Poly-omic risk scores predict inflammatory bowel disease diagnosis

Christopher H. Arehart,[1,2,3,*] John D. Sterrett,[1,4,*] Rosanna L. Garris,[1,5] Ruth E. Quispe-Pilco,[1,2] Christopher R. Gignoux,[6] Luke M. Evans,[2,3,†] Maggie A. Stanislawski[6,†]

Affiliations:

1. Interdisciplinary Quantitative Biology PhD Program, University of Colorado, Boulder, CO
2. Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO
3. Institute for Behavioral Genetics, University of Colorado, Boulder, CO, USA
4. Department of Integrative Physiology, University of Colorado, Boulder, CO
5. Department of Biochemistry, University of Colorado, Boulder, CO
6. Department of Biomedical Informatics, School of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO

***** These authors contributed equally to this work
† Joint senior authors

Abstract

Inflammatory Bowel Disease (IBD) is characterized by complex etiology and a disrupted gut microbiota. The substantial non-genetic variance for Crohn's disease and ulcerative colitis (≥25%) suggests that both genetic and environmental factors contribute to IBD development. We aim to extend the framework of genomic studies by examining gut characteristics that are affected by both genetic and environmental factors. Specifically, we train models and validate their accuracy using data that quantifies the microbiota, their transcripts, and the metabolites present in the gut. The IBD Multi-omics Database from the Human Microbiome Project 2 provided 1,785 repeated samples for 131 individuals (103 cases, 27 controls) across multiple -omics layers including metagenomics, metatranscriptomics, viromics, and metabolomics. After splitting the subjects into training and validation groups, we used mixed effects least absolute shrinkage and selection operator (LASSO) regression to determine the most relevant features for each -omic layer. These features, along with demographic covariates, were incorporated into a polygenic risk score framework to generate four separate -omic-level prediction scores. All four -omic-level scores were then combined into a final regression to assess the relative importance of individual -omics and the added benefits when considered together. Our models identified several species, pathways, and metabolites known to be associated with IBD risk. Individually, metabolomics and viromics based scores were more predictive than metagenomics or metatranscriptomics based scores, and when all four scores were combined, we predicted disease diagnosis with a Nagelkerke's R^2 of 0.46 and an AUC of 0.80 [95% CI: 0.63, 0.98].

Importance

The health burden of inflammatory bowel disease (IBD) among affected individuals is large, and its complex etiology has been studied using high throughput -omics technology. We applied a prediction framework across multiple -omics from the gut microbiome (metagenomics, metatranscriptomics, metabolomics, and viromics) to predict diagnoses of Crohn's disease and ulcerative colitis. The predicted scores from our models illustrated key features and allowed us to compare the relative utility of each -omic data type when used individually versus when combined in a multi-omics model. The individual and combined models performed well and emphasized the importance of metabolomics and viromics over metagenomics and metatranscriptomics predictive scores for IBD. The better predictive capability of metabolomics and viromics is likely because these -omics also serve as markers of lifestyle factors such as

diet. This study shows the utility of combining multiple -omic data types to disentangle complex disease etiologies and biological signatures.

Introduction

Inflammatory Bowel Disease (IBD) is characterized by complex etiology and contains multiple pathological subtypes, including Crohn's disease (CD) and ulcerative colitis (UC). Both of these subtypes have wide-ranging heritability estimates and involve disruptions of the gut mucosa as well as dysbiotic gut microbiota. Despite clustering of these diseases within family trees, genome-wide association studies have produced much smaller SNP-heritability estimates (0.37 for CD and 0.27 for UC)[1] than twin studies, which estimate heritability coefficients of 0.75 for CD and 0.67 for UC.[2] IBD affects individuals worldwide and has an incidence of 19.2 per 100,000 person-years in North America,[3] and age-adjusted prevalence of 0.40% and 0.65% for CD and UC respectively.[4] With impactful symptoms including diarrhea, abdominal pain, bloody stools, weight loss, and fatigue, the rising rates of CD and IBD are a salient public health concern.[5]

In genome-wide association studies to date, more than 120 related genes have been identified for CD, 67 of which were found to be differentially expressed for both CD and UC patients compared to non-IBD individuals.[6] These biological underpinnings have connected IBD to an array of comorbidities including asthma[7] and diabetes,[8] which are also characterized by polygenic architectures and complex environmental interactions related to immune activation. Additionally, many of the genes most strongly associated with IBD diagnosis are active in host-microbe interactions (such as pathogen-associated molecular pattern recognition, inflammatory responses, and phagocytic processes), including toll-like receptor (TLR) 4, TLR9, nucleotide binding oligomerization domain containing 2 (NOD2), interleukin-23 receptor, and tumor necrosis factor.[6] Multiple environmental risk factors for IBD have been identified, including smoking, excess body fat,[9] urban living,[10] antibiotic exposure,[10] and adverse childhood events.[11] On the other hand, studies have highlighted protective environmental factors including *Helicobacter pylori* infection,[12] breastfeeding,[9] and adequate vitamin D status.[10] Notably, many of these environmental factors are associated with changes to an individual's gut microbiome, its activity, the host's metabolome, and immune function.[13,14]

Our goals for this study were two-fold: (1) to further elucidate microbiome-related factors associated with IBD diagnosis, and (2) to apply a polygenic risk score-like framework[15] across multiple -omics data types to evaluate the performance of a combined poly-omic risk score model. To accomplish both of these aims, we reached beyond genetic data and genome-wide association studies to further uncover risk factors for IBD. This included microorganisms (both bacteria and viruses), RNA transcripts of the microbiome, and small molecules (metabolites) present in the gut. Whereas most studies including -omics analyses use univariate differential abundance methods (examining one feature at a time), our goal was to create an additive and predictive multivariable model that would incorporate features from multiple -omic data types. Given that many -omics data types are compositional,[16,17] meaning that the data are proportions or relative abundances, a decrease in one feature will correspond to an increase in others. Thus, multivariable differential abundance analysis or models for predicting disease status based on -omics data (as will be described later in this paper) may provide a unique alternative to univariate differential abundance methods.

Analogous methods to polygenic risk scores[15] have begun to target sources of biological variation beyond genotypes. A recent study in the UCLA Health biobank found that methylation based risk scores were considerably more accurate than their genetic-variant-based counterpart when imputing diagnoses and lab tests in the electronic health record.[18] This recent pairing of machine learning methods and biological data from multiple -omics levels has opened the door to better informed models that may have greater potential for disease prevention and

personalized medicine applications. The microbiome, transcriptome, virome, and metabolome, are all dynamic through time for individuals, which differs from genetic variants, which remain constant from birth. Therefore, a multi-omics-based prediction appears to be a promising approach in the context of complex and dynamic disease etiologies. Although high throughput technologies and data generation are on the rise, there are still common issues such as overfitting to training datasets, low interpretability, and the general lack of portability from one context to another.[19] We aim to address these shortcomings through generalizable and interpretable modeling methods. In the present analysis, we designed a modular multi-omics framework that allows researchers to assess the relative contributions of each -omics data type individually and then combine them together to assess overall predictive capability.

Methods

Dataset. The Human Microbiome Project 2[20] (HMP2) IBD multi-omics database (https://ibdmdb.org/tunnel/public/summary.html) contains 1,785 unique samples collected from 131 participants across 5 study sites, with metadata including each sample's participant ID, sex, race, antibiotic use, and site. Each of these samples was analyzed across up to eight different -omics layers. In the present study, we used feature count tables for samples across four of these -omics data types: metagenomics (MGN), metatranscriptomics (MTS), viromics (VRM), and metabolomics (MBL). For clarification, the term "sample" is used in this paper to denote one fecal specimen, which could have corresponding MGN, MTS, VRM, and MBL data, whereas "participant" is used to denote an individual who may have multiple samples throughout the duration of the study. Count tables for MGN and VRM were generated with MetaPhlAn2 and VirMap, respectively. MTS count tables were generated by summing pathway abundances as mapped by HUMAnN2. For MBL, four column methods (Carbon 18 (C18)-positive, C18-negative, Hydrophilic interaction liquid chromatography (HILIC)-positive, and HILIC-negative) coupled with mass spectrometry were used to isolate and detect metabolites. Only compounds successfully annotated with molecule names were included in analyses.

Data Processing and Thresholding. Compositional datasets (MGN, MTS, VRM) were normalized with center-log transformation, and MBL was normalized using a $\log_{10}$ transformation. We then removed highly sparse features (found in fewer than 5% of samples) from VRM. For MBL, we only included compounds present in >99% of samples. The difference in methods used for these two data types is due to variation in sparsity of the datasets, as the majority of viruses were found in very few samples, whereas the majority of compounds from MBL were found in most samples. Additionally, we removed one of any two highly collinear features (Pearson's $\varrho > 0.95$) from MBL and MTS datasets at random. After normalization, features with a standard deviation less than 1 were excluded from MGN and MTS, and features with a standard deviation less than 0.1 were excluded from VRM and MBL. The standard deviation threshold for each -omics data type was chosen based on a histogram of sample variation in the dataset and served to eliminate features with minimal differences across samples.

Training and Validation Split. Represented in our sample set were 130 distinct individuals (one participant was excluded due to missing data). To avoid overfitting, samples from 30 participants (23% of individuals) were set aside to serve as a validation set for performing model predictions and evaluating the resultant accuracy, as well as for use in our multi-omic model. Because the multi-omic model required all four data types, the 30 participants who had the most samples with all four -omics data types and the fewest samples with missing -omics were chosen for the validation set (Supplementary Figure 1). We ensured that no samples from these individuals were used in the training process for any of the four -omics layers. Any samples in the validation

set with missing -omics were excluded from the multi-omic model. No aspects of the experimental design indicate bias for how many samples from a participant contain all four -omics (based on any qualities of that participant), so this is assumed to represent a non-biased subset of the participants to the best of the researchers' knowledge from all available information on sample collection. The remaining training set contained 100 participants with a varying number of samples across layers. Of these 30 participants reserved for validation, 15 (499 samples) had CD, 3 (102 samples) had ulcerative colitis, and 12 (427 samples) were non-IBD controls. This left 100 participants for training. Of these 100 training participants, 50 (1,242 samples) had CD, 35 (932 samples) had UC, and 15 (520 samples) were non-IBD controls. Supplementary Table 1 shows the breakdown of training and validation samples within each -omic data type.

Individual -omic models. Features for each -omics data type were selected by applying mixed effects least absolute shrinkage and selection operator (mixed effects LASSO)[21] logistic regression using the glmmLasso function in the R package glmmLasso V1.5.1[22] using the following model, where y denotes IBD diagnosis, μ denotes the intercept, β_i denotes the effect sizes, *n* denotes the number of -omic features, *(1|x)* indicates a random effect for variable *x*, and ϵ denotes the residuals:

$$y = \mu + \sum_{i=1}^{n} \beta_i * Feature_i + \beta_{n+1} * Age + \beta_{n+2} * Sex + \beta_{n+3} * Race + \beta_{n+4} * Antibiotic_Use + (1|site) + (1|Participant_ID) + \epsilon \qquad (eqn.\ 1)$$

The mixed effects LASSO regression was run across a grid of lambda values, which control how parsimonious the models are (via penalizing additional features). A final LASSO regression was run using the lambda value at the elbow of the number of features retained by LASSO (Supplementary Figure 2). Using only the LASSO selected features, we generated a logistic mixed effects model (R package lme4 V1.1-29)[23] following eqn. 1 and requiring that the important covariates of age, sex, race, antibiotice use, site, and participant ID were included. Thus, feature weights were estimated in a context that accounted for these potentially confounding variables. Analogous to the methods for computing polygenic risk scores from a genome-wide association study,[24] the *n* feature weights (excluding participant demographic information) were then multiplied by each ith samples' feature value and summed, i.e., for sample i:

$$score_i = \beta_1 * feature_{1,i} + \beta_2 * feature_{2,i} + \ldots + \beta_n * feature_{n,i} \qquad (eqn.\ 2)$$

These raw scores were standardized to have a mean of zero and a standard deviation of one.

Baseline Models. We fit a baseline risk score model for each -omics data type based on only metadata for each participant using the lmer function via the lme4 R package where terms are the same as described above:

$$y = \mu + \beta_1 * Age + \beta_2 * Sex + \beta_3 * Race + \beta_4 * Antibiotic_Use + (1|site) + (1|Participant_ID) + \epsilon \qquad (eqn.\ 3)$$

We trained these models on participants in the training dataset and then used them to predict the diagnosis of participants in the validation dataset.

Assessing Model Validity. Within each -omic data type, predicted risk scores for each individual were averaged across their longitudinal samples in the validation set (see Supplementary Figure 3). Area under the curve (AUC) values were calculated from receiver operating characteristic curves (ROC) for each model (both baseline and feature) across -omics layers. Odds ratios (ORs) were also calculated for each of the predicted scores, and 95% confidence intervals were generated for each OR.

Multi-omic models. Prediction scores generated via feature weights within each -omics layer were averaged across samples for each 'testing' individual and then combined in the following logistic regression model:

$$y = \mu + \beta_1 * MGN + \beta_2 * MTS + \beta_3 * VRM + \beta_4 * MBL + \beta_5 * Age + \beta_6 * Sex + \epsilon \qquad (eqn.\ 4)$$

Age and sex were the two covariates included in this model due to little or no variability across the validation samples for other variables such as race or antibiotic use. Nagelkerke's R^2 was used to compare the combined logistic model to a baseline model only based on age and sex. Code for these methods is made publicly available at the following repository: https://github.com/sterrettJD/poly-omics-risk.

Results

Included Participants. This study included 130 participants, of whom, 27 did not have IBD, 65 had CD, and 38 had UC. Male and female participants were represented similarly, and the mean age of participants was 28 years old, with a standard deviation of 17 years and a range of 6 to 76 years of age. The majority of participants were white and did not use antibiotics. Table 1 elaborates with descriptive statistics of each group.

Table 1. Descriptive statistics of participants in the study. Categorical data are presented as "N (percent)", and numerical data are presented as "mean (standard deviation)". Abbreviations: SD, standard deviation, MGH, Massachusetts General Hospital.

	Healthy Control (N=27)	**Crohn's Disease (N=65)**	**Ulcerative Colitis (N=38)**	**Total (N=130)**
Sex				
Female	12 (44.4%)	32 (49.2%)	20 (52.6%)	64 (49.2%)
Male	15 (55.6%)	33 (50.8%)	18 (47.4%)	66 (50.8%)
Age				
Mean (SD)	29 (± 20)	26 (± 16)	29 (± 17)	28 (± 17)
Antibiotics				
No	27 (100%)	51 (78.5%)	33 (86.8%)	111 (85.4%)
Yes	0 (0%)	14 (21.5%)	5 (13.2%)	19 (14.6%)
Site				
Cedars-Sinai	1 (3.7%)	20 (30.8%)	12 (31.6%)	33 (25.4%)

Cincinnati	9 (33.3%)	17 (26.2%)	7 (18.4%)	33 (25.4%)
MGH	13 (48.1%)	13 (20.0%)	11 (28.9%)	37 (28.5%)
MGH Pediatrics	3 (11.1%)	8 (12.3%)	5 (13.2%)	16 (12.3%)
Emory	1 (3.7%)	7 (10.8%)	3 (7.9%)	11 (8.5%)
Race				
American Indian or Alaska Native	0 (0%)	1 (1.5%)	0 (0%)	1 (0.8%)
Black or African American	1 (3.7%)	3 (4.6%)	6 (15.8%)	10 (7.7%)
More than one race	1 (3.7%)	3 (4.6%)	1 (2.6%)	5 (3.8%)
White	25 (92.6%)	56 (86.2%)	29 (76.3%)	110 (84.6%)
Other	0 (0%)	2 (3.1%)	2 (5.3%)	4 (3.1%)

Baseline Models. Baseline models, which only include patient demographic information and were trained on the training dataset, poorly predicted the actual diagnosis of our validation dataset, as can be seen in Supplementary Table 2 (AUC [95% CI], Nagelkerke's R^2 for MGN = 0.43 [0.19, 0.67], 0.08; for VRM = 0.47 [0.23, 0.71], 0.02; for MTS = 0.38 [0.15, 0.62], 0.04; for MBL = 0.47 [0.25, 0.69], 0.01). This established that diagnosis could not be discriminated solely by participant metadata and non-omics covariates as described in equation 3 and Supplementary Table 1.

Individual -omic risk scores. Compared to the baseline models, the predicted risk scores (using feature weights derived from the training set) for each -omic data type (in addition to the baseline covariates of age and sex) demonstrated better predictive capability on the validation data as is shown in Figure 1 (AUC [95% CI], Nagelkerke's R^2 for MGN = 0.66 [0.44, 0.87], 0.12; for VRM = 0.83 [0.68, 0.98], 0.43; for MTS = 0.73 [0.53, 0.92], 0.20; for MBL = 0.82 [0.66, 0.98], 0.40). A similar plot showing AUCs and ORs when only using the -omic-derived scores to predict diagnosis is shown in Supplementary Figure 4 with similar results, suggesting that the covariates of age and sex are not driving this predictive accuracy (AUC [95% CI], Nagelkerke's R2 for MGN =0.70 [0.50, 0.91], 0.04; for VRM = 0.80 [0.63, 0.96], 0.31; for MTS = 0.64 [0.42, 0.86], 0.12; for MBL = 0.78 [0.59, 0.98], 0.32). Of the individual -omics models, VRM and MBL had the best predictive capability with little to no interquartile overlap between cases and controls. The VRM and MBL predicted scores had significant odds ratios of 9.28 [1.53, 56.13] (p = 0.02) and 4.30 [1.32, 13.98] (p = 0.02) respectively, while MTS had an odds ratio of 2.89 [0.62, 13.52] (p = 0.2), and MGN had the lowest odds ratio of 1.24 [0.50, 3.09] (p =0.6).

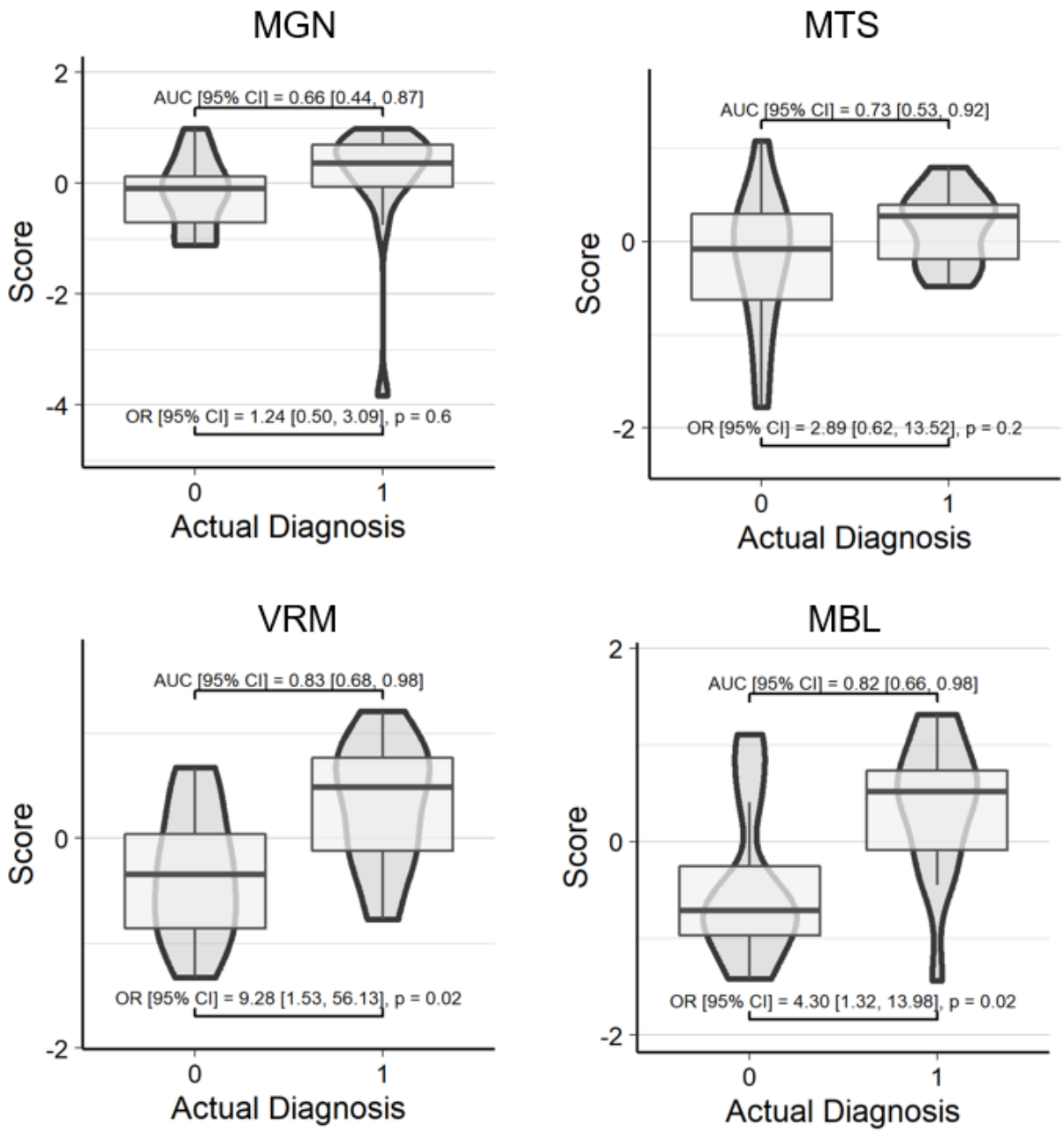

Figure 1. Risk scores predict IBD diagnosis. Z-score transformed risk scores (averaged for each participant) on the y axis are plotted against actual diagnosis on the x axis for the validation dataset. AUC and OR were calculated with basic covariates (diagnosis ~ score + age + sex). Each of the four scores shown were calculated using feature weights from a LASSO-identified mixed effect logistic regression trained in a separate set of samples/individuals: diagnosis ~ features + age + sex + race + antibiotic use + (1|site) + (1|participant ID). An actual diagnosis value of 1 indicates presence of IBD. Abbreviations: MGN, metagenomics; MTS, metatranscriptomics; VRM, viromics; MBL, metabolomics; AUC, area under the receiver operating characteristic curve; OR, odds ratio; CI, confidence interval; IBD, inflammatory bowel disease.

Selected features. Within each individual -omic model, LASSO selected between 6 and 22 features. Figure 2 shows the feature weights for each -omics data type. For MGN, *Megasphaera* sp DISK18 had the strongest negative weight (indicating association with low risk score). Additional species associated with lower risk scores included *Parabacteroides goldsteinii*, *Methanobrevibacter smithii*, *Roseburia hominis*, and *Akkermansia muciniphila*. *Firmicutes* CAG 83 was the only species identified with a positive weight. The MTS model selected 22 pathways. Of these, the strongest negative weights (associated with a lower risk score) were for the 2-methyl citrate cycle I and the superpathway of sulfur amino acid biosynthesis identified from *Saccharomyces cerevisiae*, and the strongest negative weights were for the pentose phosphate pathway and the methylerythritol phosphate pathway I. Six viruses were selected by the VRM model, four of which were positively correlated with risk score, indicating an association with IBD. The strongest positive weight was for cucumber green mottle mosaic virus and the strongest negative weight was a gokushovirus. The MBL model selected 14 metabolites. The strongest positive weights were for docosapentaenoate (identified by the C18 column) and phenyllactate and strongest negative weights were for eicosatrienoate and docosapentaenoate (identified by the HILIC column).

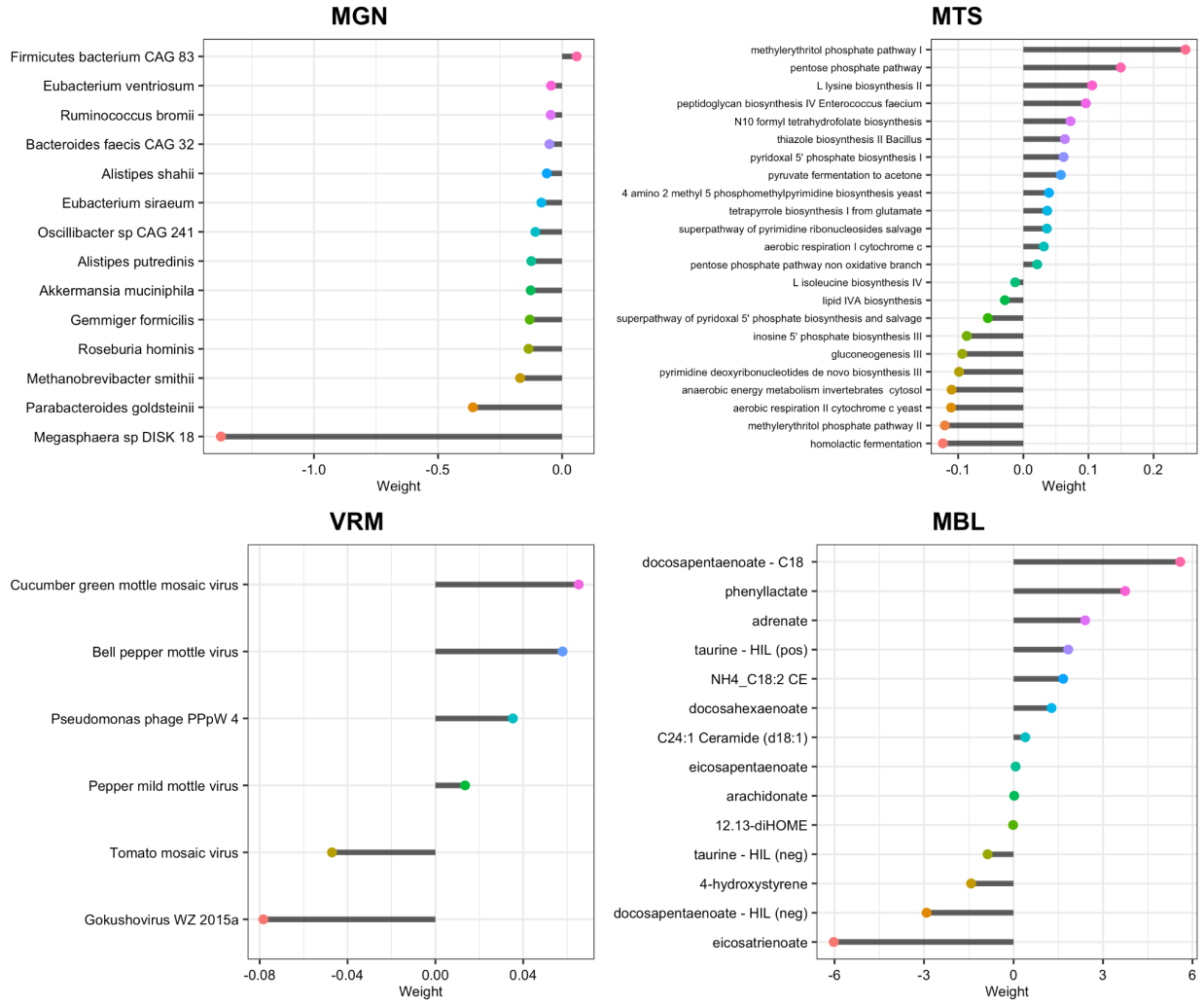

Figure 2. Feature weights for each risk score model. The effect sizes described by the x-axis were multiplied by transformed abundances and then summed to generate each -omic's score. A negative weight corresponds to lower predicted risk of IBD, whereas a positive weight would confer higher predicted risk of IBD. Abbreviations: MGN, metagenomics; MTS, metatranscriptomics; VRM, viromics; MBL, metabolomics; IBD, inflammatory bowel disease; HIL, hydrophilic interaction liquid chromatography.

Combined -omics Model. Figure 3a illustrates the odds ratios of all -omics scores in a combined framework for the 30 individuals reserved for model validation. The covariates of age and sex were only slightly predictive of IBD on their own (Nagelkerke Pseudo R^2 = 0.11); however, a multiple regression including age, sex, and all four scores produced a Nagelkerke R^2= 0.46 and an AUC of 0.80 (Figure 3c). The variance explained by the four scores without age and sex was slightly lower (R^2= 0.37), showing the importance of including demographic covariates and the predictive capacity of our combined framework to explain variance among IBD diagnoses. Despite the 95% confidence interval for all four scores' ORs overlapping with one, VRM and MBL stood out with larger values than MGN or MTS. The correlation between scores (Figure 3b), illustrates some similarity between VRM, MGN, and MBL, in addition to MTS and MBL. Analysis for a leave-one–omic-out approach is shown in Supplementary Figure 5 to further demonstrate the increased odds ratios (near the threshold for significance) for VRM and MBL compared to MGN and MTS.

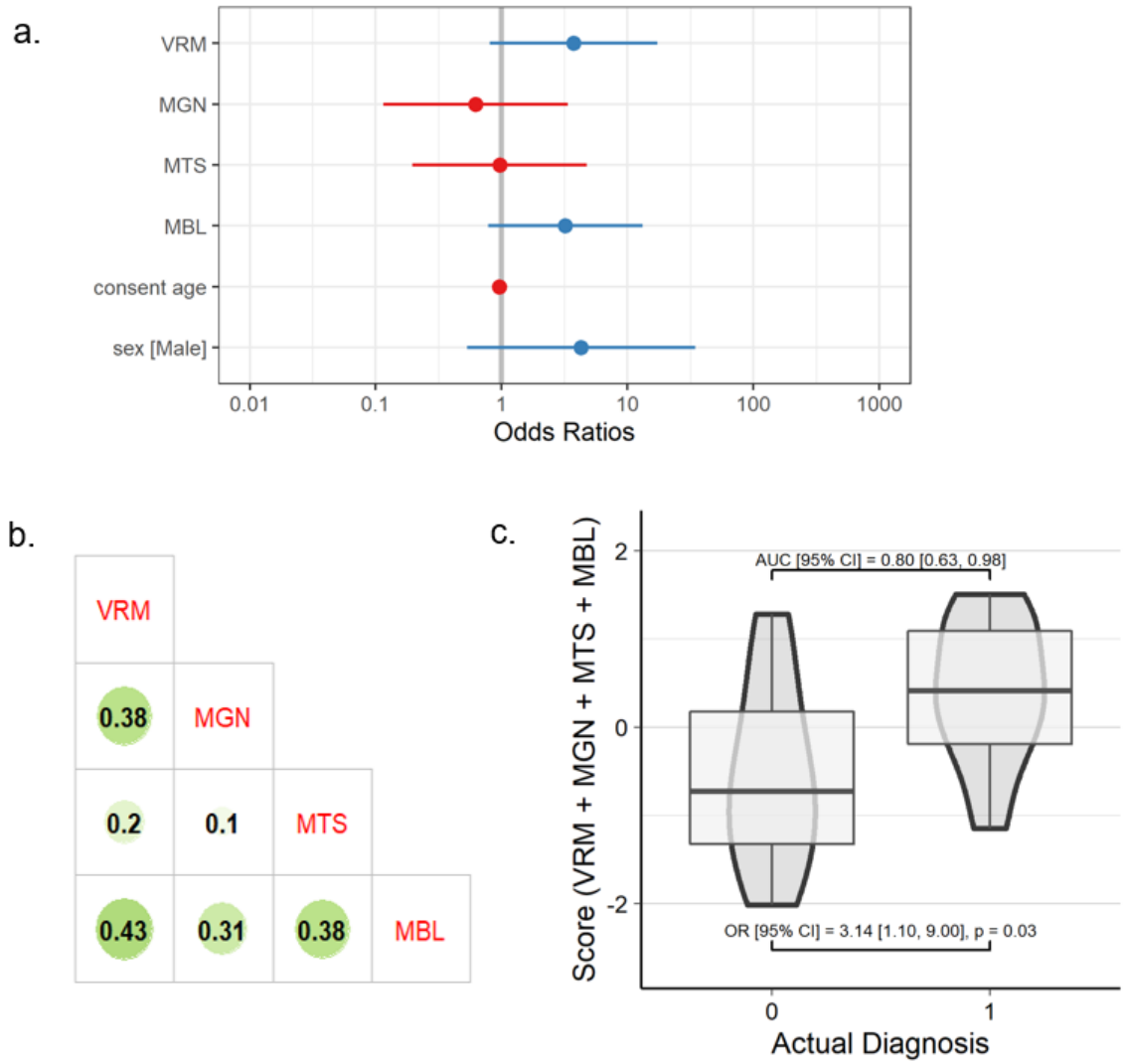

Figure 3. Results of multi-omic modeling. (a) Odds ratios and 95% confidence intervals of the predicted score (b) pearson correlation matrix between -omics scores, and (c) standardized risk scores (summed across all -omics) on the y-axis plotted against actual diagnosis on the x-axis for the validation dataset, where AUC and OR were calculated with basic covariates (diagnosis ~ score + age + sex). In panel a, points represent the odds ratio for each -omic's predicted scores in the multi-omic regression, and lines represent 95% confidence intervals of the odds ratios. In panel b, the size and darkness of the circle represents the correlation between the predicted scores for each -omic data type. In panel c, an actual diagnosis of 1 represents a case for IBD. Abbreviations: MGN, metagenomics; MTS, metatranscriptomics; VRM, viromics; MBL, metabolomics; AUC, area under the receiver operating characteristic curve; OR, odds ratio; CI, confidence interval; IBD, inflammatory bowel disease.

Discussion

Overall, we successfully utilized a polygenic risk score framework across multiple -omics data types to predict IBD diagnosis. Other studies have taken similar approaches, such as single variable differential abundance testing across multiple -omics in IBD,[25] multivariate

analysis of multi-omic interactions in participants with IBD,[26] like composite of nonsupervised multivariate analysis and principal component analysis, or multi-omic risk scores for diseases other than IBD.[27] Our results often corroborate the findings of other multi-omic studies of individuals with IBD, and they also highlight groups of features that associate with IBD when considering the abundances of other features. Additionally, our methods allow for comparison of the predictive ability of different -omics for disease, highlighting that taxonomic profiling (from MGN) might not hold as much of a signature of IBD as other -omics that provide information on active transcription (MTS) and compound presence (MBL). Additionally, though VRM-predicted IBD scores correlated with MGN, VIR held a much stronger relationship with IBD, which could be a result of agricultural viruses serving as markers of diet.

Interpretation of features identified by individual models

Metagenomics. None of the 14 MGN features selected by LASSO had significant individual associations with IBD in the multiple regression, and all but one had a negative weight. These negative effect sizes correspond to reduced risk of IBD when utilized in the scoring framework, and the estimate with the largest magnitude (-1.37) was *Megasphaera sp. DISK 18* which is known for being an early colonizer of the oral microbiome.[28] The next largest negative effect size from *Parabacteroides goldsteinii* is of key interest given recent findings in mouse experiments that a *P. goldsteinii* probiotic may be useful in treating diet-induced obesity and type 2 diabetes.[29] The links between the microbiome, diet, and IBD are important to consider in the context in which environmental exposures may manifest a genetic predisposition to the onset of IBD.[8] Our findings of *Roseburia hominis* and *Bacteroides faecis CAG 32* associating with decreased risk of IBD is consistent with existing literature.[30] We are intrigued by the importance of two features belonging to the *Alistipes* genus given the emerging connections between *Alistipes* and gut dysbiosis.[31] Lastly, *Ruminococcus bromii*'s role in the metagenomics score is worth highlighting since *R. bromii* is known to support the growth of *R. gnavus*,[32] which is a species purported to be associated with increased risk of CD.[30,33]

Metatranscriptomics. The MTS model identified a strong negative weight for transcripts belonging to the 2-methyl citrate cycle, meaning this pathway was associated with lower risk of IBD. Notably, the 2-methyl citrate cycle is responsible for metabolizing propionic acid[34] (a short chain fatty acid produced by gut commensal organisms fermenting dietary fiber) into succinate, which can enter the tricarboxylic acid (TCA) cycle. Increased transcription of this pathway could be indicative of high propionic acid in the gut, which has previously been demonstrated to be protective in the context of IBD, as propionic acid has anti-inflammatory effects through the inhibition of nuclear factor κB (NF-κB).[35] Alternatively, lower propionic acid could reflect decreased dietary fiber in participants with IBD. This result also aligns with previous analysis[36] of the HMP2 dataset (using different methods). The MTS model additionally identified a strong negative weight for sulfur amino acid biosynthesis transcripts. Given that one signature of IBD is a microbial shift toward the catabolism of taurine and cysteine to produce hydrogen sulfide, this is consistent with previous literature as well.[37]

The MTS model identified the pentose phosphate and methylerythritol phosphate pathways as having the strongest positive weights. The pentose phosphate pathway is involved in the metabolism of C5 sugars, but not much literature exists explaining a connection between it and IBD, though one study identified an increase in the pentose phosphate pathway in ileal Crohn's disease.[38] Similarly, little literature exists about the potential role of the methylerythritol phosphate pathway in the context of IBD, though its intermediates are potent activators of human gamma delta T cells, which are the first line of mucosal defense.[39] Gamma delta T cells have been implicated in intestinal inflammation and IBD in both human and animal models across a multitude of studies, though it is unclear whether their effects are protective or not.[40]

Viromics. The VRM model identified six viruses associated (either positively or negatively) with IBD risk. Four of these, Cucumber green mottle mosaic virus, bell pepper mottle virus, pepper mild mottle virus and tomato mosaic virus are all in the *Tobamovirus* genus. These are highly persistent and transmissible positive-sense single stranded RNA viruses that are known to infect various crop species and are found worldwide. Though many crops have developed resistance to the damaging effects of these viruses they are still endemic in agricultural products and have been found to make up a large portion of the gut virome.[41,42] As such, presence of these viruses could be serving as a marker of food choice in these participants, given that individuals with IBD may choose to consume different foods due to their condition. Interestingly, only one of these tobamoviruses, tomato mosaic virus had a negative feature weight. The other two viruses identified, pseudomonas phage PPpW4 and gokushovirus WV 2015a are both bacteriophages. Pseudomonas phage PPpW4 parasitizes pseudomonas bacteria, which are associated with pneumonia and post-surgical infections, and have been studied for use in phage therapy.[43] Gokushoviruses, though ubiquitous throughout the environment, are largely uncharacterized.[44]

Metabolomics. The MBL model identified fourteen metabolites associated with IBD, the majority of which were fatty acids. Metabolites with positive weights were docosahexaenoate, NH_4-18:2 cholesterol ester, C24:1-ceramide (d18:1), adrenate, phenyllactate, eicosapenatenoate, and arachidonate. Adrenate and phenyllactate have been found in increased concentrations in individuals with IBD.[20] Metabolites with negative weights were hydroxystyrene, eicosatrienoate, and 12,13-dihydroxyoctadec-9-enoic acid (DiHOME). 12,13-diHOME was also reported as an important metabolite for differentiating IBD status when applying a knockoff filtering-based multivariate approach to data from HMP2.[45]

Our model selected two compounds annotated as taurine that had been isolated using different chromatographic columns (hydrophilic interaction liquid chromatography (HILIC) negative and HILIC positive), and interestingly, these compounds had contrasting positive and negative weights. This may be related to the charge (or other aspects) of the identified compounds, but more detailed investigation to follow up the untargeted metabolomics would be needed to differentiate these two. In accordance with this uncertainty, multiple studies have found conflicting associations between taurine in the gut metabolome and IBD status.[46] Taurine, being a sulfur-containing amino acid, could have very diverse roles in the gut metabolome given its microbial metabolism to hydrogen sulfide, which can have effects ranging from pathogen inhibition at moderate concentrations to direct irritation of the gut mucosa at high concentrations.[47] Additionally, taurine plays other roles in the gut metabolome, such as conjugating with cholic acid to form the bioactive secondary bile acid taurocholic acid. Our findings are consistent with some differential abundance testing results from the original HMP2 paper, as they found taurine and taurocholic acid to be differentially abundant across IBD status.[20]

Similarly, our model selected two compounds annotated as docosapentaenoate, though in this case one with a positive weight came from the C18 column (specialized for intermediate polarity compounds, such as free fatty acids), and one with a negative weight came from the negative ion mode hydrophilic interaction liquid chromatography (HILIC) negative mode (specialized for polar compounds). According to previous studies, Docosapentaenoate is an omega-3 fatty acid generally regarded to have anti-inflammatory effects.[48]

Comparing individual -omics models. Of the individual -omic models, MBL and VRM had the highest predictive capability, while MTS provided moderate prediction, and MGN provided the lowest predictive accuracy. One explanation for these results is that MBL and VRM provide information about both the microbiome and the host; the metabolites could have been produced by the host or their diet, and the viruses could also be diet-related given that four of the six

selected viruses belong to the genus *Tobamovirus* and infect crops. In contrast, MTS and MGN only provide information about the microbiome, not the host. However, MTS specifies data on active transcription, whereas MGN only provides information about taxonomy and functional capabilities. Notably, our MGN analysis only considered taxonomy, not functional potential (i.e., no MGN pathway abundance data). These results are consistent with other findings that the metabolome predicts phenotypes better than taxonomic profiles do.[49]

Combined model. The results of our combined model mirror the AUCs and ORs of the individual -omics models. In the combined model, MGN had the lowest OR, whereas MTS had a moderate OR, and MBL and VRM had the highest ORs. We were limited by our sample size to not be able to reserve a second bonafide validation set for the combined model, but the improved R^2 from a baseline model indicates predictive capability of this combined model.

Limitations. As with many studies of this kind, batch and site effects may affect generalizability of our model. High throughput -omics technology faces many limitations due to batch effects, and future work should consider techniques for combining batches in this framework, using batch correction techniques or incorporating random effects for batches.[50] Such work to combine data from different studies could increase the robustness of this framework and increase the generalizability of its diagnostic capability, which would increase the likelihood of clinical implementation for such diagnostic -omic modeling, which should be considered an end goal.

Additionally, the number of -omics features was far greater than the number of participants, though we were able to use multiple longitudinal samples per participant (accounted for via random effects for participant ID). We took steps to reduce the number of features, such as through the sparsity, collinearity, and variance filtering steps, as well as through the use of LASSO regression for feature selection. However, such steps could be causing us to unwittingly discard useful information. Importantly, our modeling did not consider interactions between features, either within or between -omics, which is a direction for future research.

We would have liked to perform cross-validation splits of the training and testing datasets to test the robustness of these results. However, without sufficient sample sizes overlapping between -omics layers we were limited to one training-testing split of the data. In addition, exploring non-linear machine learning techniques such as random forests may have yielded better results.

The currently available dataset from HMP2 does not allow public access of host genome data, which would have provided immense context to the “outer” -omics upon which our analysis focused, and future analysis incorporating host genome data would add to these results by allowing comparison of the outer -omics to standard polygenic risk scores for IBD. Additionally, no information was available regarding the existence of any relatives within the dataset, which could have confounded results, as our analyses assumed all participants were not related or living in the same location.

Conclusions

Not only did we successfully implement a poly-omic risk score framework across four -omics layers, each of our individual models identified features known to associate with IBD risk while providing new insights into features that may influence or be influenced by IBD. The majority of IBD models do not utilize multiple -omics layers, focusing instead on host genetics.[51] Those that utilize other information often focus on predicting flare-up and other clinical outcomes among individuals with IBD.[39] This study is unique in that it combines -omics layers to predict the presence of disease.

Acknowledgements

We would like to thank Kristin Powell and Stephanie Hoyt for their support for this project, as well as Tom Cech and the BioFrontiers Institute for their support of the Interdisciplinary Quantitative Biology program. We'd also like to thank Lukas Buecherl for providing manuscript feedback and the Human Microbiome Project 2 team for data collection, processing, and public access.

Funding

This project was funded by the National Science Foundation-sponsored Interdisciplinary Quantitative Biology PhD program, the Integrated Data Science (Int dS) Graduate Training Fellowship, and the William J. Freytag Fellowship.

References

1. Chen, G.-B. *et al.* Estimation and partitioning of (co)heritability of inflammatory bowel disease from GWAS and immunochip data. *Hum. Mol. Genet.* **23**, 4710–4720 (2014).
2. Gordon, H., Trier Moller, F., Andersen, V. & Harbord, M. Heritability in Inflammatory Bowel Disease: From the First Twin Study to Genome-Wide Association Studies. *Inflamm. Bowel Dis.* **21**, 1428–1434 (2015).
3. Molodecky, N. A. *et al.* Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. *Gastroenterology* **142**, 46-54.e42; quiz e30 (2012).
4. Xu, F., Carlson, S. A., Liu, Y. & Greenlund, K. J. Prevalence of Inflammatory Bowel Disease Among Medicare Fee-For-Service Beneficiaries — United States, 2001−2018. *MMWR Morb. Mortal. Wkly. Rep.* **70**, 698–701 (2021).
5. Alatab, S. *et al.* The global, regional, and national burden of inflammatory bowel disease in 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet Gastroenterol. Hepatol.* **5**, 17–30 (2020).
6. Garza-Hernandez, D. *et al.* A systematic review and functional bioinformatics analysis of genes associated with Crohn's disease identify more than 120 related genes. *BMC Genomics* **23**, 302 (2022).
7. Kuenzig, M. E. *et al.* Co-occurrence of Asthma and the Inflammatory Bowel Diseases: A Systematic Review and Meta-analysis. *Clin. Transl. Gastroenterol.* **9**, 188 (2018).
8. Benchimol, E. I. *et al.* Asthma, type 1 and type 2 diabetes mellitus, and inflammatory bowel disease amongst South Asian immigrants to Canada and their children: a population-based cohort study. *PloS One* **10**, e0123599 (2015).
9. Carreras-Torres, R., Ibáñez-Sanz, G., Obón-Santacana, M., Duell, E. J. & Moreno, V. Identifying environmental risk factors for inflammatory bowel diseases: a Mendelian randomization study. *Sci. Rep.* **10**, 19273 (2020).
10. Piovani, D. *et al.* Environmental Risk Factors for Inflammatory Bowel Diseases: An Umbrella Review of Meta-analyses. *Gastroenterology* **157**, 647-659.e4 (2019).
11. Thomann, A., Lis, S. & Reindl, W. P796 Adverse childhood events and psychiatric comorbidity in a single-centre IBD-cohort. *J. Crohns Colitis* **12**, S514–S514 (2018).
12. Yu, Y., Zhu, S., Li, P., Min, L. & Zhang, S. Helicobacter pylori infection and inflammatory bowel disease: a crosstalk between upper and lower digestive tract. *Cell Death Dis.* **9**, 961 (2018).
13. de Souza, H. S. P. & Fiocchi, C. Immunopathogenesis of IBD: current state of the art. *Nat. Rev. Gastroenterol. Hepatol.* **13**, 13–27 (2016).
14. Verstockt, B., Smith, K. G. & Lee, J. C. Genome-wide association studies in Crohn's disease: Past, present and future. *Clin. Transl. Immunol.* **7**, e1001 (2018).
15. Choi, S. W., Mak, T. S.-H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).
16. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.* **8**, 2224 (2017).
17. Quinn, T. P. *et al.* A field guide for the compositional analysis of any-omics data. *GigaScience* **8**, giz107 (2019).
18. Thompson, M. *et al. Methylation risk scores are associated with a collection of phenotypes within electronic health record systems*. http://medrxiv.org/lookup/doi/10.1101/2022.02.07.22270047 (2022) doi:10.1101/2022.02.07.22270047.
19. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).

20. IBDMDB Investigators *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).
21. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B Methodol.* **58**, 267–288 (1996).
22. Groll, A. glmmLasso: Variable Selection for Generalized Linear Mixed Models by L1-penalized Estimation. (2017).
23. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using **lme4**. *J. Stat. Softw.* **67**, (2015).
24. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* **9**, e1003348 (2013).
25. Borren, N. Z. *et al.* Multi-"-Omics" Profiling in Patients With Quiescent Inflammatory Bowel Disease Identifies Biomarkers Predicting Relapse. *Inflamm. Bowel Dis.* **26**, 1524–1532 (2020).
26. Mars, R. A. T. *et al.* Longitudinal Multi-omics Reveals Subset-Specific Mechanisms Underlying Irritable Bowel Syndrome. *Cell* **182**, 1460-1473.e17 (2020).
27. Ota, M. & Fujio, K. Multi-omics approach to precision medicine for immune-mediated diseases. *Inflamm. Regen.* **41**, 23 (2021).
28. Nallabelli, N. *et al.* Biochemical and genome sequence analyses of Megasphaera sp. strain DISK18 from dental plaque of a healthy individual reveals commensal lifestyle. *Sci. Rep.* **6**, 33665 (2016).
29. Wu, T.-R. *et al.* Gut commensal Parabacteroides goldsteinii plays a predominant role in the anti-obesity effects of polysaccharides isolated from Hirsutella sinensis. *Gut* **68**, 248–262 (2019).
30. Glassner, K. L., Abraham, B. P. & Quigley, E. M. M. The microbiome and inflammatory bowel disease. *J. Allergy Clin. Immunol.* **145**, 16–27 (2020).
31. Parker, B. J., Wearsch, P. A., Veloo, A. C. M. & Rodriguez-Palacios, A. The Genus Alistipes: Gut Bacteria With Emerging Implications to Inflammation, Cancer, and Mental Health. *Front. Immunol.* **11**, (2020).
32. Crost, E. H. *et al.* Mechanistic Insights Into the Cross-Feeding of Ruminococcus gnavus and Ruminococcus bromii on Host and Dietary Carbohydrates. *Front. Microbiol.* **9**, 2558 (2018).
33. Henke, M. T. *et al. Ruminococcus gnavus* , a member of the human gut microbiome associated with Crohn's disease, produces an inflammatory polysaccharide. *Proc. Natl. Acad. Sci.* **116**, 12672–12677 (2019).
34. Dolan, S. K. *et al.* Loving the poison: the methylcitrate cycle and bacterial pathogenesis. *Microbiology* **164**, 251–259.
35. Tedelind, S., Westberg, F., Kjerrulf, M. & Vidal, A. Anti-inflammatory properties of the short-chain fatty acids acetate and propionate: a study with relevance to inflammatory bowel disease. *World J. Gastroenterol.* **13**, 2826–2832 (2007).
36. Medina, J. M., Fernández-López, R., Crespo, J. & Cruz, F. de la. Propionate Fermentative Genes of the Gut Microbiome Decrease in Inflammatory Bowel Disease. *J. Clin. Med.* **10**, 2176 (2021).
37. Walker, A. & Schmitt-Kopplin, P. The role of fecal sulfur metabolome in inflammatory bowel diseases. *Int. J. Med. Microbiol.* **311**, 151513 (2021).
38. Morgan, X. C. *et al.* Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* **13**, R79 (2012).
39. Altincicek, B. *et al.* Cutting Edge: Human γδ T Cells Are Activated by Intermediates of the 2- *C* -methyl-d-erythritol 4-phosphate Pathway of Isoprenoid Biosynthesis. *J. Immunol.* **166**, 3655–3658 (2001).
40. Catalan-Serra, I., Sandvik, A. K., Bruland, T. & Andreu-Ballester, J. C. Gammadelta T Cells in Crohn's Disease: A New Player in the Disease Pathogenesis? *J. Crohns Colitis* **11**,

1135–1145 (2017).
41. Kenyon, L., Kumar, S., Tsai, W.-S. & Hughes, J. d'A. Virus Diseases of Peppers (Capsicum spp.) and Their Control. in *Advances in Virus Research* vol. 90 297–354 (Elsevier, 2014).
42. Zhang, T. *et al.* RNA Viral Community in Human Feces: Prevalence of Plant Pathogenic Viruses. *PLoS Biol.* **4**, e3 (2006).
43. Kawato, Y. *et al.* Complete genome sequence analysis of two Pseudomonas plecoglossicida phages, potential therapeutic agents. *Appl. Environ. Microbiol.* **81**, 874–881 (2015).
44. Kirchberger, P. C. & Ochman, H. Resurrection of a global, metagenomically defined gokushovirus. *eLife* **9**, e51599 (2020).
45. Bin Masud, S. *et al.* Utilizing machine learning with knockoff filtering to extract significant metabolites in Crohn's disease with a publicly available untargeted metabolomics dataset. *PLOS ONE* **16**, e0255240 (2021).
46. Li, M. *et al.* Gut Microbial Metabolome in Inflammatory Bowel Disease: from Association to Therapeutic Perspectives. *Comput. Struct. Biotechnol. J.* S2001037022001131 (2022) doi:10.1016/j.csbj.2022.03.038.
47. Carbonero, F., Benefiel, A. C., Alizadeh-Ghamsari, A. H. & Gaskins, H. R. Microbial pathways in colonic sulfur metabolism and links with health and disease. *Front. Physiol.* **3**, (2012).
48. Zheng, Z., Dai, Z., Cao, Y., Shen, Q. & Zhang, Y. Docosapentaenoic acid (DPA, 22:5n-3) ameliorates inflammation in an ulcerative colitis model. *Food Funct.* **10**, 4199–4209 (2019).
49. Ho, B. *et al.* Gut Metabolites Are More Predictive of Disease and Cohoused States than Gut Bacterial Features in a Polycystic Ovary Syndrome-Like Mouse Model. *mSystems* **6**, e01149-20 (2021).
50. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 10.1038/nrg2825 (2010).
51. Nguyen, N. H. *et al.* Machine Learning-based Prediction Models for Diagnosis and Prognosis in Inflammatory Bowel Diseases: A Systematic Review. *J. Crohns Colitis* **16**, 398–413 (2022).