frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 18 January 2021
doi: 10.3389/fmolb.2020.603740

Metabolic Phenotypes as Potential Biomarkers for Linking Gut Microbiome With Inflammatory Bowel Diseases

Stanislav N. Iablokov[1,2†], Natalia S. Klimenko[3,4†], Daria A. Efimova[3], Tatiana Shashkova[3,5], Pavel S. Novichkov[6,7], Dmitry A. Rodionov[1,8] and Alexander V. Tyakht[3,4*]*

[1] A.A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia, [2] P.G. Demidov Yaroslavl State University, Yaroslavl, Russia, [3] Atlas Biomed Group—Knomics LLC, London, United Kingdom, [4] Center for Precision Genome Editing and Genetic Technologies for Biomedicine, Institute of Gene Biology, Russian Academy of Sciences, Moscow, Russia, [5] Moscow Institute of Physics and Technology, Moscow, Russia, [6] PhenoBiome Inc., San Francisco, CA, United States, [7] Lawrence Berkeley National Lab, Berkeley, CA, United States, [8] Sanford-Burnham-Prebys Medical Discovery Institute, La Jolla, CA, United States

OPEN ACCESS

Edited by:
Lu Zhang,
Hong Kong Baptist University, Hong Kong

Reviewed by:
Yiqi Jiang,
City University of Hong Kong, Hong Kong
JinQun Huang,
Beijing Genomics Institute (BGI), China

***Correspondence:**
Alexander V. Tyakht
a.tyakht@gmail.com
Dmitry A. Rodionov
rodionov@sbpdiscovery.org

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: *07 September 2020*
Accepted: *09 December 2020*
Published: *18 January 2021*

Citation:
Iablokov SN, Klimenko NS, Efimova DA, Shashkova T, Novichkov PS, Rodionov DA and Tyakht AV (2021) Metabolic Phenotypes as Potential Biomarkers for Linking Gut Microbiome With Inflammatory Bowel Diseases. Front. Mol. Biosci. 7:603740. doi: 10.3389/fmolb.2020.603740

The gut microbiome is of utmost importance to human health. While a healthy microbiome can be represented by a variety of structures, its functional capacity appears to be more important. Gene content of the community can be assessed by "shotgun" metagenomics, but this approach is still too expensive. High-throughput amplicon-based surveys are a method of choice for large-scale surveys of links between microbiome, diseases, and diet, but the algorithms for predicting functional composition need to be improved to achieve good precision. Here we show how feature engineering based on microbial phenotypes, an advanced method for functional prediction from 16S rRNA sequencing data, improves identification of alterations of the gut microbiome linked to the disease. We processed a large collection of published gut microbial datasets of inflammatory bowel disease (IBD) patients to derive their community phenotype indices (CPI)—high-precision semiquantitative profiles aggregating metabolic potential of the community members based on genome-wide metabolic reconstructions. The list of selected metabolic functions included metabolism of short-chain fatty acids, vitamins, and carbohydrates. The machine-learning approach based on microbial phenotypes allows us to distinguish the microbiome profiles of healthy controls from patients with Crohn's disease and from ones with ulcerative colitis. The classifiers were comparable in quality to conventional taxonomy-based classifiers but provided new findings giving insights into possible mechanisms of pathogenesis. Feature-wise partial dependence plot (PDP) analysis of contribution to the classification result revealed a diversity of patterns. These observations suggest a constructive basis for defining functional homeostasis of the healthy human gut microbiome. The developed features are promising interpretable candidate biomarkers for assessing microbiome contribution to disease risk for the purposes of personalized medicine and clinical trials.

Keywords: gut microbiome, metabolic phenotypes, inflammatory bowel diseases, machine learning, classifier, 16S rRNA sequencing

INTRODUCTION

Recent advances in cultivation-based approaches for studying microbial diversity like culturomics (Bilen et al., 2018) allowed to isolate and characterize phenotypes and genomes of many human gut microbial species (Forster et al., 2019). High-throughput DNA sequencing of microbiome samples is still a method of choice for total profiling of the human microbiome. It provides semiquantitative taxonomic composition as well as functional potential, including biosynthesis of small molecules (Sugimoto et al., 2019). In the gut, the functional profile is generally less variable than the taxonomic one (Eng and Borenstein, 2018). Multiple species tend to be involved in complex ecological networks, many of which arise from cross-feeding, suggesting that the metabolic potential of a single species is less important than a community-wide gene-centric metabolism—provided the completeness of the pathways, however.

Deciphering the total microbiome network of metabolic interactions became possible with the advent of whole-genome shotgun (WGS) metagenomics. Even simple mapping of the reads to global reference gene catalogs already shows a strong variability between subjects both globally and by specific gene groups including carbohydrate catabolism, antibiotic resistance, and virulence factors (Yarygin et al., 2017), suggesting a more detailed investigation of distinct metabolic pathways. The extent of realization of the metabolic potential encoded in the microbiome can be evaluated quantitatively and qualitatively using metabolomics, particularly to elucidate its alterations linked to specific disorders. For example, targeted and untargeted metabolomic analysis of fecal samples from patients with inflammatory bowel diseases (IBD) revealed dysregulated metabolism of SCFAs, bile acids, tryptophan, and other molecules (Franzosa et al., 2019), suggesting that microbiota-derived metabolites play key roles in the pathogenesis (Lavelle and Sokol, 2020). To date, metabolomic experiments are more expensive and less standardized compared to high-throughput sequencing. The concept of predicting metabolite levels from metagenomic composition based on bacterial genome-scale metabolic models has shown promising results in the context of personalizing therapeutic dietary interventions for Crohn's disease (CD) patients (Bauer and Thiele, 2018).

The amplicon 16S rRNA sequencing is still the method of choice in terms of cost and information content for large-scale microbiome surveys of links between human microbiome and diet, diseases, verification of microbiome-related health claims of food products, and individual microbiome profiling. The sequenced variable 16S regions are often organized into operational taxonomic units (OTUs), i.e., clusters of similar sequences, or merged into the exact biological sequences present in the sample, so-called amplicon sequence variants (ASVs) that are further counted to get their relative abundances and taxonomically assigned using reference 16S databases (Prodan et al., 2020). Although this approach does not allow direct measurement of microbial gene content other than 16S rRNA itself, there are algorithmic methods for inferring the functional composition of the community from such data based on an *a priori* accumulated knowledge about microbial genomes (Aßhauer et al., 2015; Louca et al., 2016; Douglas et al., 2020; Narayan et al., 2020). Metagenomic prediction tools revealed functional alterations in the human gut microbiome linked to many diseases including IBD (Imhann et al., 2018), Parkinson's disease (Cirstea et al., 2020), and nonalcoholic fatty liver disease (Boursier et al., 2016), as well as to dietary interventions (Klimenko et al., 2018; Volokh et al., 2019) and other factors. However, the tools provide low accuracy for certain groups of functions. Firstly, some functional groups of genes are subject to frequent horizontal gene transfer across distant taxa (like antibiotic resistance determinants). Secondly, many metabolic pathways are established based on general databases and are neither curated to the point of sufficient accuracy nor take into account the specifics of a particular niche (like human gut). Meanwhile, precise metabolic reconstruction provides increased precision to elucidate ecological mechanisms of the communities (Sung et al., 2017; Garza et al., 2018).

Besides total analysis of all metabolic functions carried out by the gut microbes, some of them are often examined in a targeted manner as being highly relevant to the host–microbe interactions, ecological equilibrium, and diet, and of significant interest to be explored in novel datasets using interactive online systems (Efimova et al., 2018). The majority of these groups are short-chain fatty acid (SCFA) production, carbohydrate catabolism, and synthesis of vitamins and amino acids. The major SCFAs, namely, acetate, butyrate, and propionate, are synthesized by the gut microbes and are essential for host physiology by regulating inflammation, immunity, tumorigenesis, satiety, and involvement in signal functions (Koh et al., 2016). The propensity to synthesize them and the specific metabolic pathways vary across the bacterial kingdom, as showcased by butyrate (Vital et al., 2014). In the gut of IBD patients, there is a depletion of butyrate synthesis potential (Laserna-Mendieta et al., 2018). The main substrates for SCFA production are carbohydrates (glycans) of various structural complexities (oligo- and polysaccharides) coming as dietary fibers and general food components, making them the keystones in the prebiotic action of these molecules (Gibson et al., 2017). Different bacterial taxa have different capacities toward degrading a specific fiber type, and cross-feeding based on symbiotic catabolism of a complex glycan is not uncommon (Cockburn and Koropatkin, 2016; Cerqueira et al., 2020). Examination of an individual gut microbiome's total capacity for glycan catabolism suggests a way of designing personalized microbiome-tailored dietary plans (Klimenko et al., 2018).

Another prominent group of host health-relevant metabolites are vitamins. In the gut, the microbes can synthesize vitamin K and B vitamins along with their precursors (Rodionov et al., 2019). There are reports that at least some of the vitamins are accessible to the host (LeBlanc et al., 2013) and their fecal concentrations can be associated with clinical factors (McCann et al., 2019). Recent studies indicate that the importance of gut as a source of vitamins is limited and even a greater role of these vitamins might be in maintaining a robust ecological network between the species in the gut (Rodionov et al., 2019; Sharma et al., 2019). Interestingly, investigation of vitamin synthesis from

stool metatranscriptomes of IBD patients showed an increased expression level of biotin (vitamin B7) biosynthetic enzymes compared to healthy controls (Das et al., 2019). Finally, amino acids released from undigested luminal proteins and peptides are accessible for gut microbiota that is involved in amino acid fermentation to form SCFAs and/or transformation to numerous metabolic end products such as phenols and indoles from aromatic amino acids (Davila et al., 2013). During intestinal inflammation, the microbiome potential for synthesis of amino acids, can decrease in favor of catabolism (Morgan et al., 2012). The therapeutic potential of amino acids for IBD has been proposed due to the reduction of inflammation, oxidative stress, and cell death in the gut they can evoke (Liu et al., 2017).

Development of high-precision and accuracy approaches for profiling of these selected functions will provide a valuable tool for efficient mining of biomedically relevant information from amplicon sequencing data and improving the downstream interpretations of gut microbiome data. Previously, we developed a new genomics-based methodology of predictive phenotype profiling that computes CPI (Community Phenotype Index) values as community-wide fractional representation of a limited set of basic metabolic phenotypes (such as amino acid auxotrophy and sugar utilization capabilities) deduced from *in silico* reconstruction over a large reference collection of HGM genomes and projected over 16S rRNA abundance profiles of the analyzed samples (Rodionov et al., 2019). This predictive metabolic phenotype profiling methodology was further applied to characterize the 16S rRNA amplicon-based taxonomic profiles of the fecal microbiomes in *in vivo* and *in vitro* studies and identify metabolic phenotypes that are linked to variable diets or growth conditions (Peterson et al., 2019; Sharma et al., 2019; Elmén et al., 2020; Jones et al., 2020). Here we used this *in silico* metabolic phenotype profiling approach to identify the links between the functional homeostasis of gut microbiome and disease. To assess the performance of our approach in discovering novel robust biomarkers, we applied it to the case of the IBD that are associated with the altered microbiome composition, along with the genetic, lifestyle and environmental factors (Beaugerie et al., 2018).

MATERIALS AND METHODS

Study Design and Raw Sequence Data Analysis

For our analysis, we selected the following three previously published IBD datasets with publicly available 16S rRNA gene sequencing data of stool samples. The Spanish dataset (ESP) included 34 Crohn's disease (CD) patients, 33 ulcerative colitis (UC) patients, and 101 healthy controls (HC) (Pascal et al., 2017). The Chinese dataset (CHN) included 72 CD, 51 UC, and 71 healthy individuals (Zhou et al., 2018); an additional 16 CD patients on infliximab treatment were excluded from the analysis. Both ESP and CHN datasets were further filtered to retain only one sample per individual in cases of multiple replicates. The study of 313 IBD patients from the Netherlands (NLD) included 188 patients with CD, 107 patients with UC, and 18 additional IBD patients with either intermediate or undetermined disease status (Imhann et al., 2018); the latter were excluded from the study. The healthy NLD group originally included 1,010 individuals from the LifeLines DEEP cross-sectional general population study (Tigchelaar et al., 2015) that was further reduced to 495 healthy controls selected as previously described in (Imhann et al., 2018).

TABLE 1 | Number of samples per dataset and per clinical status analyzed in this work.

IBD status/dataset	Spain (ESP)	Netherlands (NLD)	China (CHN)
Healthy controls (HC)	91	496	67
Crohn's disease (CD)	34	163	50
Ulcerative colitis (UC)	39	99	37

The raw sequence data from the CHN and ESP datasets were downloaded from the European Nucleotide Archive (www.ebi.ac.uk/ena)—project IDs PRJNA422193 and PRJEB22028, respectively. The NLD datasets were obtained from the European Phenome-Genome Archive (https://ega-archive.org/), project IDs EGAS00001002702 and EGAS00001001704. The 16S rRNA gene sequences (hypervariable region V4) were analyzed using the DADA2 plugin from the QIIME2 package (Callahan et al., 2016; Bokulich et al., 2018). Briefly, raw demultiplexed reads were quality filtered, denoised, and dereplicated into ASVs and the read counts (relative abundance values) were calculated for each ASV in each sample. Average abundance loss after DADA2 filtering was 23% for ESP, 27% for CHN, and 12% for NLD datasets. The obtained ASV abundance tables were additionally filtered to retain only the amplicons satisfying at least one of the following criteria: (i) ASV is present in >0.5% of samples; (ii) dataset-average ASV abundance >0.25%; and (iii) maximum ASV abundance >0.5% across the dataset. As a result, each of the three datasets were filtered with 1–2% average abundance loss per sample. Finally, we filtered each dataset by a minimal coverage (number of reads per sample). For ESP and NLD, the coverage threshold was >15,000 reads, whereas for the CHN, it was set to >4,000 reads due to overall low read counts for this dataset. The distributions of read numbers are shown in **Supplementary Figure 1**. The numbers of samples retained for further analysis in the three datasets are provided in **Table 1**.

Taxonomic Assignment

Taxonomic classification of ASV representative sequences was performed using the multi-taxonomic assignments (MTA) approach as described below. Each representative sequence was aligned using NCBI BLAST ToolKit against a joined reference 16S rRNA database including sequences from RDP database version 11.5 (Cole et al., 2014) and NCBI 16S database version of December 2019. Alignment results were sorted according to their identity F (as a fraction of 1), with the maximum F-value for the best hit denoted as M. Alignment hits with value of F in the range between M and $M-(1-M)/S$ and greater than a threshold value D were selected for MTA. Here, S acts as a scaling parameter,

which controls the list of taxonomic descriptions accepted for MTA based on the *F* value of the alignment and was taken equal to 4. The drop threshold parameter *D* was taken equal to 0.85. The resulting MTA for each ASV represented a list of unique regular taxonomic descriptions with equal weights assigned to each item. String representations for MTA consisted of slash-separated names of taxa on each taxonomic level. The main advantage of the MTA over the consensus-based methods consists in its ability to assign taxonomic descriptions up to the genus (and even species) level for sequences with low identity and, hence, poor genus-level resolution. In this case, the organisms with partially resolved genus-level taxonomies could also participate in machine learning, providing multi-taxonomic descriptions as features. However, the broader the list of accepted taxonomies for MTA is, the less useful the corresponding feature becomes for cross-study analysis. For example, multi-taxonomic descriptions A/B/C for an ASV in one study and A/B/D for a closely related ASV in another study are considered as different features. To increase the overlap in the sets of taxonomic features between different studies, one should aim for shorter MTA strings. This motivates the strict choice of S and D parameters, which leads to compact MTA descriptions.

Prediction of Metabolic Phenotypes in Reference Genomes

Functional gene assignments and metabolic reconstructions were performed using the SEED database/tools that allow a subsystem-based analysis of ∼6,000 bacterial genomes, including a subset of 2,662 reference human gut microbial (HGM) genomes representing 690 individual species (Overbeek et al., 2014). The subsystem-based approach for metabolic reconstruction combines protein similarity search, analysis of chromosomal gene clusters, and phylogenetic profiling (Overbeek et al., 2005). The collection of curated subsystems includes metabolic pathways for (i) biosynthesis of essential nutrients (vitamins, amino acids); (ii) uptake and fermentation of carbohydrates including mono-, oligo-saccharides, sugar acids, and alcohols; (iii) degradation of amino acids; and (iv) production of SCFAs. The metabolic subsystems were developed based on the previously published genomic studies of phylogenetic distribution of bacterial pathways for metabolism of vitamins and amino acids, utilization of carbohydrates, and production of butyrate and propionate in HGM bacteria (Rodionov et al., 2011, 2013, 2019; Ravcheev et al., 2013; Khoroshkin et al., 2016; Leyn et al., 2017; Arzamasov et al., 2018; Bouvier et al., 2019; Feng et al., 2020). Using the collection of pathway-specific logical rules (Rodionov et al., 2019), we obtained Binary Phenotype Matrix (BPM) describing 94 inferred phenotypic features (nutrient requirements, utilization capabilities, metabolite production) of each reference genome as binary ("1" or "0") phenotypes and reflecting the presence/absence of a complete catabolic or biosynthetic pathway. In addition to catabolic enzymes, the sugar utilization subsystems also included sugar-specific uptake transporters; thus, the assigned sugar utilization capability required the presence of both catabolic pathway and uptake transporter. We also obtained the distribution of 229 families of glycosyl hydrolases (GHs) in the analyzed reference genomes using dbCAN2 tool (Zhang et al., 2018). The obtained GH family distribution was converted to GH-BPM, where each column represents an individual GH family, and each binary phenotype represents the presence or absence of at least one enzyme from this family. The obtained metabolic BPM and GH-BPM for 2,662 reference genomes provided as a part the Phenotype Profiler tool (see below) were used to calculate the Community Phenotype Index (CPI) for each metabolic phenotype or GH family and each 16S rRNA sample as previously described (Rodionov et al., 2019) and explained in more details below.

Calculation of the Metabolic Phenotype Profile

To obtain phenotype profiles for analyzed 16S rRNA samples, we utilized the Phenotype Profiler tool provided by PhenoBiome Inc. (San Francisco, CA). First, we mapped each ASV obtained from the samples to a reference collection of 2,662 microbial genomes based on their 16S rRNA gene sequence match. The reference HGM genome collection was analyzed with Barrnap (https://github.com/tseemann/barrnap) to predict the location of ribosomal RNA genes and select all 16S rRNA gene sequences for each genome. Partial 16S rRNA gene sequences were replaced with corresponding complete sequences from the NCBI 16S rRNA database. In order to establish such mapping, each ASV sequence was first aligned against the reference 16S rRNA collection using the NCBI BLAST standalone toolkit. To further assign reference organisms to ASV, we used the same top hit selection criteria as in the MTA procedure with the same values for S and D as described above. The reference organisms corresponding to the selected alignment hits, therefore, constituted a mapping for each ASV with weights distributed equally.

To calculate CPI values for each sample and for each metabolic phenotype, we first obtained a probabilistic estimate p_i for a given ASV (that corresponds to one or more reference genomes) to possess a certain binary metabolic phenotype as a mapping-weighted average across BPM. CPI values were calculated as:

$$CPI = \sum_i p_i A_i$$

where the sum is taken over all ASVs and A_i represents a particular ASV's relative abundance. CPI provides a fractional representation of cells in the community possessing a specific metabolic pathway or GH family.

Phenotype alpha diversity (PAD) metric was calculated for each metabolic phenotype for each sample as an alpha diversity of microbial ASVs possessing a particular phenotype. In order to do this, we first aligned ASVs' representative sequences using MUSCLE (Edgar, 2004). Next, an unrooted tree was built from the alignment using FastTree 2 (Price et al., 2010). Finally, the tree was rooted according to the midpoint strategy and used to compute Faith's phylogenetic alpha diversity metric with the Python scikit-bio (http://scikit-bio.org) package. For each sample and each phenotype, only those ASVs that had map-averaged

expected phenotype values >0.6 were considered as the ones having the potential to express certain phenotypes.

Machine Learning for Clinical Status Prediction

Classification Setups

To construct the microbiome-based classifiers of clinical status (HC, CD, or UC), we used the taxonomic and metabolic phenotype profiles obtained for 3 datasets (CHN, ESP, NLD) as input features for Random Forest classifiers (implemented in Python's scikit-learn RandomForestClassifier). The classifiers were different by three strategies, two sets of predictors, and two IBD clinical states (CD or UC). We used the following strategies: (i) classification of each dataset separately using two-thirds of samples as a training set and one-third as a testing set (Single strategy); (ii) classification of a mixed dataset constructed from an equal number of healthy and randomly selected affected subjects from every dataset using two-thirds as a training set and one-third as a testing set (Mixed strategy); and (iii) classification of joined datasets constructed by combining each dataset pair as a training set and the remaining third dataset as a testing set (leave-one-out strategy, L1O). By applying these strategies to the three analyzed datasets (CHN, ESP, NLD), we obtained seven variants of classifiers (including three variants for the L1O strategy, three variants for the Single strategy, and one variant for the Mixed strategy). We also considered two different sets of predictors including (i) taxonomic names at the genus level and their corresponding abundances and (ii) metabolic phenotypes and their corresponding CPI values, and two IBD clinical states CD and UC) compared with healthy controls (HC). As a result, we designed 28 types of RF classifiers.

The following parameters were used to build each RF classifier:

1) tree depth = 3;
2) number of trees = 200;
3) percent of features in each split = 50%;
4) balanced class weights.

The use of balanced class weights for cost function calculation ensured that classes (clinical status) were weighted inversely proportional to their frequency in the training set.

For the Single and Mixed strategies, we performed random subsampling and 10 cross-validation iterations to evaluate mean performance characteristics. In the L1O strategy, two combined datasets served as a training set with the remaining one being a testing set; thus, we performed three cross-validation iterations without random subsampling.

Feature Filtration and Extraction

For each cross-validation iteration, we also performed feature filtration and feature extraction (**Figure 1**). These procedures were based only on the information contained in the training part of the data. Feature filtration was done according to the following sets of rules, for taxonomic as well as for the phenotypic features. For taxonomic features, the taxa satisfying at least one of the following criteria were filtered out: (i) nonzero abundance in <5 samples and (ii) maximum abundance across the training set <1%. For phenotypic features, phenotypes satisfying at least one of the following criteria were filtered out: (i) mean CPI value out of the range of [0.1, 0.9]; (ii) CPI value >0.05 in <5 samples; and (iii) mean PAD value <3.5. For the strategies including more than one dataset (MC, L1O), all above filtering conditions were required to be satisfied in each dataset separately. Feature filtration was followed by the feature extraction including 10 sub-iterations of classification with the same classifier parameters as before. After each sub-iteration, we evaluated feature importance (decrease of Gini impurity) based on the training set data. The top 20 features with the highest mean importance values were used in the final classification.

Subsampling

The groups of samples were subsampled for each of the aforementioned classification strategies except for the Single strategy in order to equalize the contribution of each dataset to the training set. For the Mixed strategy, this resulted in the following numbers of samples per IBD clinical status in each dataset: 67 for HC, 34 for CD, and 37 for UC. For the L1O strategy, these numbers were dependent on a particular strategy variant (**Table 2**). The disproportion in the number of samples between the CD/UC/HC classification groups was accounted for by using balanced class weights.

Performance Evaluation

To assess classification quality, we conducted the ROC curve analysis on each cross-validation iteration and calculated the following metrics: Area Under the Curve (AUC), sensitivity, and specificity. Then, for each of 28 classifier types, we estimated mean values and corresponding standard deviations of these metrics.

Analysis of Stable Predictors

We identified a collection of stable predictors (phenotypes or genera) for each disease status using feature importance analysis. For each classifier type, mean feature importances (mean decrease of Gini impurity) were calculated across 10 cross-validation iterations. If some feature was not extracted (during filtration and extraction procedure) in one or more cross-validation iterations, then its importance value was set to zero. A predictor was defined as stable for a given disease status if it satisfied the following criteria: (i) nonzero importance values in at least six out of seven classifier types corresponding to possible combinations of classification strategy and dataset and (ii) the same-sign difference between mean predictor values (CPI or abundance) for HC and CD/UC groups in each dataset. The importance for each stable predictor was taken to be its average value across possible combinations of classification strategy and dataset.

To investigate the relationship between stable predictors and clinical status, we constructed additional classifiers based only on stable predictors for the Mixed classification strategy for each combination of predictor set and disease status. For each such classifier, 20 cross-validation iterations were performed. We also constructed single-feature partial dependence plots (PDP) in each iteration for each of the stable predictors with resolution of

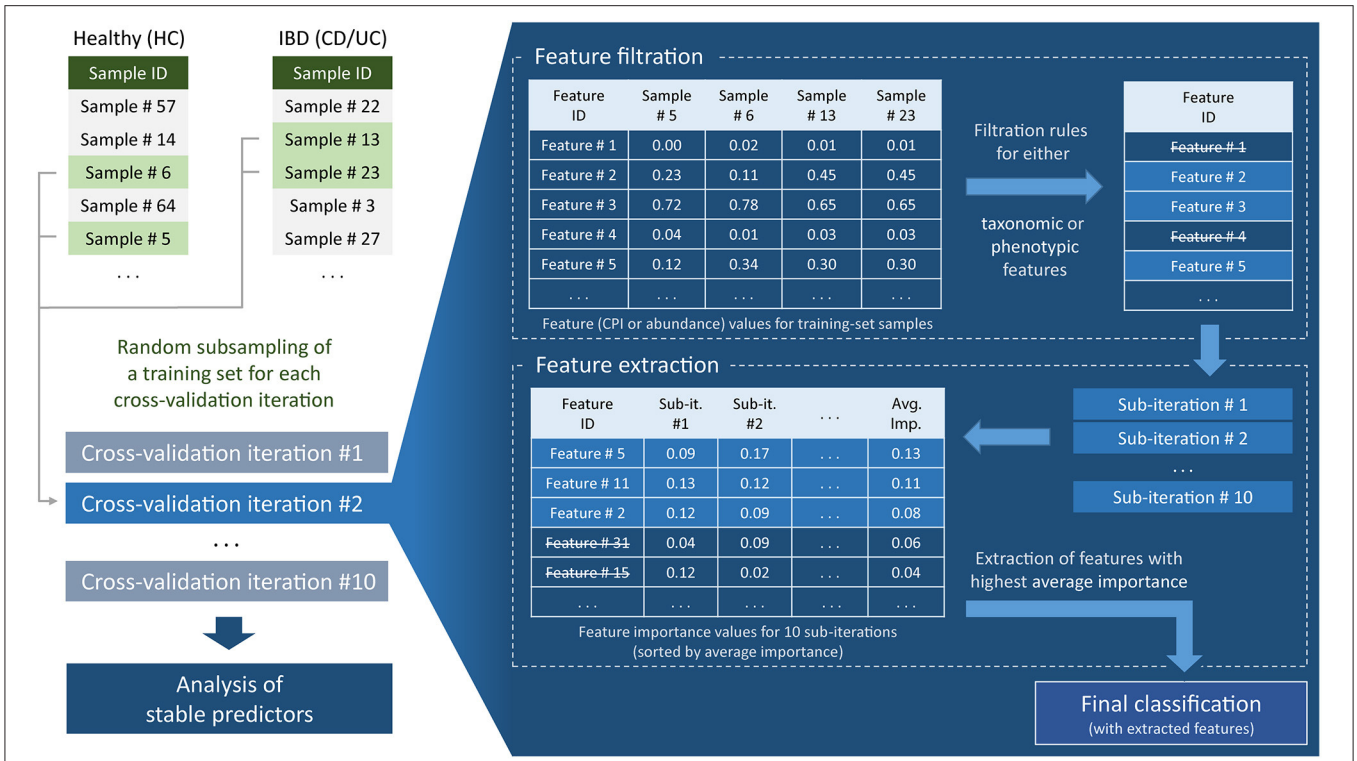

FIGURE 1 | Flowchart of machine learning approach for clinical status prediction based on microbiome data.

TABLE 2 | Number of samples per dataset and per classification group used for the L1O classification strategy.

Strategy	L1O:ESP			L1O:NLD			L1O:CHN		
Description	CHN+NLD (training) ESP (testing)			CHN+ESP (training) NLD (testing)			NLD+ESP (training) CHN (testing)		
Dataset	CHN	NLD	ESP	CHN	ESP	NLD	NLD	ESP	CHN
HC	67	67	91	67	67	496	91	91	67
CD	50	50	34	34	34	163	34	34	50
UC	37	37	39	37	37	99	39	39	37

20 grid steps and probability of CD or UC outcome as a response. The mean dependency was calculated across the iterations. Based on the form of the dependency, the predictors were classified into five categories—sharply decreasing, sharply increasing, smoothly decreasing, smoothly increasing, and unclassified. The classification into categories was performed by analyzing the differences of probability values (Δi) between the (i+5)-th and i-th step (for each i from 1 to 15) in the following way. First, the direction of dependence (increasing or decreasing) was determined by the sign of Δ_i with the highest absolute value: positive sign corresponds to the increasing PDP and negative—for decreasing. If one predictor had at the same time positive and negative Δ_i, and the highest ratio between their absolute values was < 2, then the predictor was defined as unclassified. Second, the form of dependence (sharp or smooth) was determined by the following rule: if at least one of the absolute Δ_i values was $>= 3$ times higher than the maximum probability differences calculated for outer lower [1, i) and upper (i+5,20] intervals, then the sharp form was chosen. Otherwise a smooth form was chosen.

RESULTS

Predicted Metabolic Phenotype and Taxonomy Profiles of IBD Samples

We selected three previously published fecal microbiome datasets representing geographically distinct cohorts of IBD patients from China (CHN) (Zhou et al., 2018), Spain (ESP) (Pascal et al., 2017), and the Netherlands (NLD) (Imhann et al., 2018). Each analyzed dataset included two groups of IBD subjects with either Crohn's disease (CD) or ulcerative colitis (UC) diagnosis and also a group of healthy control (HC) subjects from the same geographical population. Having filtered the

three datasets to remove duplicate samples representing the same subject and samples from subjects that were treated with immunosuppressants, we have analyzed raw 16S rRNA amplicon sequence data for each dataset individually. We applied QIIME2's DADA2 plugin to obtain the ASV abundance profiles and then filtered out the samples with low read counts as described in section Materials and Methods. The number of remaining samples per dataset and per clinical status group is provided in **Table 1**.

ASVs' taxonomies were obtained using the multi-taxonomic assignment (MTA) approach. The analysis of abundance distribution for the top 20 taxonomic genera demonstrated a much larger variability between the analyzed three datasets and moderate variations between groups of samples with a different clinical status within each dataset. Further, we analyzed the obtained ASV profiles using the metabolic phenotype profiling approach (Rodionov et al., 2019) to calculate the sample-wise Community Phenotype Indices (CPIs) for 94 metabolic phenotypes from BPM constructed from the reference collection of 2,662 HGM genomes. Using the same approach, we calculated CPI values for 229 GH families from GH-BPM constructed from the genomic distribution of GH enzymes in reference genomes (see section Materials and Methods). Additionally, for each phenotype (and GH family), we estimated the corresponding Phenotype Alpha Diversity (PAD) values, which was defined as the phylogenetic alpha diversity of a subcommunity of corresponding phenotype carrier (as described in section Materials and Methods). The obtained CPI and PAD values for the analyzed metabolic phenotypes and GH families across all samples in the three analyzed datasets are shown in **Supplementary Tables 1, 2**, while the genus-level taxonomic profiles are presented in **Supplementary Table 3**. We performed principal component analysis (PCA) of the phenotypic features and principal coordinate analysis (PCoA) of the taxonomic features calculated for samples from the three IBD datasets and revealed visible differences between microbiome composition of samples from each of these datasets (**Figures 2A,B**). Distributions of top 10 variable features among taxonomic genera and phenotypic CPIs across the three IBD datasets are provided in **Figures 2C,D**. Among the most variable taxonomic genera are *Bacteroides* (in CHN and ESP datasets), *Blautia* (in NLD and ESP), and *Bifidobacterium* (in NLD), while the list of most variable metabolic phenotypes is largest in the CHN dataset and includes biosynthesis of vitamins (biotin/B7, queuosine/Q, lipoate) and propionate, utilization of N-acetylglucosamine (GlcNAc), degradation of threonine (Thr_D), and the presence of specific GH families (GH94, GH43_12, GH43_10, GH13).

Classification of IBD Status Based on Taxonomy and Microbial Phenotypes

The obtained metabolic phenotype profiles (CPI and PAD values) and genus-level taxonomic profiles were used to examine the potential of microbiome features to predict IBD clinical status using the Random Forest (RF) classification approach. As an input for the RF classifier, we used two types of features, either relative taxonomic abundances at the genus level or CPI values for predicted metabolic phenotypes. Both types of features were filtered during the cross-validation procedure according to a set of rules (see sections Materials and Methods, **Figure 1**). Particularly, the corresponding PAD values served as one of the filtering criteria to eliminate phenotypes with contribution to CPI coming from a phylogenetically narrow group of organisms. The IBD clinical status (CD, or UC and HC) was used as the RF classifier output. For each of the three analyzed datasets (NLD, ESP, CHN), we applied three division strategies (Single, Mixed, and L1O, see section Materials and Methods) to obtain seven different variants of division of the analyzed datasets on training and testing sets. In total, we obtained 28 classifiers for two pairs of IBD clinical states (CD vs. HC, and UC vs. HC), two sets of predictors (taxonomies or phenotypes) and three division strategies.

Crohn's Disease Classifiers (CD-vs.-HC)

As a result, we obtained 14 CD-vs.-HC classifiers, with corresponding performance characteristics listed in **Table 3**. In general, the taxonomy-based classifiers demonstrated higher AUC and sensitivity values than the phenotypes-based classifiers (**Figures 3A,B**). On the contrary, the latter showed higher specificity values. By comparing features that were selected after feature filtration and feature extraction steps in different strategies (see **Figure 1** and section Materials and Methods for more details), we extracted 15 phenotypic and 12 taxonomic features that work as stable predictors (**Figures 3C,D**). The selected taxonomic groups and phenotypes demonstrated variable average importance values across different classification strategies and datasets. For the majority of stable taxonomic predictors (10 out of 12 taxa), mean abundance values were higher in healthy controls than in CD patients. However, for the majority of phenotypic predictors (12 out of 15 phenotypes), mean CPI values were lower in healthy controls than in CD patients.

To investigate the influence of each stable predictor on the classification result, the partial dependence plot (PDP) analysis was performed. Using the Mixed strategy, we constructed taxonomy-based and phenotype-based classifiers, with only stable predictors as input features. For each stable predictor, a PDP was obtained, and the predictors were classified based on their PDP form into five categories: sharply decreasing, sharply increasing, smoothly decreasing, smoothly increasing, and unclassified (see section Materials and Methods). In these descriptions, the term "increasing" means that the probability of CD outcome is greater for greater predictor values, and vice versa for the term "decreasing." The examples of PDP forms are shown in **Figure 4**. The grouping of predictors into PDP form categories is listed in **Table 4**, and all PDPs are shown in **Supplementary Figures 2, 3**.

The majority of stable taxonomic predictors (10 out of 12 taxa) for CD-vs-HC classifiers demonstrated sharply decreasing PDP forms (**Supplementary Figure 2**, **Table 4**). Moreover, for each of them, the abundance threshold for the sharp decrease of CD output probability was close to zero (e.g., for *Oscillospira*,

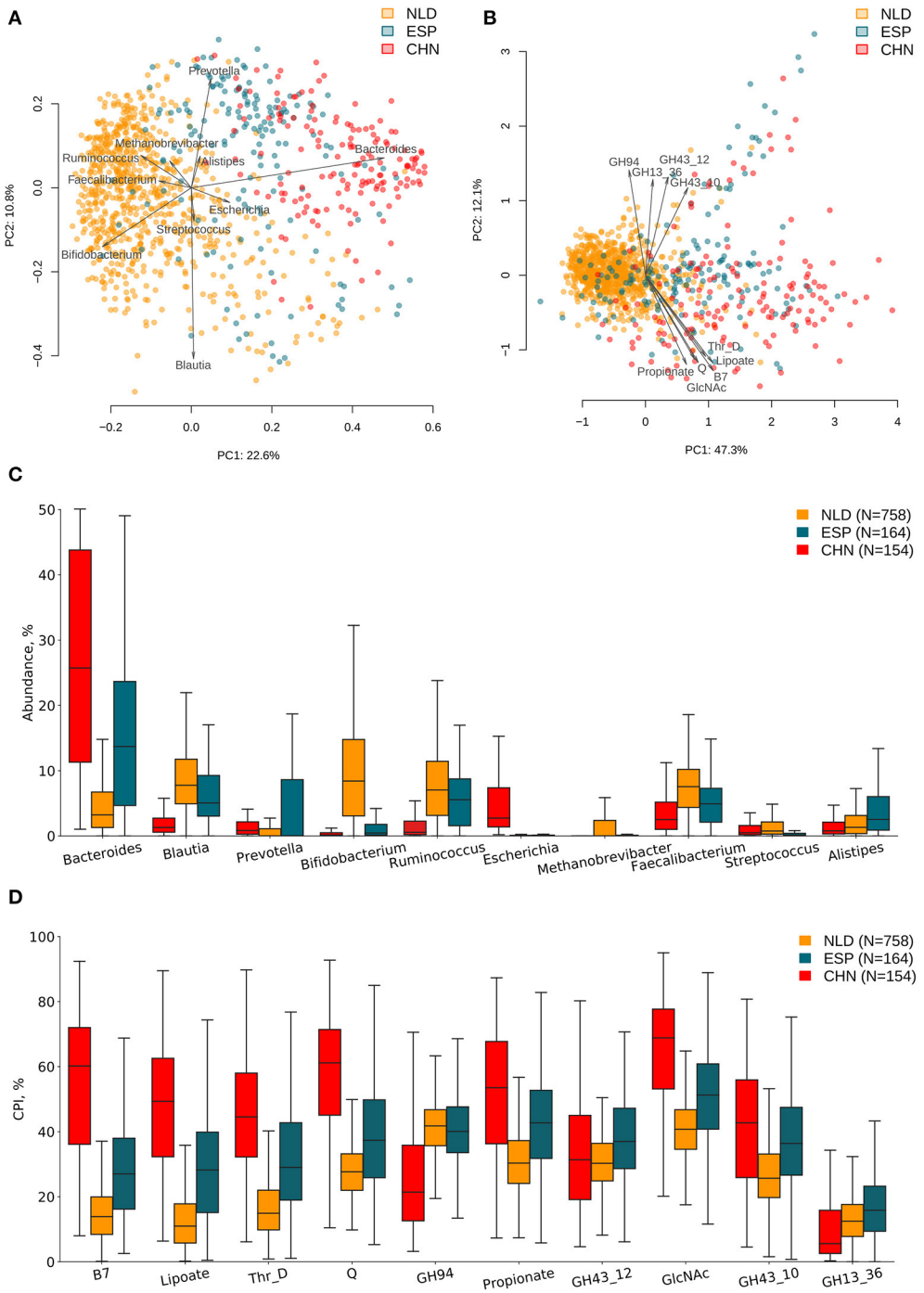

FIGURE 2 | Differences in microbiome composition and metabolic functions of samples from three investigated IBD datasets. **(A)** PCoA analysis of samples using taxonomic features (genera abundance) and Bray–Curtis dissimilarity metric for distance calculation. **(B)** PCA analysis of samples using metabolic phenotype features (CPI) for distance calculation. Arrows show top 10 features in terms of the explained variance in the given axes. Arrow length is proportional to the percent of variance explained by the feature. Distribution of taxonomic abundances **(C)** and Community Phenotype Index (CPI) values **(D)** for the corresponding top 10 features across three IBD datasets.

see **Supplementary Figure 2**). It suggests that the probability of CD as an outcome was high in the complete absence of the corresponding taxon in the community and dropped sharply even with the small increase of its abundance; the further increase of the abundance did not considerably affect the output. Noteworthily, the majority of phenotypic predictors (12 out of 15) had an increasing PDP form, smooth or sharp (**Table 4**, **Supplementary Figure 3**).

TABLE 3 | Mean performance characteristics of the taxonomy- and phenotype-based CD classifiers over 10 classification iterations.

Strategy/dataset	Taxonomy-based classifier			Phenotypes-based classifier		
	Mean sensitivity	Mean specificity	Mean AUC	Mean sensitivity	Mean specificity	Mean AUC
Single:CHN	0.7867	0.7857	0.8463	0.7333	0.7571	0.8463
Single:ESP	0.7900	0.9179	0.9511	0.7700	0.9036	0.915
Single:NLD	0.7755	0.9161	0.9344	0.6918	0.9020	0.8920
L1O:CHN	0.9440	0.4358	0.7985	0.850	0.6164	0.7725
L1O:ESP	0.9353	0.8165	0.9484	0.8176	0.8516	0.8895
L1O:NLD	0.4252	0.9871	0.8567	0.3534	0.9899	0.7632
Mixed: all datasets	0.7548	0.8083	0.8736	0.7065	0.8350	0.8339
Mean across all variants	0.7731	0.8096	0.8870	0.7032	0.8365	0.8446

In the L1O strategy description, the name of the dataset corresponds to the testing dataset (for example, the "L1O:CHN" description means that the classifier was trained on the ESP and NLD datasets and tested on the CHN dataset). Cell color reflects the characteristics' values (greater values correspond to darker colors).

Ulcerative Colitis Classifiers (UC-vs.-CD)

Similar procedures were performed to obtain 14 UC-vs.-HC classifiers. Their performance characteristics are listed in **Table 5**. In general, taxonomy-based classifiers demonstrated higher AUC, sensitivity, and specificity values than phenotype-based classifiers (**Figures 5A,B**). For the L1O:NLD classification variant, AUC values for the phenotype-based classifier (0.43) was close to that of a random guess (0.5). Overall, all constructed UC-vs.-HC classifiers demonstrated lower performance characteristics when compared to the corresponding CD-vs.-HC classifiers.

Despite the fact that in some strategies the corresponding AUC values were close to 0.5, we still observed stable predictors (defined the same way as for CD-vs.-HC classifier analysis, see section Materials and Methods). However, the number of stable predictors for UC-vs.-HC classifiers (seven taxonomic and four phenotypic) was much lower than for the respective CD-vs.-HC cases (**Figures 5C,D**).

The PDP analysis for the predictors of UC-vs.-HC classifiers revealed the relationships between the predictors' values and disease probability that were structurally similar to the CD-vs.-HC case (**Table 4**, **Supplementary Figures 4, 5**). For taxonomic features, PDP forms for three predictors were sharply decreasing, and, similarly to the CD-vs.-HC analysis, the sharp drop threshold was close to zero predictor value. The *Gemmiger* and *Ruminococcus* genera showed a similar behavior in the CD-vs.-HC classifier. For four taxonomic genera, their PDPs were smoothly increasing (**Table 4**, **Supplementary Figure 4**). For the phenotype-based UC-vs.-HC classifier, half of the predictors' PDPs were smoothly increasing, while the other half were smoothly decreasing (**Table 4**, **Supplementary Figure 5**).

DISCUSSION

High-throughput sequencing surveys of the human gut microbiome provide amplicon or WGS sequence datasets that are promising for noninvasive disease risk prediction; however, the task of extracting meaningful biomarkers from such data is still challenging. Machine learning (ML) methods including deep neural networks are powerful for understanding connections of the human gut microbiome to human health (Zhou and Gallins, 2019). During meta-analysis of composition profiles, the corresponding sets of features (e.g., OTUs or ASVs) may be incomparable across different studies due to different laboratory methods/protocols, such as 16S rRNA sequencing region, and quality control (QC) parameters. Traditionally, this problem is solved by aggregating raw ASV (or OTU) features into more biologically meaningful taxonomic features (clade names at the family, genus, or species level). Nevertheless, this aggregation is not universal because it depends on the taxonomic resolution provided by the sequenced gene region. For instance, depending on the ASV length, one might assign unambiguous taxonomic descriptions either up to the species level (*Escherichia coli*) or up to the genus level (*Escherichia*), or even up to the family level (*Enterobacteriaceae*) when the taxonomic resolution is insufficient to distinguish one microorganism from another (e.g., *Escherichia coli* and *Salmonella enterica*). Such gross aggregation leads to the loss of details in phylogenetic description. This is partially remedied through the use of the multi-taxonomic assignment (MTA) approach (see section Materials and Methods), which consists in resolving taxonomic ambiguities not by aggregation on a higher taxonomic level but via listing of all organisms phylogenetically related to the considered ASV (e.g., *Bacteroides ovatus/vulgatus*).

In this study, we introduced yet another approach for feature aggregation, which utilizes the metagenomically predicted metabolic features that are computed for each 16S sample using our previously developed metabolic phenotype profiling tool. The computed CPI values represent the expected fraction of bacterial cells in the community possessing a certain metabolic capability (a phenotype). A collection of CPI values for a selected group of phenotypes, termed Community Phenotype Profile (CPP), therefore constitutes an alternative set of features that can be used for ML. This approach has the following obvious advantages. Firstly, the set of phenotypic features is universal to all organisms and thus can serve as a basis for a cross-study analysis. Secondly, the separation of groups (HC vs. CD/UC in this study) is performed based on the bacterial metabolism and, hence, leads to a straightforward biological interpretation. Thirdly, it allows to refine the classification results

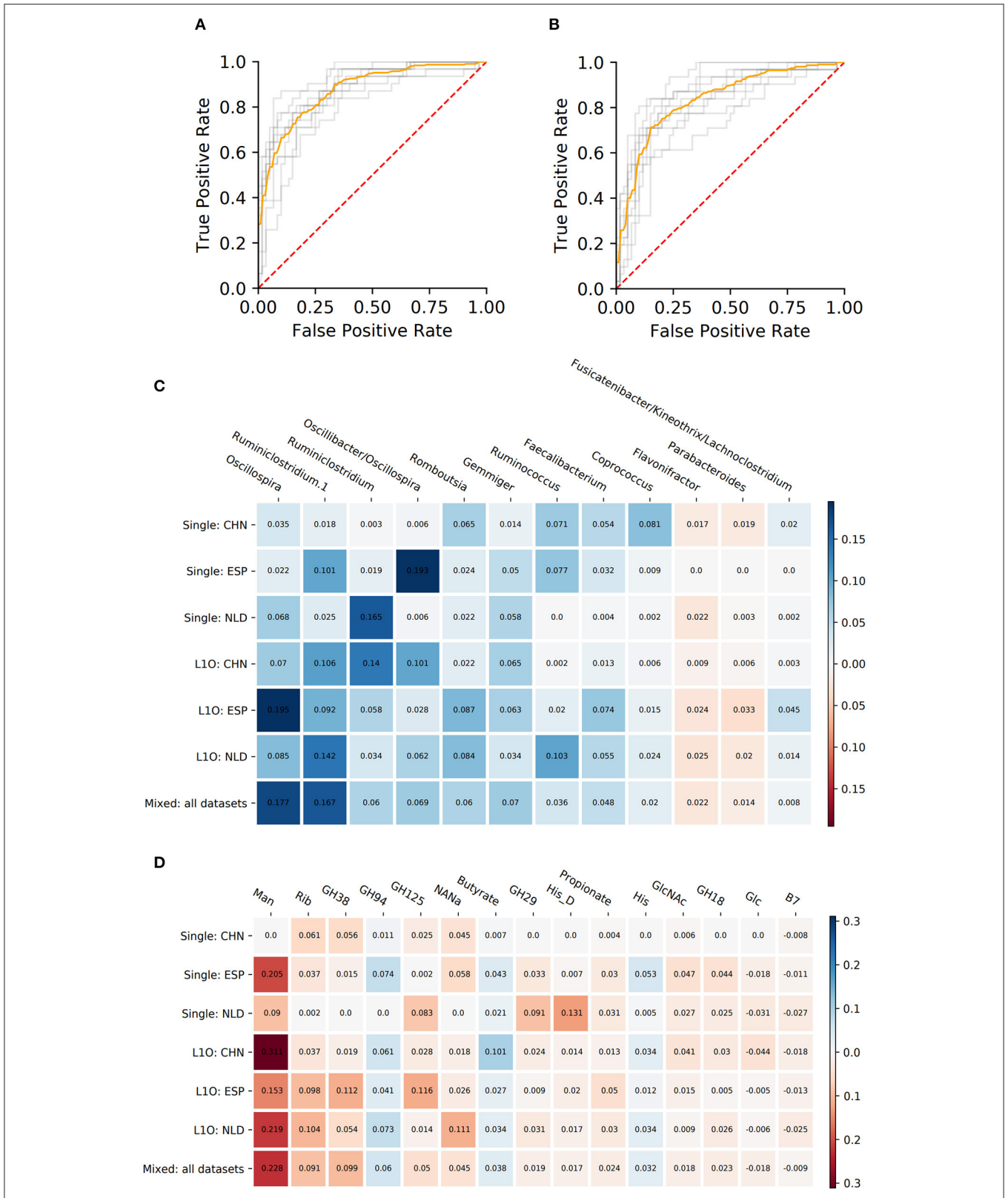

FIGURE 3 | Performance characteristics and stable predictors for the CD-vs.-HC classifier. **(A,B)** ROC curves for the Mixed strategy for CD-vs.-HC classifiers with taxonomic **(A)** and phenotypic **(B)** predictors. **(C,D)** Stable predictor importance values in different classification variants for taxonomic **(C)** and phenotypic **(D)** predictors. Color saturation corresponds to the mean importance of the predictor (see color key). Color hue corresponds to the direction of the difference between HC and CD means (blue—increased in HC, red—increased in CD).

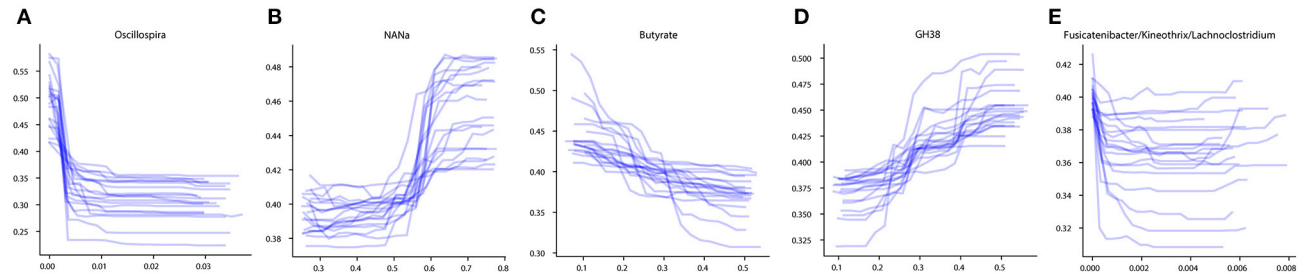

FIGURE 4 | Examples of different PDP forms. **(A)** Sharply decreasing PDP. **(B)** Sharply increasing PDP. **(C)** Smoothly decreasing PDP. **(D)** Smoothly increasing PDP. **(E)** Unclassified PDP. The *x*-axis of the plots denotes the taxonomic or phenotypic feature abundance, and the *y*-axis shows the probability of the CD classification outcome.

TABLE 4 | The forms of stable predictors PDPs (term "increasing" means increasing probability of the disease).

PDP form type	CD (taxonomy; phenotypes)	UC (taxonomy; phenotypes)
Sharply decreasing PDP	*Ruminiclostridium.1*, *Oscillospira*, *Oscillibacter/Oscillospira*, *Gemmiger*, *Romboutsia*, *Ruminococcus*, *Ruminiclostridium*, *Coprococcus*, *Fusicatenibacter/Kineothrix/Lachnoclostridium*, *Faecalibacterium*, GH94, His	*Gemmiger*, *Ruminococcus*, *Akkermansia*; –
Smoothly decreasing PDP	Butyrate	–; His, Arg
Sharply increasing PDP	NANa, Rib, GH18	–; –
Smoothly increasing PDP	*Flavonifractor*; Man, GH125, GH38, GlcNAc, GH29, Propionate, Glc, His_D, B7	*Streptococcus*, *Oscillibacter*, *Flavonifractor*, *Flintibacter;* Man, GH125
Unclassified	*Parabacteroides*	–; –

in the future when new phenotypic features are added to the CPP. Finally, being based on curated metabolic pathways refined for the species of a specific niche (here, the human gut), our method largely resolves the low-accuracy limitation of many existing metagenomic prediction methods mentioned in the Introduction. However, the limitation related to poor prediction of horizontally transferred genes is not targeted by our approach.

Another problem one deals with in metagenomics (and in ML in general) is high dimensionality of data with low sample count. In ML applications like computer vision, this is solved by the extensive use of data augmentation techniques which allow to drastically enlarge sample count. Unfortunately, they are not applicable to the microbiome studies due to the uniqueness of each sample. Therefore, the only remaining option is the reduction of feature space. Mixing of features by PCA-like methods is not suitable if one wants to preserve biological interpretability; therefore, aggregation of features into taxonomies or CPIs is a preferable choice for dimension reduction. However, in the latter case this aggregation may not be metabolism—but rather phylogeny—driven due to low phylogenetic diversity of organisms possessing a certain phenotype. This implies that the CPI value for such phenotypes would reflect, in the first place, the relative abundance of the phylogenetically narrow group of the phenotype carrier and would obstruct metabolism-driven interpretability. In order to eliminate phenotypic features with low phylogenetic diversity of their contributors from further analysis, we developed and applied a concept of Phenotype Alpha Diversity. Using the calculated PAD values, we filtered out the phenotypes with corresponding contributions to their CPI values coming from phylogenetically narrow groups of organisms, thus retaining phenotypic features which are robust for metabolism-driven inference.

We assessed the applicability of our methodology for differentiating microbiome samples from healthy subjects and patients, using the example of two most prominent IBD conditions, CD and UC, the diseases linked to profound shifts in the microbial community structure. First of all, the selected groups of phenotypes almost completely reflect information about the differences between the microbial communities' taxonomy of healthy people and patients with IBD. The difference in phenotype-based and taxonomy-based classifier performance characteristics (sensitivity, specificity, and AUC) was lower than 0.08 for each of the explored conditions (UC and CD) (**Figures 3A**, **5A**). Thus, we can conclude that despite the limited set of phenotypes selected for the analysis, in the case of IBD they contain most of the information about microbial signature of the investigated diseases.

The second question we wanted to answer was if the phenotypes could introduce new information useful for interpretation of disease relation to microbiome compared to taxonomy. For this purpose, the stable predictors obtained for taxonomic and phenotypic classifiers were compared. The introduction of new information using phenotypes is well demonstrated by the difference in forms and direction of stable predictors PDPs.

The majority of CD taxonomic stable predictors showed negative associations with CD output (**Figure 3C**). This is consistent with previous observations that IBD condition can

TABLE 5 | Mean performance characteristics of the taxonomy- and phenotype-based UC classifiers over 10 classification iterations.

Strategy: dataset	Taxonomy-based classifier			Phenotype-based classifier		
	Mean sensitivity	Mean specificity	Mean AUC	Mean sensitivity	Mean specificity	Mean AUC
Single:CHN	0.7909	0.9048	0.9277	0.8091	0.8857	0.9281
Single:ESP	0.3000	0.8667	0.6583	0.3250	0.7741	0.5815
Single:NLD	0.6300	0.9376	0.9240	0.5500	0.8678	0.8232
L1O:CHN	0.6324	0.5955	0.6791	0.4189	0.8791	0.6832
L1O:ESP	0.3051	0.8626	0.6607	0.3308	0.7714	0.5751
L1O:NLD	0.7354	0.4972	0.6944	0.8364	0.1022	0.4341
Mixed: all datasets	0.5970	0.8590	0.8153	0.5667	0.7836	0.7240
Mean across all variants	0.5701	0.7891	0.7656	0.5481	0.7234	0.6785

In the L1O strategy description, the name of the dataset corresponds to the test set (for example, the "L1O:CHN" description means that the classifier was trained on the ESP and NLD datasets and tested on the CHN dataset). Cell color reflects the characteristics' values (greater values correspond to darker colors).

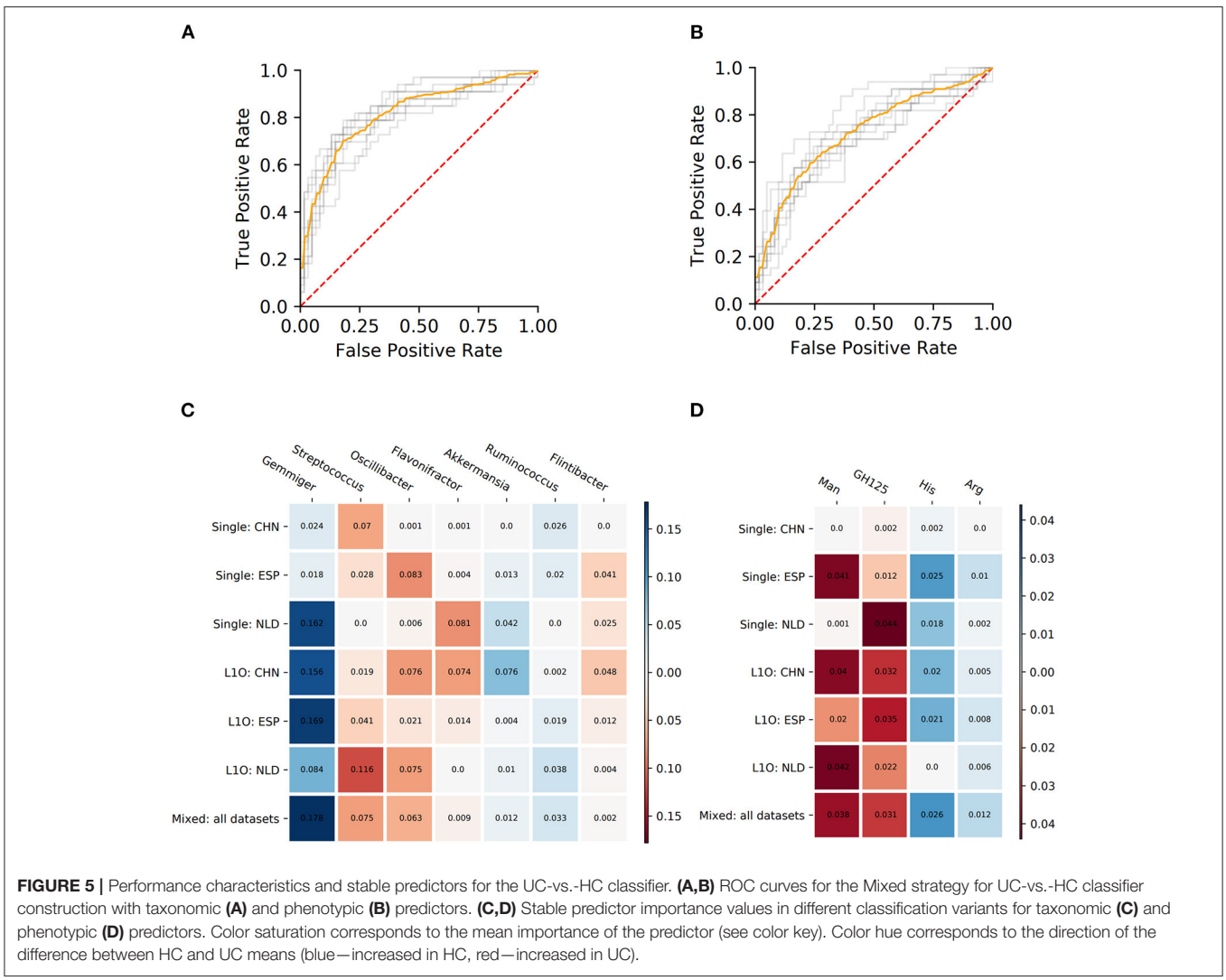

FIGURE 5 | Performance characteristics and stable predictors for the UC-vs.-HC classifier. **(A,B)** ROC curves for the Mixed strategy for UC-vs.-HC classifier construction with taxonomic **(A)** and phenotypic **(B)** predictors. **(C,D)** Stable predictor importance values in different classification variants for taxonomic **(C)** and phenotypic **(D)** predictors. Color saturation corresponds to the mean importance of the predictor (see color key). Color hue corresponds to the direction of the difference between HC and UC means (blue—increased in HC, red—increased in UC).

be characterized by the depletion of beneficial taxa rather than by the prevalence of pro-inflammatory ones (Duvallet et al., 2017; Wirbel et al., 2020). Such observations were made not only for IBD. More generally, the "Anna Karenina principle" was proposed in application to animal microbiomes: "healthy microbiomes are all alike; each unhealthy microbiome is unhealthy in its own way" (Zaneveld et al., 2017). This principle was proposed for the taxonomic composition of

microbiomes. However, it is unknown whether it can be implemented if we consider the functional potential of the community. The majority of CD taxonomic predictors showed a sharply decreasing PDP form with a sharp descent near zero abundance (**Supplementary Figure 2**). This suggests that the CD condition is associated with the complete absence of these taxa. Among such predictors, there is one of the main primary polysaccharide degraders of the human gut microbiome—*Ruminococcus* (Koropatkin et al., 2012). The primary degradation of fibers is essential for butyrate production. The group also includes other well-known [*Coprococcus* (Pryde et al., 2002; Louis et al., 2014), *Faecalibacterium*] and potential [*Oscillospira* (Gophna et al., 2017)] butyrate producers. *Faecalibacterium* is known for its anti-inflammatory properties, being depleted in the gut of CD patients (Quévrain et al., 2016). Interestingly, the production of butyrate is not thought to be the key anti-inflammatory feature of the bacterium (Sokol et al., 2008; Miquel et al., 2013; Quévrain et al., 2016). It was suggested that *F. prausnitzii* can influence immune response through the production of other metabolites (Sokol et al., 2008; Breyner et al., 2017). Only one taxon, *Flavonifractor*, showed an increasing PDP form. The taxon is known to have proinflammatory properties and was previously shown to be increased in some CD patients (Tyakht et al., 2018).

Unlike the taxonomic predictors for the CD classifier, the majority of phenotypic predictors show increasing PDP forms (**Supplementary Figure 3**). It can be suggested that each increasing phenotypic predictor is a functional representation of few increasing taxonomic predictors. However, the filter applied to the diversity of phenotypes excludes the use of phenotypes represented in a small number of taxa. Thus, we can speculate that despite the fact that the taxonomic signature is composed mainly of commensal taxa, functionally the microbiome of patients is characterized to a greater extent by pro-inflammatory predictors. In this case, we see that the "Anna Karenina principle" for taxonomic composition of the microbiome (Zaneveld et al., 2017) is not fulfilled for its functional potential. It was previously shown that a healthy gut microbial community is characterized by functional homeostasis. This means that despite the differences in the taxonomic composition of different people communities, the metabolic potential of these communities is quite similar (Eng and Borenstein, 2018). Our observations in the case of Crohn's disease support this concept. Taxonomically, the imbalance included the increase of various taxa in different individuals, but functionally, we see the increase of the same functions. Thus, the application of microbial phenotypes allowed us to identify universal markers qualitatively different from taxonomic ones and to provide a new layer of information for further interpretations.

The phenotypic stable predictors positively associated with CD were generally linked to degradation of host-derived carbohydrates (**Figure 3D**). Firstly, it was evident at the level of specific families of glycoside hydrolases; thus, we observed an increase of microbial community representation of the GH38, GH18, and GH125 families involved in catabolism of N-linked glycans that are constituents of the host mucus (El Kaoutari et al., 2013; Engevik et al., 2019). Another family of glycosyl hydrolases with increased representation in CD samples, GH29, includes exo-acting α-fucosidases that are involved in the degradation of O-linked glycans, in particular mucin (Tailford et al., 2015). In particular, GH29 is present in the mucin-dwelling bacteria from the genera *Ruminococcus* (Crost et al., 2013) and *Akkermansia* (El Kaoutari et al., 2013). These changes were reflected by the increased propensity toward utilization of five monosaccharides including neuraminic acid (NANa), N-acetylglucosamine (GlcNAc), mannose (Man), and glucose (Glc) that constitute O- and N-linked glycans, as well as ribose (Rib). The latter monosaccharide is utilized by many gut bacteria such as *Bacteroides* spp., while ribose and ribose-containing molecules such as nucleic acids may serve as nutrients for these gut symbionts (Glowacki et al., 2020). The increased potential of mannose metabolism was reported in patients with ileal CD (Morgan et al., 2012). The list of the phenotypes increased in CD samples also included histidine (His) amino acid degradation—apparently, reflecting the high availability of host-derived amino acids from the inflamed tissue. Further, the increased vitamin B7 (biotin) synthesis potential was in line with the reported upregulation of the respective biosynthetic enzymes in stool metatranscriptomes of IBD patients (Das et al., 2019). Overall, these observations suggest that the CD-associated microbiome is prominent by its ability to degrade the mucus, apparently due to highly inflammatory milieu and excessive shedding of intestinal epithelial cells (Png et al., 2010; Blander, 2016). Another phenotypic feature increased in CD was propionate production potential. Although considered to be anti-inflammatory as butyrate (Tedelind et al., 2007), some of its effects on the immune cells are opposite to the ones of butyrate (Cavaglieri et al., 2003). Recently, propionate was shown to promote the virulent properties of CD-associated *Escherichia coli* (Ormsby et al., 2020; Pobeguts et al., 2020), the key taxon linked to the disease in previous studies.

On the other hand, there were only a few phenotypic stable features negatively associated with CD, including the butyrate production, histidine biosynthesis, and the GH94 family of glycoside hydrolases (**Figure 3D**). The butyrate is one of the most studied beneficial microbiome-derived metabolites with anti-inflammatory potential. First of all, it serves as an energy source for the colonic epithelium, therefore preventing mucosal atrophy (Hamer et al., 2008). In addition, butyrate possesses some direct immuno-modulatory effects like suppression of nuclear factor kappa B (NF-κB) activation (Hamer et al., 2008), signaling through G-protein-coupled receptors (Hamer et al., 2008) and GPR109A (Singh et al., 2014). Interestingly, among the stable predictors of CD, most taxa known as butyrate producers manifested sharp PDP form, while the butyrate production phenotype CPI was smooth—apparently reflecting the averaging of their contributions. Depletion of the histidine synthesis potential in CD samples is in line with the abovementioned increase of the histidine degradation phenotype. Strikingly, unlike GH families positively associated with CD that were mainly involved in mucus degradation, the GH94 family phosphorylases that cleave β-glycosidic bonds in cellobiose and cellodextrin are involved in plant cell wall degradation (Cantarel et al., 2012).

Ulcerative colitis, the second major IBD condition we investigated, is characterized by a less pronounced microbiome disruption compared with the Crohn's disease—resulting in a generally lower classification performance. It is in line with the previous reports (Halfvarson et al., 2017; Imhann et al., 2018; Franzosa et al., 2019; Clooney et al., 2020). The reason for this can be the different epidemiological and clinicopathological pictures of the diseases. In terms of epidemiology, it was hypothesized that in Crohn's disease etiology, the early-life abnormal cross-talk between microbiome and immune system plays the essential role, while in ulcerative colitis it is dysbiosis that occurred at any time of life (Beaugerie et al., 2018). In terms of clinical picture differences, in ulcerative colitis inflammation foci are located in rectum and colon, while Crohn's disease can also involve the upper parts of the gastrointestinal tract. IBD location was shown to affect dysbiosis pictures of the diseases (Imhann et al., 2018).

The UC classifiers demonstrated generally lower performance compared to the CD and produced fewer stable predictors, most of which are shared with CD predictors. In hand with the lower performance of the UC-vs.-HC classifier, the identified stable predictors were also less reliable compared to the CD stable predictors. Three taxonomic genera, *Gemmiger*, *Flavonifractor*, and *Ruminococcus*, are shared between the CD and UC predictors. Among microbial genera that are specific for the UC-vs.-HC classifier, there were the mucin-degrading *Akkermansia* with sharply decreasing PDP form, and three taxa with smoothly increasing PDP form which are *Streptococcus*, *Oscillibacter*, and *Flintibacter*. Mucolytic taxa were previously shown to be enriched in the intestinal epithelium of IBD patients compared to healthy controls (Png et al., 2010), with the only exception for *Akkermansia*, which had significantly higher abundance in control samples, in line with our findings. Several other studies showed the protective role of *Akkermansia* in IBD (Bian et al., 2019; Earley et al., 2019). Interestingly, neither the butyrate-producing taxa nor the phenotype of butyrate synthesis itself were detected among the stable predictors for UC. It resonates with the previous studies showing that microbiome butyrate-synthetic capacity was reduced only in patients with active UC, but not in ones with inactive stage of disease, while for Crohn's disease, the association was observed for both stages (Lasserna-Mendieta et al., 2018). The list of phenotypic stable predictors positively associated with UC is shorter than the respective list for the CD and overlaps with the latter by including the mannose utilization (Man) and the glycoside hydrolase family GH125 of exo-mannosidases involved in N-glycan degradation. However, while for the CD we observed a sharply increasing PDP form for Man (with the threshold around 0.2–0.3 phenotype abundance), for the UC the form was smoothly increasing. The decrease of histidine (His) and arginine (Arg) amino acid synthesis potentials in UC samples is likely linked to a higher abundance of free amino acids originating from the inflamed host tissue in the gut.

We discovered that the classification performance varies across different studies. A classifier trained on one geographic population might be not that precise for a cohort from another country. Besides possible technological differences, this effect could be contributed to by the geography-specific features not only of the healthy microbiome composition but also in the patients with diseases. Similar effects were reported before, e.g., for type 2 diabetes (Karlsson et al., 2013). Further extended regional multicenter studies with large cohorts of healthy and affected subjects are required to elucidate the universal character of microbial phenotypes' robustness and concordance with the taxonomic features.

CONCLUSIONS

We developed a novel computational approach that uses a concept of metabolic phenotypes toward the microbiome-based classification of clinical status and assessed its performance using 16S amplicon sequencing data from multiple IBD studies. Although the set of assessed metabolic functions and pathways was limited to metabolism of sugars (including GH enzymes involved in polysaccharide degradation and SCFA production pathways), amino acids, and vitamins, our results suggest that the performance of the microbial phenotype-based classification was comparable to the state-of-art taxonomic approach.

Feature design in machine learning algorithms, which is based on cumulative metabolic potential of microbiome species (estimated via CPIs), can provide additional functional insights on the deviations of microbiome from homeostasis in disease. To provide truly metabolism-driven inference, these metabolic features are likely to account for the collective action of a phylogenetically diverse community of a particular phenotype carrier, which is reflected in the Phenotype Alpha Diversity (PAD) metric. In CD, while the community structure was characterized by the depletion of many commensal taxa rather than the presence of specific opportunists, the functional imbalance was revealed as an enrichment of inflammation-related phenotypes not reflected at the taxonomic level.

The major indicators of functional imbalance of microbiome in IBD reflect the adaptation to the inflammatory environment by including increased potential for degradation of mucin-derived carbohydrates and amino acids and propionate synthesis, while the healthy gut is characterized by enriched degradation of dietary complex carbohydrates and synthesis of butyrate and amino acids. Analysis of the abundance-dependent contribution of each feature to the classification outcome using PDP suggests that the presence of most taxa negatively associated with IBD is more important than their abundance. Further, the PDP patterns reflect how a plethora of taxa (each showing sharply decreasing PDP form in IBD) can functionally "convolve" into a single phenotype (with smoothly decreasing PDP), as exemplified by the case of butyrate producers.

Our results show that the analysis based on microbial phenotypes can provide interpretable insights into the host–microbiome mechanisms of disease. Extension of the phenotype list to include metabolism of specific polysaccharides, lipids, and bile acids will provide further insights into the possible mechanisms of gut microbiome metabolic contribution to the risks and onset and development of the disease. Further expansion of the reference microbial genomes database with predicted metabolic phenotypes will allow one to apply the phenotype profiling approach to microbial communities of other

human body sites, and more generally to various environmental and industrial microbiomes.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: www.ebi.ac.uk/ena: project IDs PRJNA422193 and PRJEB22028, ega-archive.org: project IDs EGAS00001002702 and EGAS00001001704.

AUTHOR CONTRIBUTIONS

AT and DR conceived and designed the research project. TS performed primary analysis of the sequencing data. SI, PN, and DR developed CPI/MTA/PAD concepts and analyzed metabolic phenotypes. NK, SI, DE, and AT performed machine learning analysis of the obtained taxonomic and metabolic phenotype profiles. NK, SI, DR, and AT wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This research was supported by the Russian Science Foundation (grant #19-14-00305). NK, DE, and AT acknowledge their finances from Knomics LLC. NK and AT were supported by a grant 075-15-2019-1661 from the Ministry of Science and Higher Education of the Russian Federation allocated to the Center for Precision Genome Editing and Genetic Technologies for Biomedicine under Federal Research Programme for Genetic Technologies Development for 2019–2027. The authors declare that this study received funding from Atlas Biomed Group—Knomics LLC. The funder had the following involvement with the study: study design, writing of the article, and the decision to submit for publication.

ACKNOWLEDGMENTS

We are grateful to PhenoBiome Inc. (San Francisco, CA) for providing us with the access to the Phenotype Profiler tool. The authors would also like to acknowledge the LifeLines DEEP consortium for providing access to the LifeLines DEEP cohort data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2020.603740/full#supplementary-material

REFERENCES

Arzamasov, A. A., van Sinderen, D., and Rodionov, D. A. (2018). comparative genomics reveals the regulatory complexity of bifidobacterial arabinose and arabino-oligosaccharide utilization. *Front. Microbiol.* 9:776. doi: 10.3389/fmicb.2018.00776

Aßhauer, K. P., Wemheuer, B., Daniel, R., and Meinicke, P. (2015). Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics* 31, 2882–2884. doi: 10.1093/bioinformatics/btv287

Bauer, E., and Thiele, I. (2018). From metagenomic data to personalized *in silico* microbiotas: predicting dietary supplements for Crohn's disease. *NPJ Syst. Biol. Appl.* 4, 27. doi: 10.1038/s41540-018-0063-2

Beaugerie, L., Langholz, E., Nyboe-Andersen, N., Pigneur, B., Sokol, H., and ECCO Epicom. (2018). Differences in epidemiological features between ulcerative colitis and Crohn's disease: the early life-programmed versus late dysbiosis hypothesis. *Med. Hypotheses* 115, 19–21. doi: 10.1016/j.mehy.2018.03.009

Bian, X., Wu, W., Yang, L., Lv, L., Wang, Q., Li, Y., et al. (2019). Administration of *Akkermansia muciniphila* ameliorates dextran sulfate sodium-induced ulcerative colitis in mice. *Front. Microbiol.* 10:2259. doi: 10.3389/fmicb.2019.02259

Bilen, M., Dufour, J.-C., Lagier, J.-C., Cadoret, F., Daoud, Z., Dubourg, G., et al. (2018). The contribution of culturomics to the repertoire of isolated human bacterial and archaeal species. *Microbiome* 6, 1–11. doi: 10.1186/s40168-018-0485-5

Blander, J. M. (2016). Death in the intestinal epithelium-basic biology and implications for inflammatory bowel disease. *FEBS J.* 283, 2720–2730. doi: 10.1111/febs.13771

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., et al. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* 6:90. doi: 10.1186/s40168-018-0470-z

Boursier, J., Mueller, O., Barret, M., Machado, M., Fizanne, L., Araujo-Perez, F., et al. (2016). The severity of nonalcoholic fatty liver disease is associated with gut dysbiosis and shift in the metabolic function of the gut microbiota. *Hepatology* 63, 764–775. doi: 10.1002/hep.28356

Bouvier, J. T., Sernova, N. V., Ghasempur, S., Rodionova, I. A., Vetting, M. W., Al-Obaidi, N. F., et al. (2019). Novel metabolic pathways and regulons for hexuronate utilization in proteobacteria. *J. Bacteriol.* 201, e00431–0018. doi: 10.1128/JB.00431-18

Breyner, N. M., Michon, C., de Sousa, C. S., Vilas Boas, P. B., Chain, F., Azevedo, V. A., et al. (2017). Microbial anti-inflammatory molecule (MAM) from *Faecalibacterium prausnitzii* shows a protective effect on DNBS and DSS-induced colitis model in mice through inhibition of NF-κB pathway. *Front. Microbiol.* 8:114. doi: 10.3389/fmicb.2017.00114

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13, 581–583. doi: 10.1038/nmeth.3869

Cantarel, B. L., Lombard, V., and Henrissat, B. (2012). Complex carbohydrate utilization by the healthy human microbiome. *PLoS One* 7:e28742. doi: 10.1371/journal.pone.0028742

Cavaglieri, C. R., Nishiyama, A., Fernandes, L. C., Curi, R., Miles, E. A., and Calder, P. C. (2003). Differential effects of short-chain fatty acids on proliferation and production of pro- and anti-inflammatory cytokines by cultured lymphocytes. *Life Sci.* 73, 1683–1690. doi: 10.1016/S0024-3205(03)00490-9

Cerqueira, F. M., Photenhauer, A. L., Pollet, R. M., Brown, H. A., and Koropatkin, N. M. (2020). Starch digestion by gut bacteria: crowdsourcing for carbs. *Trends Microbiol.* 28, 95–108. doi: 10.1016/j.tim.2019.09.004

Cirstea, M. S., Yu, A. C., Golz, E., Sundvick, K., Kliger, D., Radisavljevic, N., et al. (2020). Microbiota composition and metabolism are associated with gut function in Parkinson's disease. *Mov. Disord.* 35, 1208–1217. doi: 10.1002/mds.28052

Clooney, A. G., Eckenberger, J., Laserna-Mendieta, E., Sexton, K. A., Bernstein, M. T., Vagianos, K., et al. (2020). Ranking microbiome variance in inflammatory bowel disease: a large longitudinal intercontinental study. *Gut.* doi: 10.1136/gutjnl-2020-321106

Cockburn, D. W., and Koropatkin, N. M. (2016). Polysaccharide degradation by the intestinal microbiota and its influence on human health and disease. *J. Mol. Biol.* 428, 3230–3252. doi: 10.1016/j.jmb.2016.06.021

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., et al. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* 42, D633–D642. doi: 10.1093/nar/gkt1244

Crost, E. H., Tailford, L. E., Le Gall, G., Fons, M., Henrissat, B., and Juge, N. (2013). Utilisation of mucin glycans by the human gut symbiont Ruminococcus gnavus is strain-dependent. *PLoS One* 8:e76341. doi: 10.1371/journal.pone.0076341

Das, P., Babaei, P., and Nielsen, J. (2019). Metagenomic analysis of microbe-mediated vitamin metabolism in the human gut microbiome. *BMC Genomics* 20:208. doi: 10.1186/s12864-019-5591-7

Davila, A.-M., Blachier, F., Gotteland, M., Andriamihaja, M., Benetti, P.-H., Sanz, Y., et al. (2013). Intestinal luminal nitrogen metabolism: role of the gut microbiota and consequences for the host. *Pharmacol. Res*. 68, 95–107. doi: 10.1016/j.phrs.2012.11.005

Douglas, G. M., Maffei, V. J., Zaneveld, J., Yurgel, S. N., Brown, J. R., Taylor, C. M., et al. (2020). PICRUSt2: An improved and customizable approach for metagenome inference. *BioRxiv*. doi: 10.1101/672295. [Epub ahead of print].

Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A., and Alm, E. J. (2017). Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat. Commun*. 8:1784. doi: 10.1038/s41467-017-01973-8

Earley, H., Lennon, G., Balfe, Á., Coffey, J. C., Winter, D. C., and O'Connell, P. R. (2019). The abundance of *Akkermansia muciniphila* and its relationship with sulphated colonic mucins in health and ulcerative colitis. *Sci. Rep*. 9:15683. doi: 10.1038/s41598-019-51878-3

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32, 1792–1797. doi: 10.1093/nar/gkh340

Efimova, D., Tyakht, A., Popenko, A., Vasilyev, A., Altukhov, I., Dovidchenko, N., et al. (2018). Knomics-Biota - a system for exploratory analysis of human gut microbiota data. *BioData Min*. 11:25. doi: 10.1186/s13040-018-0187-3

El Kaoutari, A., Armougom, F., Gordon, J. I., Raoult, D., and Henrissat, B. (2013). The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nat. Rev. Microbiol*. 11, 497–504. doi: 10.1038/nrmicro3050

Elmén, L., Zlamal, J. E., Scott, D. A., Lee, R. B., Chen, D. J., Colas, A. R., et al. (2020). Dietary emulsifier sodium stearoyl lactylate alters gut microbiota *in vitro* and inhibits bacterial butyrate producers. *Front. Microbiol*. 11:892. doi: 10.3389/fmicb.2020.00892

Eng, A., and Borenstein, E. (2018). Taxa-function robustness in microbial communities. *Microbiome* 6:45. doi: 10.1186/s40168-018-0425-4

Engevik, M. A., Luk, B., Chang-Graham, A. L., Hall, A., Herrmann, B., Ruan, W., et al. (2019). *Bifidobacterium dentium* fortifies the intestinal mucus layer via autophagy and calcium signaling pathways. *MBio* 10:e01087-19. doi: 10.1128/mBio.01087-19

Feng, L., Raman, A. S., Hibberd, M. C., Cheng, J., Griffin, N. W., Peng, Y., et al. (2020). Identifying determinants of bacterial fitness in a model of human gut microbial succession. *Proc. Natl. Acad. Sci. U. S. A*. 117, 2622–2633. doi: 10.1073/pnas.1918951117

Forster, S. C., Kumar, N., Anonye, B. O., Almeida, A., Viciani, E., Stares, M. D., et al. (2019). A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nat. Biotechnol*. 37, 186–192. doi: 10.1038/s41587-018-0009-7

Franzosa, E. A., Sirota-Madi, A., Avila-Pacheco, J., Fornelos, N., Haiser, H. J., Reinker, S., et al. (2019). Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat Microbiol.* 4, 293–305. doi: 10.1038/s41564-018-0306-4

Garza, D. R., van Verk, M. C., Huynen, M. A., and Dutilh, B. E. (2018). Towards predicting the environmental metabolome from metagenomics with a mechanistic model. *Nat. Microbiol.* 3, 456–460. doi: 10.1038/s41564-018-0124-8

Gibson, G. R., Hutkins, R., Sanders, M. E., Prescott, S. L., Reimer, R. A., Salminen, S. J., et al. (2017). Expert consensus document: the International Scientific Association for Probiotics and Prebiotics (ISAPP) consensus statement on the definition and scope of prebiotics. *Nat. Rev. Gastroenterol. Hepatol*. 14, 491–502. doi: 10.1038/nrgastro.2017.75

Glowacki, R. W. P., Pudlo, N. A., Tuncil, Y., Luis, A. S., Sajjakulnukit, P., Terekhov, A. I., et al. (2020). A ribose-scavenging system confers colonization fitness on the human gut symbiont bacteroides thetaiotaomicron in a diet-specific manner. *Cell Host Microbe* 27, 79–92.e9. doi: 10.1016/j.chom.2019.11.009

Gophna, U., Konikoff, T., and Nielsen, H. B. (2017). Oscillospira and related bacteria - From metagenomic species to metabolic features. *Environ. Microbiol*. 19, 835–841. doi: 10.1111/1462-2920.13658

Halfvarson, J., Brislawn, C. J., Lamendella, R., Vázquez-Baeza, Y., Walters, W. A., Bramer, L. M., et al. (2017). Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat. Microbiol.* 2:17004. doi: 10.1038/nmicrobiol.2017.4

Hamer, H. M., Jonkers, D., Venema, K., Vanhoutvin, S., Troost, F. J., and Brummer, R.-J. (2008). Review article: the role of butyrate on colonic function. *Aliment. Pharmacol. Ther*. 27, 104–119. doi: 10.1111/j.1365-2036.2007.03562.x

Imhann, F., Vich Vila, A., Bonder, M. J., Fu, J., Gevers, D., Visschedijk, M. C., et al. (2018). Interplay of host genetics and gut microbiota underlying the onset and clinical presentation of inflammatory bowel disease. *Gut* 67, 108–119. doi: 10.1136/gutjnl-2016-312135

Jones, R., Berger, P. K., Plows, J. F., Alderete, T. L., Millstein, J., Fogel, J., et al. (2020). Lactose-reduced infant formula with added corn syrup solids is associated with a distinct gut microbiota in hispanic infants. *Gut Microbes, in press*. doi: 10.1080/19490976.2020.1813534

Karlsson, F. H., Tremaroli, V., Nookaew, I., Bergström, G., Behre, C. J., Fagerberg, B., et al. (2013). Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* 498, 99–103. doi: 10.1038/nature12198

Khoroshkin, M. S., Leyn, S. A., Van Sinderen, D., and Rodionov, D. A. (2016). Transcriptional regulation of carbohydrate utilization pathways in the bifidobacterium genus. *Front. Microbiol*. 7:120. doi: 10.3389/fmicb.2016.00120

Klimenko, N. S., Tyakht, A. V., Popenko, A. S., Vasiliev, A. S., Altukhov, I. A., Ischenko, D. S., et al. (2018). Microbiome responses to an uncontrolled short-term diet intervention in the frame of the citizen science project. *Nutrients* 10:576. doi: 10.3390/nu10050576

Koh, A., De Vadder, F., Kovatcheva-Datchary, P., and Bäckhed, F. (2016). From dietary fiber to host physiology: short-chain fatty acids as key bacterial metabolites. *Cell* 165, 1332–1345. doi: 10.1016/j.cell.2016.05.041

Koropatkin, N. M., Cameron, E. A., and Martens, E. C. (2012). How glycan metabolism shapes the human gut microbiota. *Nat. Rev. Microbiol*. 10, 323–335. doi: 10.1038/nrmicro2746

Laserna-Mendieta, E. J., Clooney, A. G., Carretero-Gomez, J. F., Moran, C., Sheehan, D., Nolan, J. A., et al. (2018). Determinants of reduced genetic capacity for butyrate synthesis by the gut microbiome in Crohn's disease and ulcerative colitis. *J. Crohns. Colitis* 12, 204–216. doi: 10.1093/ecco-jcc/jjx137

Lavelle, A., and Sokol, H. (2020). Gut microbiota-derived metabolites as key actors in inflammatory bowel disease. *Nat. Rev. Gastroenterol. Hepatol*. 17, 223–237. doi: 10.1038/s41575-019-0258-z

LeBlanc, J. G., Milani, C., de Giori, G. S., Sesma, F., van Sinderen, D., and Ventura, M. (2013). Bacteria as vitamin suppliers to their host: a gut microbiota perspective. *Curr. Opin. Biotechnol*. 24, 160–168. doi: 10.1016/j.copbio.2012.08.005

Leyn, S. A., Maezato, Y., Romine, M. F., and Rodionov, D. A. (2017). Genomic reconstruction of carbohydrate utilization capacities in microbial-mat derived consortia. *Front. Microbiol*. 8:1304. doi: 10.3389/fmicb.2017.01304

Liu, Y., Wang, X., and Hu, C.-A. A. (2017). Therapeutic potential of amino acids in inflammatory bowel disease. *Nutrients* 9:920. doi: 10.3390/nu9090920

Louca, S., Parfrey, L. W., and Doebeli, M. (2016). Decoupling function and taxonomy in the global ocean microbiome. *Science* 353, 1272–1277. doi: 10.1126/science.aaf4507

Louis, P., Hold, G. L., and Flint, H. J. (2014). The gut microbiota, bacterial metabolites and colorectal cancer. *Nat. Rev. Microbiol*. 12, 661–672. doi: 10.1038/nrmicro3344

McCann, A., Jeffery, I. B., Ouliass, B., Ferland, G., Fu, X., Booth, S. L., et al. (2019). Exploratory analysis of covariation of microbiota-derived vitamin K and cognition in older adults. *Am. J. Clin. Nutr*. 110, 1404–1415. doi: 10.1093/ajcn/nqz220

Miquel, S., Martín, R., Rossi, O., Bermúdez-Humarán, L. G., Chatel, J. M., Sokol, H., et al. (2013). *Faecalibacterium prausnitzii* and human intestinal health. *Curr. Opin. Microbiol*. 16, 255–261. doi: 10.1016/j.mib.2013.06.003

Morgan, X. C., Tickle, T. L., Sokol, H., Gevers, D., Devaney, K. L., Ward, D. V., et al. (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol*. 13:R79. doi: 10.1186/gb-2012-13-9-r79

Narayan, N. R., Weinmaier, T., Laserna-Mendieta, E. J., Claesson, M. J., Shanahan, F., Dabbagh, K., et al. (2020). Piphillin predicts metagenomic composition and dynamics from DADA2-corrected 16S rDNA sequences. *BMC Genomics* 21:56. doi: 10.1186/s12864-020-6537-9

Ormsby, M. J., Johnson, S. A., Carpena, N., Meikle, L. M., Goldstone, R. J., McIntosh, A., et al. (2020). Propionic Acid promotes the virulent phenotype of Crohn's disease-associated adherent-invasive *Escherichia coli*. *Cell Rep.* 30, 2297–305.e5. doi: 10.1016/j.celrep.2020.01.078

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H.-Y., Cohoon, M., et al. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* 33, 5691–5702. doi: 10.1093/nar/gki866

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., et al. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* 42, D206–D214. doi: 10.1093/nar/gkt1226

Pascal, V., Pozuelo, M., Borruel, N., Casellas, F., Campos, D., Santiago, A., et al. (2017). A microbial signature for Crohn's disease. *Gut* 66, 813–822. doi: 10.1136/gutjnl-2016-313235

Peterson, C. T., Sharma, V., Iablokov, S. N., Albayrak, L., Khanipov, K., Uchitel, S., et al. (2019). 16S rRNA gene profiling and genome reconstruction reveal community metabolic interactions and prebiotic potential of medicinal herbs used in neurodegenerative disease and as nootropics. *PLoS One* 14:e0213869. doi: 10.1371/journal.pone.0213869

Png, C. W., Lindén, S. K., Gilshenan, K. S., Zoetendal, E. G., McSweeney, C. S., Sly, L. I., et al. (2010). Mucolytic bacteria with increased prevalence in IBD mucosa augment *in vitro* utilization of mucin by other bacteria. *Am. J. Gastroenterol.* 105:2420. doi: 10.1038/ajg.2010.281

Pobeguts, O. V., Ladygina, V. G., Evsyutina, D. V., Eremeev, A. V., Zubov, A. I., Matyushkina, D. S., et al. (2020). Propionate induces virulent properties of crohn's disease-associated *Escherichia coli*. *Front. Microbiol.* 11:1460. doi: 10.3389/fmicb.2020.01460

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 - approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490. doi: 10.1371/journal.pone.0009490

Prodan, A., Tremaroli, V., Brolin, H., Zwinderman, A. H., Nieuwdorp, M., and Levin, E. (2020). Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PLoS ONE* 15:e0227434. doi: 10.1371/journal.pone.0227434

Pryde, S. E., Duncan, S. H., Hold, G. L., Stewart, C. S., and Flint, H. J. (2002). The microbiology of butyrate formation in the human colon. *FEMS Microbiol. Lett.* 217, 133–139. doi: 10.1111/j.1574-6968.2002.tb11467.x

Quévrain, E., Maubert, M. A., Michon, C., Chain, F., Marquant, R., Tailhades, J., et al. (2016). Identification of an anti-inflammatory protein from Faecalibacterium prausnitzii, a commensal bacterium deficient in Crohn's disease. *Gut* 65, 415–425. doi: 10.1136/gutjnl-2014-307649

Ravcheev, D. A., Godzik, A., Osterman, A. L., and Rodionov, D. A. (2013). Polysaccharides utilization in human gut bacterium Bacteroides thetaiotaomicron: comparative genomics reconstruction of metabolic and regulatory networks. *BMC Genomics* 14:873. doi: 10.1186/1471-2164-14-873

Rodionov, D. A., Arzamasov, A. A., Khoroshkin, M. S., Iablokov, S. N., Leyn, S. A., Peterson, S. N., et al. (2019). Micronutrient requirements and sharing capabilities of the human gut microbiome. *Front. Microbiol.* 10:1316. doi: 10.3389/fmicb.2019.01316

Rodionov, D. A., Novichkov, P. S., Stavrovskaya, E. D., Rodionova, I. A., Li, X., Kazanov, M. D., et al. (2011). Comparative genomic reconstruction of transcriptional networks controlling central metabolism in the *Shewanella genus*. *BMC Genomics* 12(Suppl 1):S3. doi: 10.1186/1471-2164-12-S1-S3

Rodionov, D. A., Rodionova, I. A., Li, X., Ravcheev, D. A., Tarasova, Y., Portnoy, V. A., et al. (2013). Transcriptional regulation of the carbohydrate utilization network in *Thermotoga maritima*. *Front. Microbiol.* 4:244. doi: 10.3389/fmicb.2013.00244

Sharma, V., Rodionov, D. A., Leyn, S. A., Tran, D., Iablokov, S. N., Ding, H., et al. (2019). B-vitamin sharing promotes stability of gut microbial communities. *Front. Microbiol.* 10:1485. doi: 10.3389/fmicb.2019.01485

Singh, N., Gurav, A., Sivaprakasam, S., Brady, E., Padia, R., Shi, H., et al. (2014). Activation of Gpr109a, receptor for niacin and the commensal metabolite butyrate, suppresses colonic inflammation and carcinogenesis. *Immunity* 40, 128–139. doi: 10.1016/j.immuni.2013.12.007

Sokol, H., Pigneur, B., Watterlot, L., Lakhdari, O., Bermúdez-Humarán, L. G., Gratadoux, J.-J., et al. (2008). *Faecalibacterium prausnitzii* is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc. Natl. Acad. Sci. U. S. A.* 105, 16731–16736. doi: 10.1073/pnas.0804812105

Sugimoto, Y., Camacho, F. R., Wang, S., Chankhamjon, P., Odabas, A., Biswas, A., et al. (2019). A metagenomic strategy for harnessing the chemical repertoire of the human microbiome. *Science*. 366:eaax9176. doi: 10.1126/science.aax9176

Sung, J., Kim, S., Cabatbat, J. J. T., Jang, S., Jin, Y.-S., Jung, G. Y., et al. (2017). Global metabolic interaction network of the human gut microbiota for context-specific community-scale analysis. *Nat. Commun.* 8:15393. doi: 10.1038/ncomms15393

Tailford, L. E., Crost, E. H., Kavanaugh, D., and Juge, N. (2015). Mucin glycan foraging in the human gut microbiome. *Front. Genet.* 6:81. doi: 10.3389/fgene.2015.00081

Tedelind, S., Westberg, F., Kjerrulf, M., and Vidal, A. (2007). 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. doi: 10.3748/wjg.v13.i20.2826

Tigchelaar, E. F., Zhernakova, A., Dekens, J. A. M., Hermes, G., Baranska, A., Mujagic, Z., et al. (2015). Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open* 5:e006772. doi: 10.1136/bmjopen-2014-006772

Tyakht, A. V., Manolov, A. I., Kanygina, A. V., Ischenko, D. S., Kovarsky, B. A., Popenko, A. S., et al. (2018). Genetic diversity of *Escherichia coli* in gut microbiota of patients with Crohn's disease discovered using metagenomic and genomic analyses. *BMC Genomics* 19:968. doi: 10.1186/s12864-018-5306-5

Vital, M., Howe, A. C., and Tiedje, J. M. (2014). Revealing the bacterial butyrate synthesis pathways by analyzing (meta)genomic data. *MBio* 5:e00889. doi: 10.1128/mBio.00889-14

Volokh, O., Klimenko, N., Berezhnaya, Y., Tyakht, A., Nesterova, P., Popenko, A., et al. (2019). Human gut microbiome response induced by fermented dairy product intake in healthy volunteers. *Nutrients* 11:547. doi: 10.3390/nu11030547

Wirbel, J., Zych, K., Essex, M., Karcher, N., and Kartal, E. (2020). SIAMCAT: user-friendly and versatile machine learning workflows for statistically rigorous microbiome analyses. *bioRxiv*. doi: 10.1101/2020.02.06.931808

Yarygin, K., Tyakht, A., Larin, A., Kostryukova, E., Kolchenko, S., Bitner, V., et al. (2017). Abundance profiling of specific gene groups using precomputed gut metagenomes yields novel biological hypotheses. *PLoS One* 12:e0176154. doi: 10.1371/journal.pone.0176154

Zaneveld, J. R., McMinds, R., and Thurber, R. V. (2017). Stress and stability: applying the Anna Karenina principle to animal microbiomes. *Nature Microbiol.* 2:17121. doi: 10.1038/nmicrobiol.2017.121

Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., et al. (2018). dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* 46, W95–W101. doi: 10.1093/nar/gky418

Zhou, Y., Xu, Z. Z., He, Y., Yang, Y., Liu, L., Lin, Q., et al. (2018). Gut microbiota offers universal biomarkers across ethnicity in inflammatory bowel disease diagnosis and infliximab response prediction. *mSystems* 3. doi: 10.1128/mSystems.00188-17

Zhou, Y.-H., and Gallins, P. (2019). A review and tutorial of machine learning methods for microbiome host trait prediction. *Front. Genet.* 10:579. doi: 10.3389/fgene.2019.00579

Conflict of Interest: PN was employed by the company PhenoBiome Inc.; AT, NK, DE, and TS were employed by the company Atlas Biomed Group—Knomics LLC. DR and PN are co-founders of PhenoBiome Inc.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.