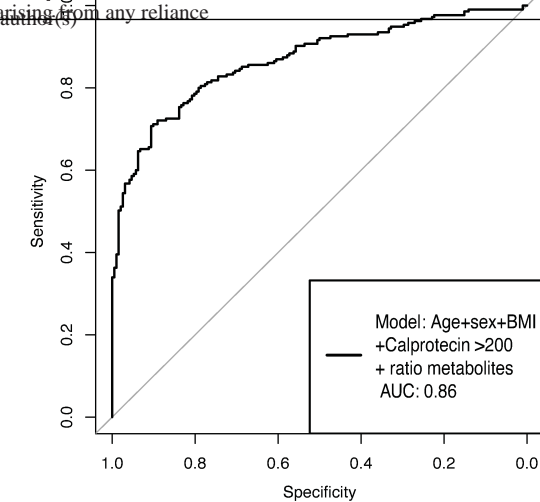
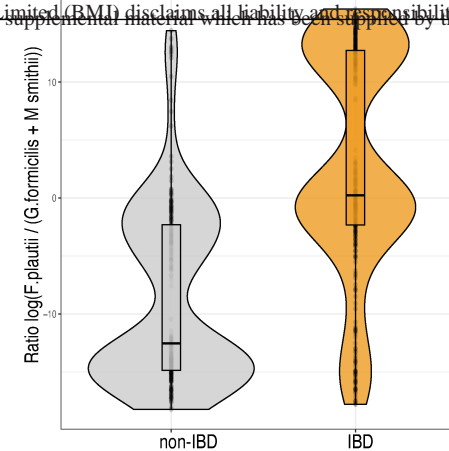
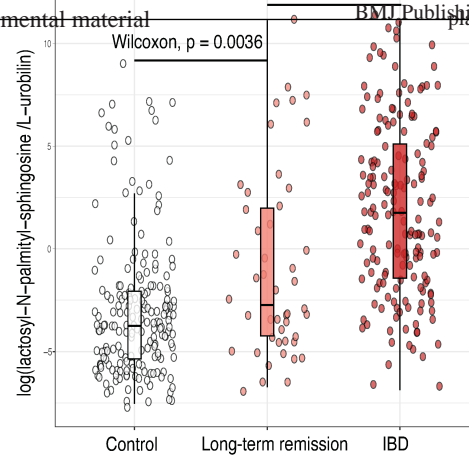
1. Klaassen, M. A. Y. *et al.* Anti-inflammatory Gut Microbial Pathways Are Decreased During Crohn's Disease Exacerbations. *Journal of Crohn's & colitis* (2019) doi:10.1093/ecco-jcc/jjz077.
2. Bolte, L. A. *et al.* Long-term dietary patterns are associated with pro-inflammatory and anti-inflammatory features of the gut microbiome. *Gut* **70**, 1287–1298 (2021).
3. Siebelink, E., Geelen, A. & Vries, J. H. M. de. Self-reported energy intake by FFQ compared with actual energy intake to maintain body weight in 516 adults. *British Journal of Nutrition* **106**, 274–281 (2011).
4. Willett, W. C., Howe, G. R. & Kushi, L. H. Adjustment for total energy intake in epidemiologic studies. *The American Journal of Clinical Nutrition* **65**, 1220S-1228S (1997).
5. Beghini, F. *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife* **10**, e65088 (2021).
6. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biology* **20**, 257 (2019).
7. Andreu, V. P. *et al.* A systematic analysis of metabolic pathways in the human gut microbiota. *bioRxiv* 2021.02.25.432841 (2021) doi:10.1101/2021.02.25.432841.
8. Andreu, V. P. *et al.* BiG-MAP: an Automated Pipeline To Profile Metabolic Gene Cluster Abundance and Expression in Microbiomes. *mSystems* (2021) doi:10.1128/mSystems.00937-21.
9. Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655 (2019).
10. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
11. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

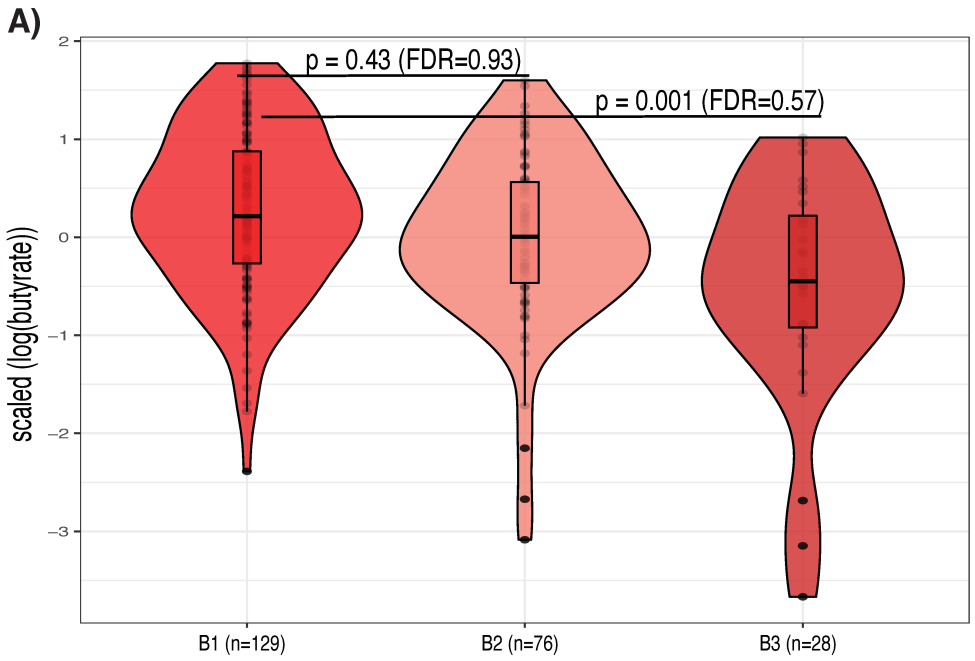
12. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* **81**, 559–575 (2007).
13. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* **48**, 1443–1448 (2016).
14. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284–1287 (2016).
15. Gordon-Rodriguez, E., Quinn, T. P. & Cunningham, J. P. Learning sparse log-ratios for high-throughput sequencing data. *Bioinformatics* **38**, 157–163 (2022).
16. Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* **37**, 852–857 (2019).
17. Morton, J. T. *et al.* Learning representations of microbe–metabolite interactions. *Nat Methods* 1–9 (2019) doi:10.1038/s41592-019-0616-3.



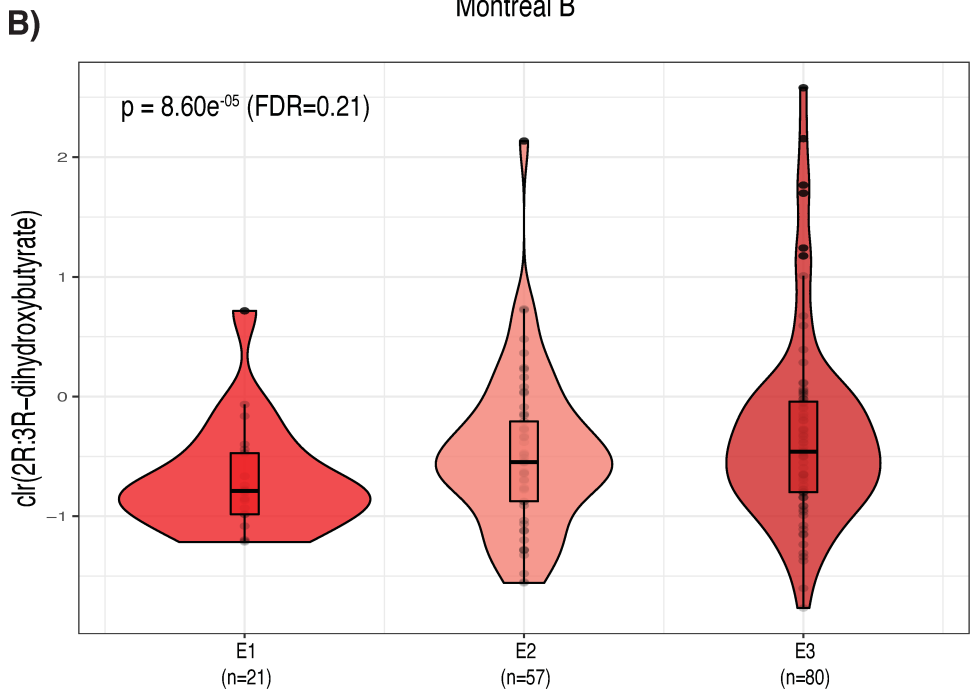








## Montreal B



## Montreal E

