frontiers
in Microbiology

ORIGINAL RESEARCH
published: 25 May 2021
doi: 10.3389/fmicb.2021.653314

Binary Metabolic Phenotypes and Phenotype Diversity Metrics for the Functional Characterization of Microbial Communities

Stanislav N. Iablokov[1]*, *Pavel S. Novichkov*[2], *Andrei L. Osterman*[3] *and Dmitry A. Rodionov*[1,3]*

[1] *A.A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia,* [2] *PhenoBiome Inc., Walnut Creek, CA, United States,* [3] *Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA, United States*

The profiling of 16S rRNA revolutionized the exploration of microbiomes, allowing to describe community composition by enumerating relevant taxa and their abundances. However, taxonomic profiles alone lack interpretability in terms of bacterial metabolism, and their translation into functional characteristics of microbiomes is a challenging task. This bottom-up approach minimally requires a reference collection of major metabolic traits deduced from the complete genomes of individual organisms, an accurate method of projecting these traits from a reference collection to the analyzed amplicon sequence variants (ASVs), and, ultimately, an approach to a microbiome-wide aggregation of predicted individual traits into physiologically relevant cumulative metrics to characterize and compare multiple microbiome samples. In this study, we extended a previously introduced computational approach for the functional profiling of complex microbial communities, which is based on the concept of *binary metabolic phenotypes* encoding the presence ("1") or absence ("0") of various measurable physiological properties in individual organisms that are termed phenotype carriers or non-carriers, respectively. Derived from complete genomes via metabolic reconstruction, binary phenotypes provide a foundation for the prediction of functional traits for each ASV identified in a microbiome sample. Here, we introduced three distinct mapping schemes for a microbiome-wide phenotype prediction and assessed their accuracy on the 16S datasets of mock bacterial communities representing human gut microbiome (HGM) as well as on two large HGM datasets, the American Gut Project and the UK twins study. The 16S sequence-based scheme yielded a more accurate phenotype predictions, while the taxonomy-based schemes demonstrated a reasonable performance to warrant their application for other types of input data (e.g., from shotgun metagenomics or qPCR). In addition to the abundance-weighted Community Phenotype Indices (CPIs) reflecting the fractional representation of various phenotype carriers in microbiome samples, we employ metrics capturing the diversity of phenotype carriers, Phenotype Alpha Diversity (PAD) and Phenotype Beta Diversity (PBD). In combination with CPI, PAD allows to classify the robustness of metabolic phenotypes by their anticipated stability in

OPEN ACCESS

Edited by:
Alla Lapidus,
Saint Petersburg State University,
Russia

Reviewed by:
Vladimir Ulyantsev,
ITMO University, Russia
Dmitry A. Ravcheev,
National University of Ireland Galway,
Ireland

***Correspondence:**
Stanislav N. Iablokov
biology@iablokov.ru
Dmitry A. Rodionov
rodionov@sbpdiscovery.org

Specialty section:
This article was submitted to
Evolutionary and Genomic
Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 14 January 2021
Accepted: 06 April 2021
Published: 25 May 2021

Citation:
Iablokov SN, Novichkov PS,
Osterman AL and Rodionov DA
(2021) Binary Metabolic Phenotypes
and Phenotype Diversity Metrics
for the Functional Characterization
of Microbial Communities.
Front. Microbiol. 12:653314.
doi: 10.3389/fmicb.2021.653314

the face of potential environmental perturbations. PBD provides a promising approach for detecting the metabolic features potentially contributing to disease-associated metabolic traits as illustrated by a comparative analysis of HGM samples from healthy and Crohn's disease cohorts.

Keywords: predictive functional profiling, metagenomic, 16S rRNA sequencing, metabolic phenotypes, microbiome, phenotype diversity

INTRODUCTION

Microbial communities are known to colonize various habitats, such as water, soil, and higher organisms including humans. Understanding various types of interactions within microbial communities (or microbiomes) and with an environmental niche is of obvious fundamental and practical importance, especially in the field of human health and disease. Thus, the problems of development and homeostasis of the human gut microbiome (HGM) attracted the most attention from many research groups due to many established associations between dysbiosis and various pathological conditions such as obesity (Cani and Van Hul, 2020), diabetes (Gurung et al., 2020), cancer (Guven et al., 2020), inflammatory bowel diseases (IBD) (Caruso et al., 2020), and neurological disorders (Grochowska et al., 2019). Additional tentative associations of HGM and human health include the regulation of blood pressure (Marques et al., 2018), neurodevelopment (Borre et al., 2014), bile acid metabolism (Ramirez-Perez et al., 2017), and immune homeostasis (Wang et al., 2019). The knowledge of HGM taxonomic composition and functional profiles would enable the identification of relevant features associated with such conditions, opening new diagnostic and therapeutic opportunities.

Rapid advancement of genomic technology enabled a comprehensive coverage of HGM by thousands of completely sequenced reference genomes (O'Leary et al., 2016; Forster et al., 2019; Poyet et al., 2019; Zou et al., 2019). Likewise, massive amounts of fecal samples from a variety of clinical and population studies have been taxonomically profiled by amplicon (16S rDNA) sequencing methodology. A comparative and correlative analysis of these taxonomic profiles versus various types of clinical and other metadata provides new powerful approaches toward the diagnostic and rational selection of probiotics. However, taxonomic profiles alone, no matter how useful for certain practical tasks, lack interpretability in terms of functional, most importantly metabolic, properties and interactions of microbial communities. A genome-based reconstruction of bacterial metabolic pathways and networks (Rodionov et al., 2010; Ravcheev et al., 2013; Khoroshkin et al., 2016; Romine et al., 2017) is a well-established methodology enabling a predictive metabolic modeling (Zhang et al., 2009; Pan and Reed, 2018). However, the extension of this methodology toward complex and widely variable microbial communities, despite some first encouraging steps (Zhang and Reed, 2014; Magnusdottir et al., 2017), represents a substantial challenge. To establish a reliable quantitative description (let alone mathematical model) of the metabolic potential of a microbial community from the 16S amplicon sequencing data using a bottom-up approach (from genome-wide to microbiome-wide metabolic reconstruction), we need to successfully address at least three critical issues.

First, a scalable functional profiling approach should adopt a standard language, i.e., a set of functional traits that can be confidently deduced from the genomes of individual species reflecting their measurable physiological, biochemical, or other properties. These traits (or rather presence and absence thereof) assigned to each genome in a representative collection of reference genomes would provide a foundation for converting the 16S rRNA gene profiles to functional profiles of microbial communities. The commonly used state-of-the-art tools, e.g., PICRUSt2 (Douglas et al., 2020) and Tax4Fun2 (Wemheuer et al., 2020), report the abundance of either gene families or automatically assigned metabolic pathways, deriving reference data from existing genomic databases of biochemical pathways such as the KEGG (Kanehisa et al., 2012). Previously, we introduced a different approach to the predictive metabolic profiling of microbial communities based on the concept of *binary metabolic phenotypes* (Rodionov et al., 2019), which are deduced from the complete genomes of reference bacteria using a subsystems-based metabolic reconstruction (Overbeek et al., 2005). Binary phenotypes represent measurable physiological properties (traits) of individual species such as the ability or inability to produce or consume certain metabolite or nutrient and encoded as "1" for a particular phenotype's carriers and as "0" for non-carriers. The main distinctive aspect of this approach [originally introduced and described for the example of predicted B-vitamin prototrophy and auxotrophy (Rodionov et al., 2019)] is in aggregating data on reconstructed and curated metabolic pathways into a single binary (1 or 0) encoding of a particular functional trait, which has a straightforward biological interpretation.

A second challenge of genomic reference-based functional profiling is the limited coverage of the HGM species by complete reference genomes (amenable to metabolic reconstruction) exacerbated by an even more limited precision of 16S profiling. Indeed, even in the most studied environmental niches such as HGM, taxonomic profiling of detected amplicon sequence variants (ASVs) provides only a partial strain-level resolution, with many ASVs having taxonomic descriptions at the species, genus, or even family level. It leads to a bioinformatic problem of accurate projection of the current knowledge on the presence/absence of functional features from reference genomes of individual strains to real-life ASVs (ensembles of taxonomically unresolved species and strains). Due to the intrinsic variations of many features within microheterogeneous ASVs, for computational purposes, such projection should be considered probabilistic and corresponding maps of ASVs to reference genomes should be accompanied by an estimated error reflecting the uncertainty due to these variations.

In previous studies, we have used a simple taxonomy-based mapping approach, which projects ASVs to the reference database of genomes with their respective binary phenotypes (termed *Binary Phenotype Matrix* or BPM) using a level-by-level comparison of the ASV's taxonomic description with the taxonomies of the reference genomes. BPM entries with the best match were then taken to form the ASV-to-BPM map, with weights equally distributed across unique species. Furthermore, for each binary phenotype and ASV, a Phenotype Index (PI, a value between 0 and 1) was calculated as the map-weighted average binary phenotype of reference organisms across the ASV-to-BPM map. Each PI represents a probability for a given binary phenotype to be associated with the ASV and is accompanied by the corresponding prediction error due to an imprecise mapping. To improve the precision (minimize the uncertainty) of PI assignment to ASVs, we have introduced two additional mapping methods, the first of which relies on the so-called Multi-Taxonomic Assignment (MTA) scheme. In contrast to the commonly used Consensus-Based Taxonomic Assignment (CBTA) scheme [e.g., in consensus-blast plugin in QIIME2 (Bolyen et al., 2019)], where ambiguities are resolved by keeping taxonomic descriptions with sufficient consensus, MTA preserves multiple best-match species-level taxonomies in order to enhance taxonomic resolution. Alternatively, the second, sequence-based approach (SEQ) does not rely on intermediate taxonomies and aligns ASVs directly to the 16S reference database, thus, allowing to obtain a strain-level resolution for some ASVs. Here, we report the benchmarking of our phenotype prediction approach using all the above mapping schemes. For a dataset of 1,000 mock bacterial communities representing HGM with defined taxonomic composition and functional profiles, SEQ scheme demonstrated overall insignificant prediction uncertainties and, as anticipated, largely outperformed the CBTA and MTA schemes. The taxonomy-based schemes, however, showed a reasonable prediction accuracy for a subset of phylogenetically homogeneous phenotypes, thus justifying their use in the phenotypes-based analysis of taxonomic profiles in the absence of the 16S sequencing data (e.g., originating from whole genome shotgun sequencing).

Finally, a third critical aspect of functional bottom-up profiling methodology is the robust computational method of microbiome-wide aggregation of the functional traits of individual species into the community's cumulative metrics, which should have an explicit biological interpretation. Such method would enable a computational comparative analysis of multiple HGM samples to support applications in diagnostics and rational development of dietary supplements for the prevention or correction of dysbiosis-related syndromes. A potential practical utility of such analysis based on binary phenotype encoding of metabolic properties can be illustrated by studies of defined microbial consortia in gnotobiotic mice model aimed at developing therapeutic food supplements for infants with dysbiosis triggered by malnutrition (Blanton et al., 2016; Gehrig et al., 2019; Raman et al., 2019).

Recently, we have introduced the computational approach to a microbiome-wide aggregation of metabolic properties assigned to ASVs (as outlined above) by calculating the Community Phenotype Indices (CPIs), i.e., abundance-weighted PIs. For a given phenotype, the CPI value corresponds to a fractional representation of the phenotype carriers in a microbiome sample. This simple metric was applied to the analysis of B-vitamin biosynthetic potential over large collections of 16S-profiled HGM samples from the Human Microbiome Project and American Gut Project studies (Human Microbiome Project Consortium, 2012; McDonald et al., 2018) and allowed us to detect a significant abundance of B-vitamins auxotrophs, in accordance with the micronutrient sharing hypothesis (Rodionov et al., 2019). This hypothesis was further supported by the studies in humanized gnotobiotic mice model and via anaerobic *in vitro* culturing in the context of extreme variations of B-vitamin supply (Sharma et al., 2019). The CPI-based functional profiling of HGM samples was applied to several other 16S rRNA metagenomic datasets, including the *in vitro* fermentation of fecal microbiomes (Peterson et al., 2019; Elmen et al., 2020) and the comparative analysis of metabolic properties in microbiomes of infants as a function of breast-feeding vs. formula (Jones et al., 2020), allowing us to link the metabolic phenotypes with variable environmental/growth conditions. Finally, predicted metabolic phenotypes were used for the classification of HGM samples from healthy versus IBD patients providing interpretable insights into the host-microbiome mechanisms of disease (Iablokov et al., 2021).

Here, we extended this bioinformatics approach for the metabolic phenotype profiling of HGM samples by incorporating novel metrics for a diversity-based description of the phenotype carriers, namely, Phenotype Alpha Diversity (PAD) and Phenotype Beta Diversity (PBD). These metrics were applied for the metabolic phenotype profiling of several large metagenomic datasets, with PAD serving as a measure of stability for a given functional trait (phenotype) and PBD being a promising method for identification of driving phenotypes.

MATERIALS AND METHODS

Raw Data Analysis

Raw 16S rRNA gene sequencing data from two large metagenomic studies representing the general population, namely, from the American Gut Project (AGP) (McDonald et al., 2018) and the UK Twins study (UKT) (Goodrich et al., 2016), as well as from three IBD-related studies conducted in China (CHN) (Zhou et al., 2018), Spain (ESP) (Pascal et al., 2017), and Netherlands (NLD) for the IBD group (Imhann et al., 2018) and for healthy controls (Tigchelaar et al., 2015), were analyzed using the QIIME2's dada2 plugin (Bolyen et al., 2019). 16S amplicons were quality-filtered and dereplicated into amplicon sequence variants (ASVs) with default parameters. Samples with reads counts below a certain threshold were discarded. For the AGP dataset, we additionally removed samples with high levels of blooms. Summary for the analyzed datasets, with read count thresholds and the remained number of samples, is presented in **Table 1**.

In silico Mock Communities

A set of 1,000 mock communities was randomly generated (*in silico*) from the reference collection of 2,662 bacterial HGM genomes (Rodionov et al., 2019; Iablokov et al., 2021) that are available in the SEED database (Overbeek et al., 2014). The number of unique species (S) for each community was sampled from the normal distribution (mean = 30, std = 5) with the restriction 10 < S < 60. Unique species names were then sampled from the list of 120 most abundant species in the UKT dataset, with weights equal to their mean abundance (A) across the dataset. For each unique species-level description, N organisms (strains) with the corresponding taxonomy were uniformly sampled from the reference HGM genome database. The number N was uniformly sampled between 1 and the ceiling value of the square root of the total number of organisms with the given species-level description. The respective relative abundance (R) of each species in a given mock community was sampled from a normal distribution (mean = A, std = A × 0.4) with the restriction A × 0.1 < R < A × 3. Values of R for each community were then rescaled to sum up to 1. The respective relative abundance fractions (Q) of strains within a species were uniformly sampled and rescaled to sum up to 1. Total amplicon count (T) for each community was sampled from a normal distribution (mean = 20,000; std = 4,000) with the restriction 4,000 < T < 40,000. Individual 16S amplicon counts (C) for the organisms (strains) in each community were then obtained as C = Q × R × T. These generated bacterial communities comprised the MOCK TRUE dataset with a confidently known one-to-one association (map) between each 16S sequence and the reference genome database. To model a real experiment, the corresponding 16S rRNA gene sequences were further truncated to the V3–V4 variable regions (flanked by 341F/806R primers). Truncated amplicons with identical sequences were collapsed into a single amplicon sequence variant (ASV) with aggregated abundance, comprising the MOCK AGGR dataset. The resulting ASV abundance tables and respective 16S sequences (for both the TRUE and AGGR mock datasets) are presented in **Supplementary Table 1**.

TABLE 1 | Summary on the number of samples retained in each analyzed dataset after filtration by the minimum number of reads threshold.

Dataset	Number of samples	Min. number of reads
AGP	2,868	10,000
UKT	3,288	10,000
CHN	134 HC/75 CD	4,000
ESP	154 HC/140 CD	15,000
NLD	966 HC/163 CD	15,000

For the IBD-related datasets (CHN, ESP, and NLD), the respective numbers are given for healthy (HC) and Crohn's disease (CD) cohorts separately.

Taxonomic Profiling and Abundance Renormalization

Taxonomic profiling of ASV representative sequences was performed following the Multi-Taxonomic Assignment (MTA) scheme. Specifically, 16S amplicons were aligned using NCBI BLAST+ (Camacho et al., 2009) against a joined reference 16S rRNA database with sequences from the RDP database version 11.5 (Cole et al., 2014) and NCBI 16S database version of December 2019. Alignment results were sorted according to the fraction (from 0 to 1) of their identity F, with the maximum *F* value for the alignment denoted as M. Top alignment hits with value of F in the range from M to M-(1-M)/S and a threshold greater than the 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 limits the alignment quality from below, and was taken equal to 0.85. The strict choice of value for S is motivated by the necessity to investigate 16S sequences (and their corresponding organisms) in a small neighborhood of the top match, while keeping the MTA description compact. The value of D was chosen to discard the ASVs with a poor taxonomic resolution, for which metabolic phenotype predictions are highly inaccurate due to the high degree of phenotype heterogeneity within a broader phylogenetic group. The resulting multi-taxonomy for each ASV was a list of unique species-level taxonomic descriptions with equal weights assigned to each item. String representations for MTA consisted of slash-separated names of taxa on each taxonomic level, e.g., *Bacteroides ovatus/vulgatus* or *Escherichia coli/Salmonella enterica*.

Based on the MTA profile, we additionally performed a Consensus-Based Taxonomic Assignment (CBTA) by choosing 51%-consensus on each taxonomic level, leaving blank the assignment entries for taxonomic levels with an insufficient consensus.

Original dada2-derived 16S amplicon counts were further renormalized to account for the different 16S rRNA gene copy numbers (GCN) in microbial genomes. Average GCN values for taxonomic entities at different ranks were extracted from the rrndb-5.6 database (Stoddard et al., 2015) and mapped to ASVs using their MTA-based taxonomic descriptions. For each ASV, a simple mean GCN value was calculated and used as a factor to normalize the ASV's abundance. Thus, obtained abundances were then re-scaled to sum up to 1.

Binary Phenotypes for Reference Genomes

The comparative genome analysis and reconstruction of target metabolic pathways in the reference genome database containing 2,662 HGM genomes (representing ∼770 individual species) was previously conducted using the subsystems approach (Overbeek et al., 2005; Osterman et al., 2010) implemented in the SEED/RAST platform (Overbeek et al., 2014) as previously described (Rodionov et al., 2019; Iablokov et al., 2021). Briefly, the metabolic subsystems were manually built as groups of functional roles for enzymes, transporters, and transcriptional regulators that are involved in a specific aspect of the cellular machinery such as a metabolic pathway. Functional gene assignments and metabolic reconstructions were performed using the following three genome context techniques to functionally link a set of genes to a single pathway: (i) clustering of genes on the chromosome (operons), (ii) co-regulation of genes by a common regulator [regulons, as captured in the RegPrecise database (Novichkov et al., 2013)], and (iii) co-occurrence of genes in a set of related genomes (Overbeek et al., 2007). For the functional gene annotation and building metabolic subsystem in SEED, we combined the existing annotations with information from literature accessed via the PaperBLAST tool (Price and Arkin, 2017) and reference databases including the UniProtKB/Swiss-Prot for characterized proteins (Boutet et al., 2016), KEGG for reference metabolic pathways (Kanehisa et al., 2012), TCDB for transporter classification (Elbourne et al., 2017), and CAZy for classification of Glycosyl Hydrolases (Lombard et al., 2014).

Curated across HGM genomes metabolic subsystems include the central biochemical pathways classified into four categories: (i) biosynthesis of vitamins and cofactors, (ii) biosynthesis of protein-forming amino acids, (iii) utilization of carbohydrates and other carbon sources, and (iv) production of fermentation products including short-chain fatty acids (SCFAs) such as acetate, butyrate, and propionate. In many subsystems, we captured distinct biochemical pathway variants and numerous non-orthologous enzymes and transporters. Using pathway-specific logical rules that account for both the variable pathway and signature genes, we assigned binary phenotypes (1/0) for each metabolic phenotype and target genome. The resulting Binary Phenotype Matrix (BPM) contained 94 metabolic phenotypes reflecting the presence/absence of a complete catabolic or biosynthetic pathway. The following 24 representative phenotypes were a subject of analysis in this work (**Figure 2**): biosynthesis of B-vitamins (B1, B2, B3, B5, B6, B7, B9, B12), lipoate, and vitamin K; production of SCFAs (butyrate and propionate); utilization of carbohydrates (glucose, galactose, fructose, mannose, xylose, arabinose, fucose, rhamnose, ribose, lactose); and biosynthesis of amino acids (His, Trp). These phenotypes were chosen by a combination of criteria such as physiological relevance, extensive knowledge ensuring confidence of phenotype inference, and significant and variable representation across microbiome samples.

ASV Mapping

To obtain the phenotype profiles for the analyzed 16S rRNA samples, we utilized a development version of the Phenotype Profiler tool provided by PhenoBiome Inc. (Walnut Creek, CA, United States[1]). Mapping of ASVs to the BPM was performed using three different schemes. Two of them are taxonomy-based (CBTA and MTA), and they use a level-by-level comparison of the ASV taxonomic descriptions with taxonomies of the reference genomes. Matches on the deepest taxonomic level were added to the ASV-to-BPM map with weights equally distributed (and summing up to 1) across unique species. Within each unique species, the weights were equally distributed between all strains. ASVs with a match on at least the family level were considered as mapped, with the rest being non-mapped. For multi-taxonomies, this procedure was applied for each simple taxonomic description in the MTA with additional weighting using the MTA weights. The third scheme is sequence-based (SEQ) and employs an ASV sequence alignment against the 16S database for reference genomes (from which phenotypes were derived). This process mirrors the MTA scheme with the same values for S and D parameters. ASVs with F values greater than D were considered as mapped. ASVs with all reference sequences having their *F* value below D were considered as non-mapped. The average values of ASV coverage by reference genomes, i.e., the abundance of mapped ASVs, for all analyzed datasets is presented in **Table 2**.

Assignment of Phenotype Indices

For each ASV and metabolic phenotype, we assigned a corresponding Phenotype Index (PI, from 0 to 1), representing the probability for the phenotype to be associated with the given ASV. PIs were calculated using the ASV-to-BPM map as the map-weighted averages of binary phenotypes from the BPM: $PI = \sum W_i \times P_i$, where W_i and P_i are the mapping weight (from 0 to 1) and respective binary phenotype value (0 or 1) for the ith organism in the ASV-to-BPM map. Assuming that the binary phenotypes follow the binomial distribution, we calculated variance of $Var\,(PI) = PI \times (1 - PI)$, and have taken it as an estimate of PI prediction uncertainty. To measure the cumulative properties of microbial communities with respect to a given phenotype, we computed the Community Phenotype Indices CPIs as abundance-weighted average PIs for all phenotypes, $CPI = \sum A_i \times PI_i$, where A_i and PI_i are the relative abundance (from 0 to 1) and respective Phenotype Index (from 0 to 1) of the ith ASV in the sample. Under the assumption of independent co-occurrence of ASVs, we calculated variance of CPI, $Var(CPI) = \sum A_i^2 \times Var\,(PI_i)$, and have taken it as an estimate of CPI prediction error. The relative CPI prediction uncertainty (rSTD) was then calculated as the ratio of its standard deviation (square root of variance) and CPI itself. The computed CPI values for the *in silico* mock, AGP, and UKT datasets, as well as for the three IBD-related studies are provided in **Supplementary Tables 2A,B, 7**, respectively.

Phenotype Diversity

The Phenotype Alpha Diversity (PAD) and Phenotype Beta Diversity (PBD) were estimated as, respectively, the alpha and beta diversity of the sub-communities of carriers of a particular phenotype. Briefly, a multiple alignment of ASV

TABLE 2 | Mapping coverage for the analyzed datasets.

Scheme	AGP	UKT	CHN	ESP	NLD
CBTA	81%	89%	–	–	–
MTA	84%	93%	–	–	–
SEQ	79%	86%	99%	97%	94%

For each dataset (and mapping scheme where appropriate), the average abundance of ASVs mapped to the BPM is shown as percentage.

[1] www.phenobiome.com

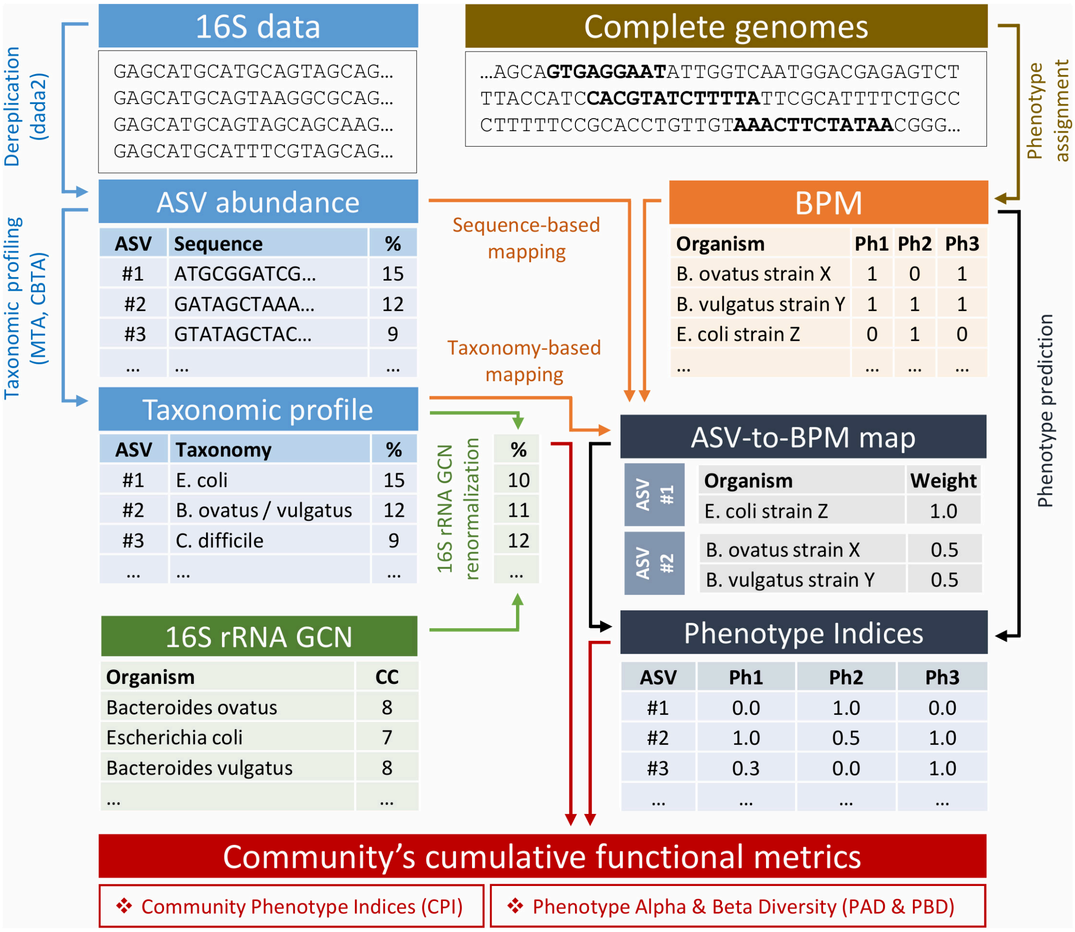

FIGURE 1 | Flowchart of the phenotype profiler pipeline.

representative sequences was performed using MUSCLE (Edgar, 2004), followed by the construction of an unrooted phylogenetic tree with FastTree 2 (Price et al., 2010). The resulting tree was rooted according to the midpoint strategy and used (with ASV abundances and ASV-to-BPM map) in calculation of both PAD and PBD metrics with the Python's scikit-bio package[2]. Sub-communities of the respective phenotype carriers were determined according to a PI > 0.6 threshold, with the abundances of selected ASVs normalized to 1 for each sample. Faith Phylogenetic Diversity (Faith, 1992) and Weighted UniFrac (Lozupone and Knight, 2005) were chosen as the respective alpha and beta diversity metrics. The computed PAD values for AGP and UKT datasets are provided in **Supplementary Table 4**.

RESULTS AND DISCUSSION

A computational pipeline for metabolic profiling of complex microbial communities represented by the 16S amplicon sequencing data is based on the probabilistic prediction of metabolic phenotypes for each ASV identified in the microbiome sample (see Methods). Thus, for each phenotype, an ASV is assigned a respective Phenotype Index (PI, from 0 to 1), i.e., a probability for the phenotype to be associated with a given ASV. The principal computational scheme of our pipeline is shown in **Figure 1**. It includes the following steps: (i) filtering of 16S rRNA amplicons and their dereplication into ASVs; (ii) taxonomic profiling; (iii) abundance renormalization due to variations in 16S rRNA gene copy numbers (GCN) among different taxa; (iv) mapping of ASVs to the reference genomes; (v) phenotype prediction; and (vi) calculation of community's cumulative characteristics, including the Community Phenotype Index (CPI), Phenotype Alpha Diversity (PAD), and Phenotype Beta Diversity (PBD). For characteristics requiring phylogenetic data (PAD and PBD), the additional procedures of ASV multiple alignment and tree construction are implemented (not shown). In the following section, we assess the phenotype prediction method using the *in silico* generated mock communities with precisely known taxonomic composition and functional profiles and compare different approaches for the mapping of ASVs to the collection of reference genomes. Next, we discuss the potential application of the recently introduced Phenotype Alpha Diversity

[2] http://scikit-bio.org

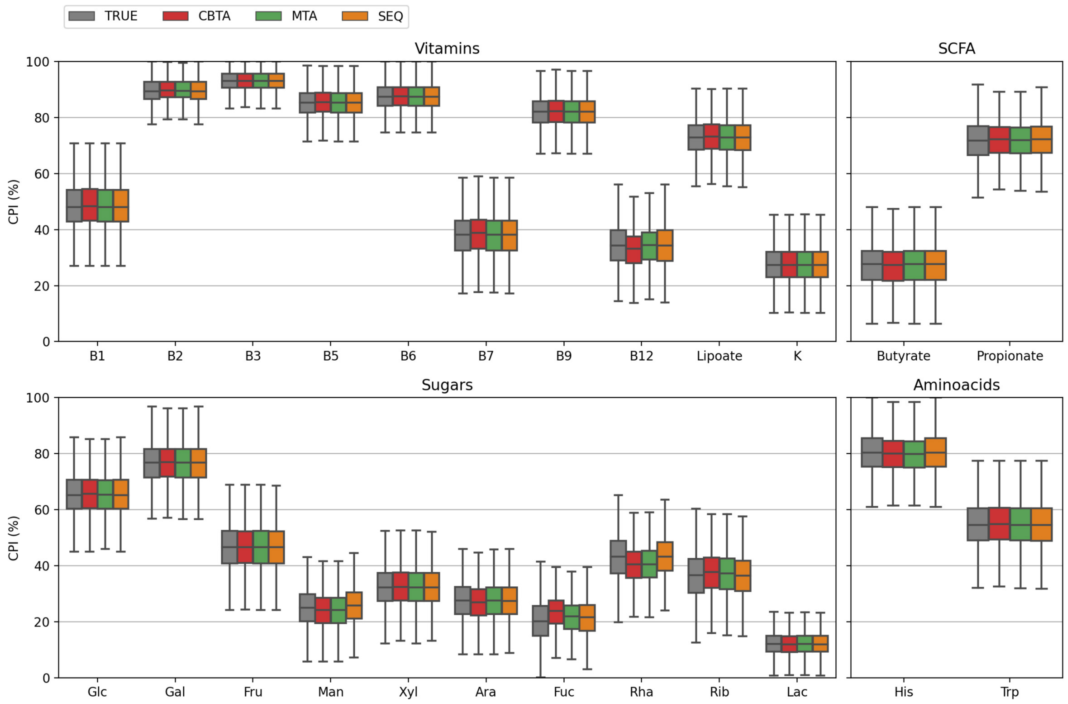

FIGURE 2 | CPI value distributions for 1,000 *in silico* generated mock communities. Phenotype Indices (PI) are predicted according to four ASV-to-BPM mapping schemes: (i) TRUE, with true Pis; (ii) CBTA and MTA, with PIs predicted using taxonomy-based mapping schemes; (iii) SEQ, with PIs predicted using a sequence-based mapping scheme.

(PAD) metric as a measure of phenotype stability with respect to environmental change. Finally, we propose a similar concept of Phenotype Beta Diversity (PBD), which provides a diversity-based approach for the detection of metabolic features which are presumably associated with clinical status.

ASV Mapping and Phenotype Prediction

Throughout previous studies, we used three different schemes for the prediction of phenotypes, one scheme being an upgrade of another. Two of them employ taxonomic assignments to map ASVs to reference genomes (ASV-to-BPM map), with the level-by-level comparison of respective taxonomic strings (see Methods). The naivest scheme makes use of the Consensus-Based Taxonomic Assignment (CBTA) approach, which resolves ambiguities by keeping the descriptions only for the taxonomic levels with a sufficient consensus. A somewhat advanced scheme is based on the Multi-Taxonomic Assignment (MTA) approach, which retains all relevant simple taxonomic descriptions equally weighted, with string representations for MTA consisting of slash-separated taxonomies, e.g., *Bacteroides ovatus/vulgatus* or *Escherichia coli/Salmonella enterica*. Lastly, the third scheme utilizes the sequence-based (SEQ) mapping approach, which relies on the alignment of ASVs directly against the 16S rRNA sequences of the reference organisms.

In this study, we assessed the performance of our computational approach by analyzing the accuracy of phenotype prediction for 1,000 *in silico* generated mock bacterial communities representing HGM with defined taxonomic composition and functional profiles. We calculated the Community Phenotype Indices, i.e., abundance-weighted PIs, and relative CPI prediction uncertainties (see Methods) for the three proposed ASV mapping schemes (CBTA, MTA, and SEQ) as well as for the true mapping scheme (TRUE) known upon the generation of mock communities (**Supplementary Table 2A**). The obtained CPI distributions (**Figure 2**) demonstrated a high degree of similarity, therefore, making a dataset-wise description of functional traits independent of the mapping scheme. From a sample-by-sample comparison of the predicted CPIs with their true values (**Figure 3**), it is apparent that the SEQ mapping scheme significantly outperforms (**Supplementary Table 3**) the taxonomy-based schemes (CBTA and MTA). This is somewhat an expected result, because the V3–V4 variable regions of the 16S rRNA gene allows for a partial strain-level taxonomic resolution, thus, decreasing the phenotype prediction error. Among the latter two schemes, MTA showed overall lower values in the relative change of CPI, most likely due to the phenotype averaging across a phylogenetically narrower group of organisms as compared to CBTA.

Despite the generally poorer performance of the taxonomy-based mapping approaches, they nonetheless demonstrated a reasonable phenotype prediction accuracy for the majority of the considered phenotypes, with typical discrepancies from the true values of the order of 5%. This suggests the potential use of the taxonomy-based mapping for functional profiling of metagenomic samples lacking a 16S sequencing data, however, described by taxonomic profiles derived from, e.g., shotgun whole-metagenomic sequencing. This seems especially promising for the profiling of highly conservative phenotypes (such as B-vitamin or amino acid biosynthesis), i.e., functional traits with less variability in taxonomically close microbial genomes. Another important observation is that for some

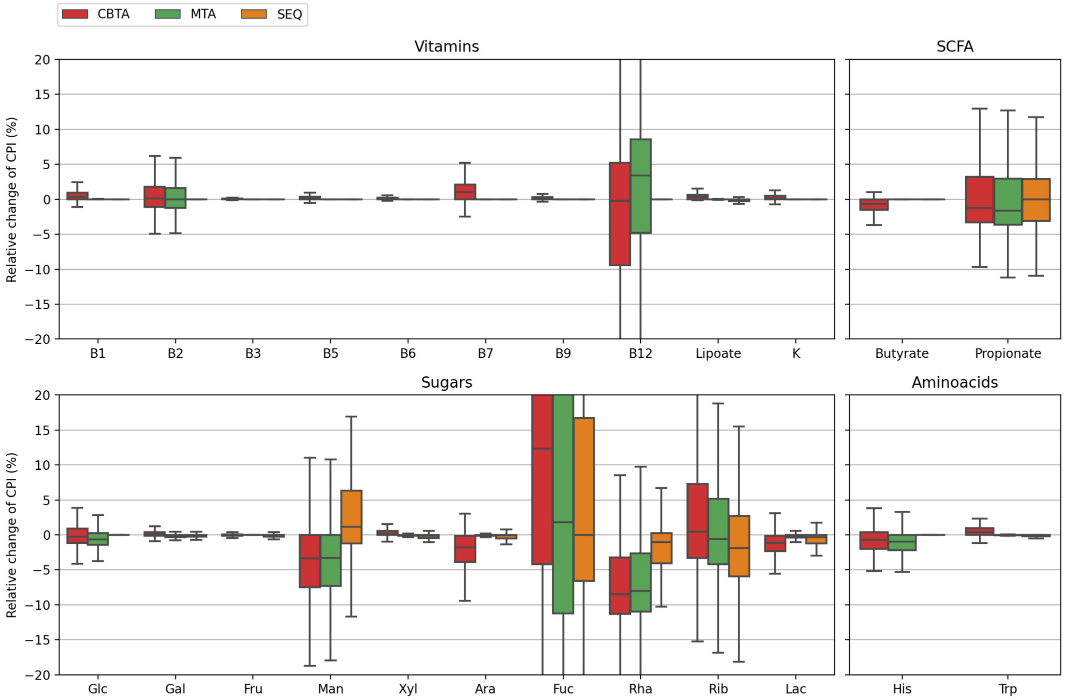

FIGURE 3 | Distributions of the relative change in the CPI values for 1,000 *in silico* generated mock communities. CPI values are calculated using three ASV-to-BPM mapping schemes (CBTA, MTA, SEQ) and compared with true CPIs.

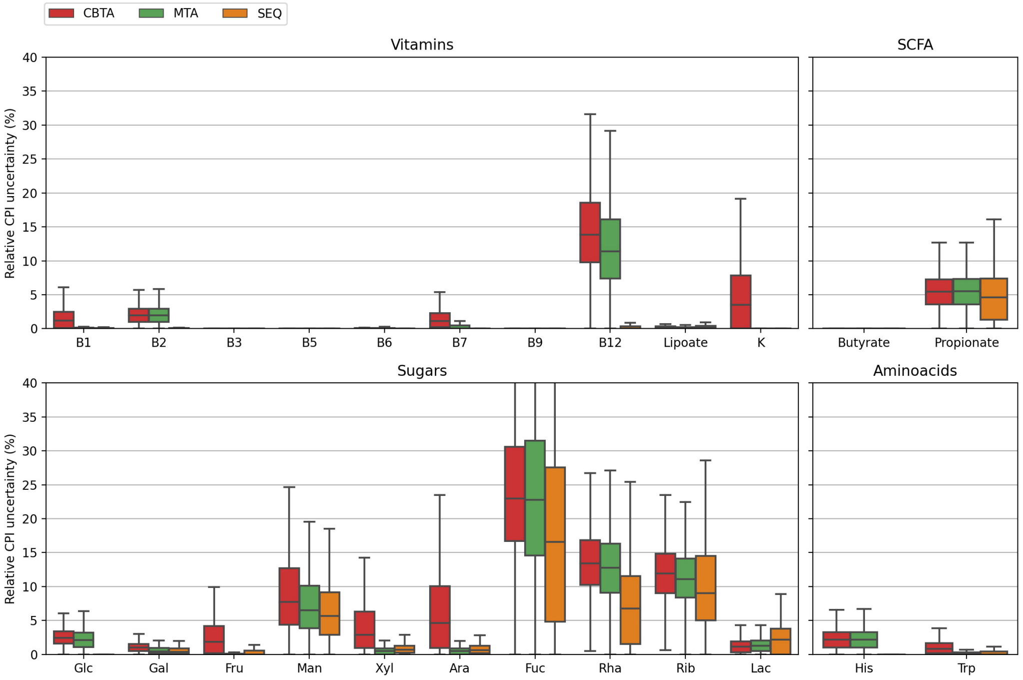

FIGURE 4 | Distributions of the relative CPI prediction uncertainties calculated for 1,000 *in silico* generated mock communities using three mapping schemes.

phenotypes, namely, fucose degradation (Fuc) and ribose degradation (Rib), even the SEQ mapping scheme demonstrated inaccurate phenotype predictions of the order of 10% and larger. This is due to the high degree of their phylogenetic microheterogeneity which is straightforwardly observed when considering the corresponding distributions for the relative CPI prediction uncertainties (**Figure 4**). Unlike the genuine prediction errors (**Figure 3**) estimated with respect to the true CPI values of mock communities, the relative CPI uncertainties were calculated based solely on the ASV mapping schemes (CBTA, MTA, and SEQ). The observed correlation between the true phenotype prediction errors and relative CPI uncertainties

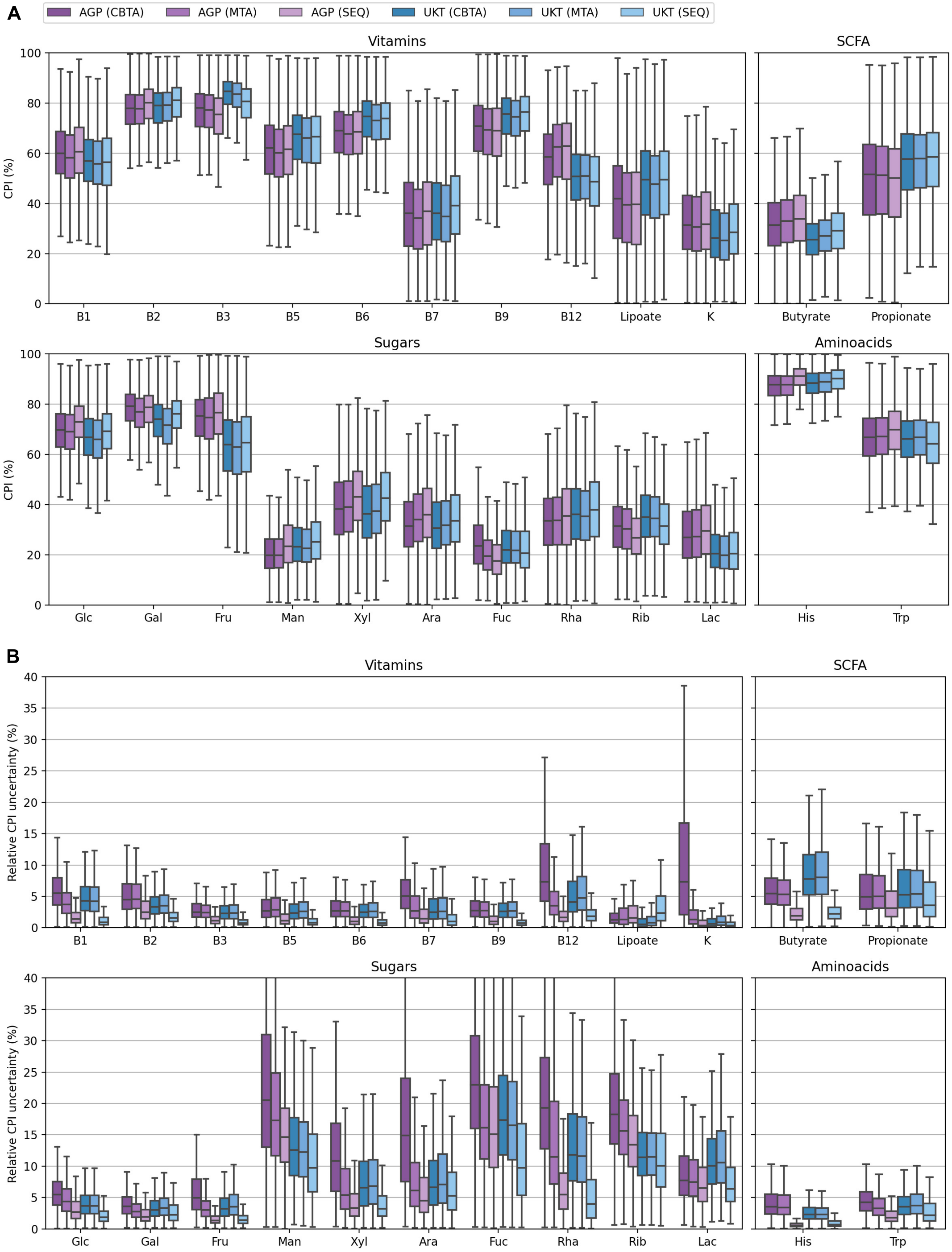

FIGURE 5 | Distributions of CPI values **(A)** and their relative uncertainty values **(B)** for the AGP and UKT datasets calculated using three ASV-to-BPM mapping schemes.

strongly suggests the use of the latter as a measure of phenotype prediction ignorance.

To assess the performance of our computational approach for functional profiling of complex microbial communities on real metagenomic data, we analyzed gut samples from two large HGM datasets, namely the American Gut Project (AGP) and UK Twins study (UKT). We calculated the respective CPI values (**Figure 5A**) and corresponding relative CPI prediction uncertainties (**Figure 5B**) using three ASV-to-BPM mapping schemes, namely, CBTA, MTA, and SEQ (**Supplementary Table 2B**). The resulting CPI distributions showed a striking similarity between these mapping schemes for both datasets, which is in good agreement with our previous discussion. The magnitudes of relative CPI prediction uncertainty for both datasets mirror those observed for the mock communities, with the SEQ mapping scheme demonstrating an overall greater performance. The only minor exception is that of small, however, non-zero, values of the relative CPI prediction uncertainty, that were measured in both datasets even for the phenotypes with a high level of phylogenetic homogeneity. These uncertainties are driven by the presence of taxa, which are poorly covered by reference genomes, thus, leading to the ambiguous ASV-to-BPM maps. Finally, the overall level of CPI prediction uncertainty was usually higher for the AGP samples than for the UKT samples, which is due to the lower length of the sequenced 16S rRNA gene region in the AGP dataset (150 nts) as compared to the UKT study (292 nts), thus, explaining the superior taxonomic resolution in the latter case.

In summary, both taxonomy-based (CBTA and MTA) and sequence-based (SEQ) mapping schemes demonstrated a reasonable phenotype prediction accuracy for the majority of the metabolic phenotypes in both *in silico* generated mock communities and samples from large HGM studies (AGP and UKT), with the SEQ mapping scheme significantly outperforming the taxonomy-based approaches. It is expected that the use of full-length 16S amplicons will further increase the prediction accuracy of SEQ, thus, allowing for the reliable profiling of metabolic phenotypes even with a high degree of phylogenetic microheterogeneity. In further analysis, the use of the SEQ mapping scheme is assumed.

TABLE 3 | Phenotype categories and their observed characteristics.

Category	I	II	III	IV
Phenotypes	B7, Lipoate, K, Xyl, Ara, Rha, Propionate	Man, Rib, Lac, Butyrate	B2, B3, B6, B9, Glc, Gal, Trp	B1, B5, B12, Fru
CPI variation	Moderate to large (35.6–49.5)	Moderate (27.7–33.0)	Moderate to small (21.4–32.5)	Moderate (35.1–40.1)
Median PAD	Small (3.3–6.4)	Small to moderate (5.1–7.6)	Large (9.8–11.4)	Moderate to large (7.9–10.2)

Phenotype Alpha Diversity

In a recent paper (Iablokov et al., 2021), we introduced a concept of Phenotype Alpha Diversity (PAD), which serves to describe the alpha diversity for a sub-community of carriers of a particular phenotype, thus, reflecting how phylogenetically broad or narrow this sub-community is. To further develop this concept, we computed the PAD values for the AGP and UKT datasets (**Supplementary Table 4**) and investigated the CPI-vs.-PAD relationship. The respective scatterplots are shown in **Figure 6** for six selected phenotypes and in **Supplementary Figure 1** for the rest of the phenotypes. Based on a set of descriptive statistics for the CPI and PAD distributions (**Supplementary Table 5**), we clustered 22 phenotypes into four categories according to the shapes of their respective CPI-vs.-PAD clouds, described by median PAD values (PAD Q50), median CPI values (CPI Q50), and 10–90 percentile range (10–90 PR) of CPI

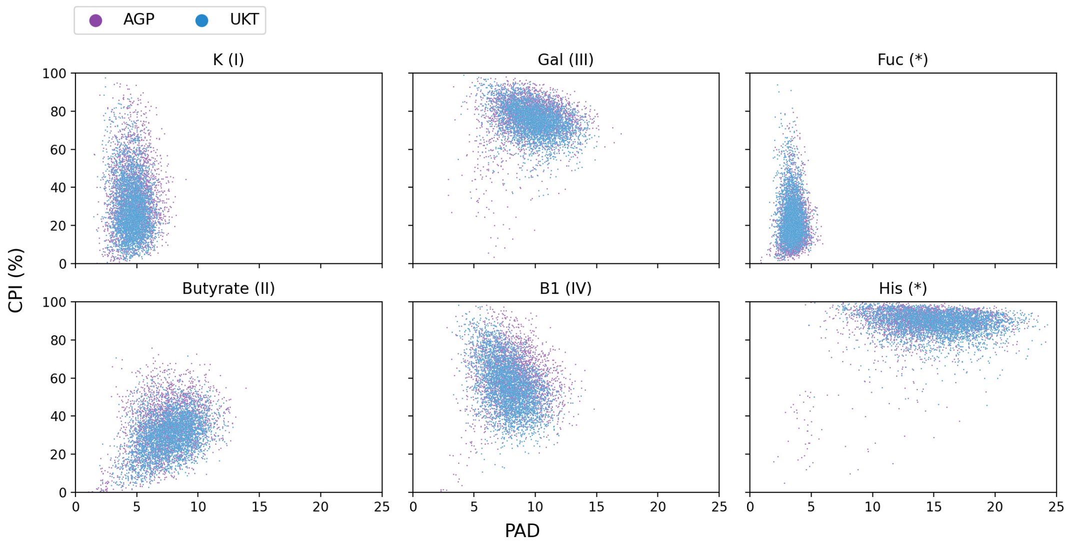

FIGURE 6 | CPI-vs.-PAD scatterplots for six selected metabolic phenotypes calculated for the AGP and UKT datasets.

variation (**Table 3**). The remaining two phenotypes (Fuc and His) remained unclassified since they demonstrated marginal features.

The CPI and PAD metrics allow us to relate the variation of phenotype abundance with its diversity, the latter likely accounting for the phenotype stability in the face of environmental perturbations. For phenotypes with a generally low PAD (e.g., as in categories I and II), one expects moderate to large variations in the respective CPI values due to taxonomic shifts caused by such environmental changes. At the same time, well-diversified phenotypes (such as in category III and His) are expected to demonstrate large median CPI values, leaving little room for a significant CPI variation. This hypothesis is indeed supported by the overall decreasing trend in the CPI 10-90 PR vs. median PAD plot (**Figure 7A**) for the analyzed phenotypes.

We also notice that phenotypes with large PAD values (categories III and IV) have a strong tendency to show large PAD variations (**Figure 7B**) across the analyzed samples. Thus, for these phenotypes, PAD itself can be used as a non-redundant description of microbial communities. However, this is not entirely true for the histidine biosynthesis (His) phenotype. Despite the great variation in its PAD values (10–90 PR = 8), Phenotype Alpha Diversity for the His phenotype becomes a less efficient metabolism-driven description due to its significant correlation with the total alpha diversity. This conclusion should also hold for the majority of other amino acid biosynthesis phenotypes, which are even more abundant than His. Lastly, the fucose utilization (Fuc) phenotype has both a low diversity (PAD Q50 = 3.5) and abundance (CPI Q50 = 19.3). It also demonstrates an exceptionally low CPI variation (10–90 PR = 25.9), thus, not entirely following the "low diversity–high variation" pattern observed for other phenotypes. This is probably due to the fact that fucose is a rare monosaccharide, which is present as a minor constituent in host-derived glycans such as mucin (Tailford et al., 2015) and human milk oligosaccharides (Bode, 2009), and fucose utilization represents a somewhat rare functional capability among HGM bacteria. Similar PAD vs. CPI dependence is expected for other phenotypes describing the utilization of rare carbohydrates (data not shown).

TABLE 4 | Intragroup similarity of carriers of selected three metabolic phenotypes in the HC and CD groups.

Dataset	BD (HC)	BD (CD)	PBD (HC)	PBD (CD)	rPBD (HC)	rPBD (CD)
(A) Butyrate producers						
CHN	0.44	0.39	0.36	0.46	0.92	1.29
ESP	0.47	0.49	0.27	0.44	0.65	0.99
NLD	0.44	0.46	0.26	0.45	0.64	1.00
Mean	**0.45**	**0.44**	**0.30**	**0.45**	**0.73**	**1.10**
(B) Lactose utilizers						
CHN	0.44	0.39	0.34	0.39	0.85	1.10
ESP	0.47	0.49	0.21	0.39	0.51	0.85
NLD	0.44	0.46	0.29	0.39	0.69	0.88
Mean	**0.45**	**0.44**	0.28	0.39	0.68	0.94
(C) Vitamin B12 producers						
CHN	0.44	0.39	0.37	0.41	0.91	1.09
ESP	0.47	0.49	0.28	0.42	0.67	0.87
NLD	0.44	0.46	0.25	0.36	0.62	0.79
Mean	0.45	0.44	0.30	0.40	0.73	0.92

Pairwise mean values for total beta diversity (BD), Phenotype Beta Diversity (PBD), and relative Phenotype Beta Diversity (rPBD) are shown for each IBD dataset for both healthy (HC) and Crohn's disease (CD) groups. Bold font was used to emphasize the last row with mean values. Coloring of cells reflects the corresponding magnitudes in order to facilitate visual perception.

Overall, Phenotype Alpha Diversity provides a measure of diversity for the sub-communities of phenotype carriers, similar to the functional redundancy index (FRI) used in Tax4Fun2 (Wemheuer et al., 2020). Despite the conceptual resemblance with PAD, the latter employs data on gene families [such as from the KEGG database (Kanehisa et al., 2012)], while in the present computational approach, the considered functional traits are the actual biological phenotypes, e.g., a capability for a vitamin biosynthesis or a sugar utilization. It should be noted that the PAD metric has an important application for sample classification tasks. For classifiers with phenotype metrics (such as CPI) used as features, PAD can serve as a filtering criterion, allowing one to discard phenotypes with insufficient diversity of carriers across

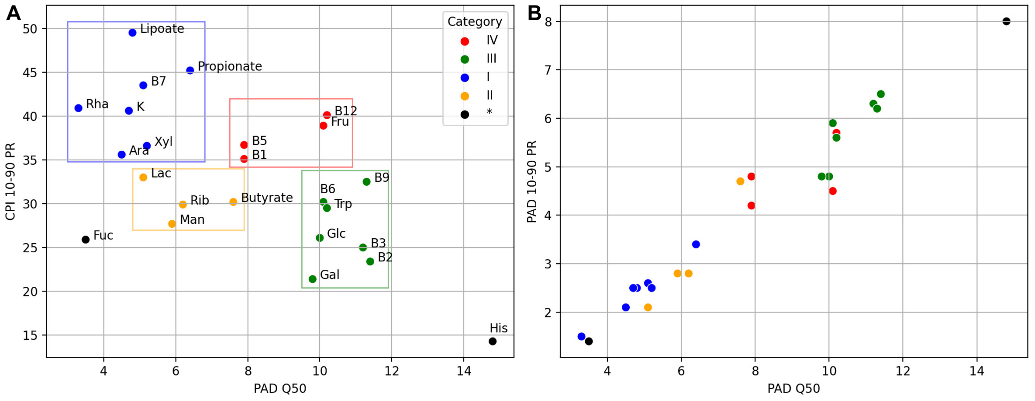

FIGURE 7 | Mutual dependence of the CPI and PAD values obtained for the AGP and UKT datasets. **(A)** Dependence between median PAD (PAD Q50) and variation of CPI (CPI 10–90 percentile range). A total 22 metabolic phenotypes are clustered into four categories, with two phenotypes remained unclassified. **(B)** Dependence between median PAD (PAD Q50) and variation of PAD (PAD 10–90 percentile range).

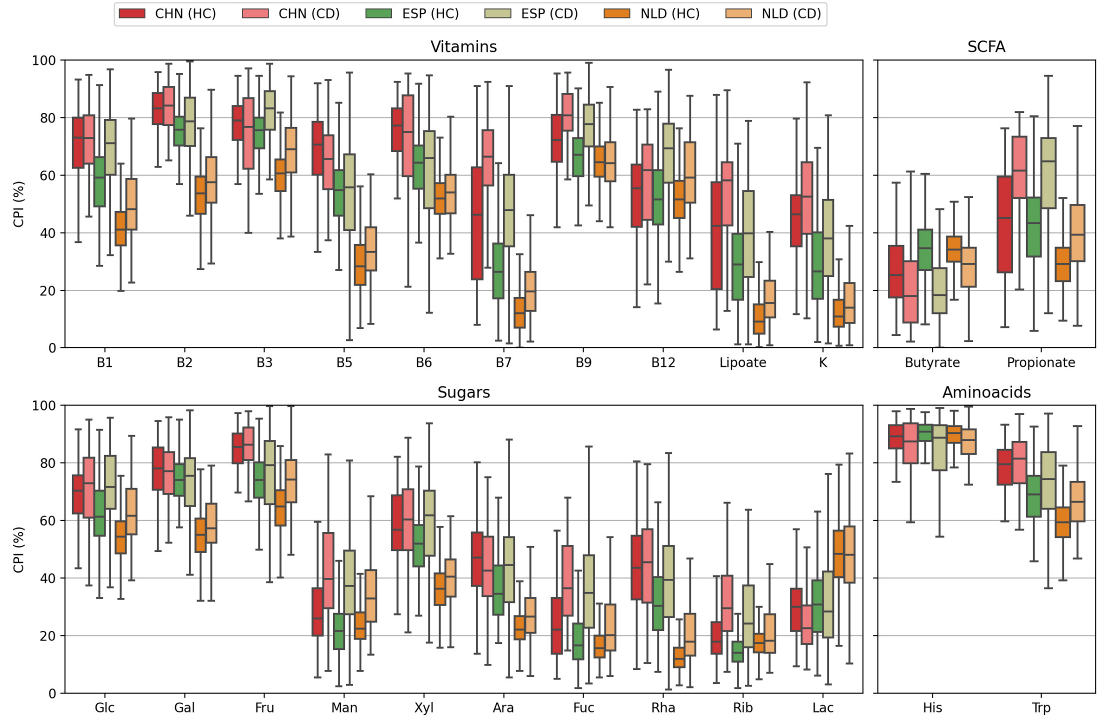

FIGURE 8 | Distributions of the CPI values calculated for samples from heathy (HC) and Crohn's diseases (CD) cohorts in the CHN, ESP, and NLS datasets.

the analyzed samples. We successfully applied this approach to construct classifiers for healthy vs. Crohn's disease subjects with only phylogenetically well-diversified phenotypes (Iablokov et al., 2021). This permitted us to interpret the classification outcome in a truly metabolism-driven manner, i.e., in terms of potentially driving phenotypes.

Phenotype Beta Diversity

In a complete analogy to Phenotype Alpha Diversity, we introduce the concept of Phenotype Beta Diversity (PBD) as the beta diversity (i.e., distances between samples) for the sub-communities of phenotype carriers. To show the potential applications of PBD, we analyzed gut samples in the healthy (HC) and Crohn's disease (CD) groups from three inflammatory bowel disease (IBD) studies conducted in China (CHN), Spain (ESP), and Netherlands (NLD). The respective PBD values for the analyzed phenotypes, as well as the total beta diversity (BD), were calculated using the Weighted UniFrac beta diversity metric. To estimate the intragroup similarity between samples within the HC and CD groups, we computed the respective mean pairwise distances using both BD distance matrix and PBD distance matrices for each phenotype (**Supplementary Table 6**). To account for the inheritance of the diversity scale by the phenotype carriers' sub-community from the total community, we also calculated the relative PBD (rPBD) as the ratio of the PBD value for a given phenotype over the total BD, for each of the HC and CD groups.

The values of rPBD that are close to 1.0 correspond to the same level of dissimilarity between samples for either of the two approaches, BD or PBD. Any deviation from 1.0 serves as an indicator for an increase or decrease in dissimilarity when passing from the total microbial communities to the sub-communities of phenotype carriers. Among all analyzed phenotypes, the butyrate synthesis (Butyrate), lactose (Lac) degradation, and B12 vitamin synthesis (B12) demonstrated a significant drop in both PBD and rPBD values for the HC samples when compared with the CD samples (**Tables 4A–C**). This observation suggests that in healthy subjects, the sub-communities of each of the above phenotype carriers are more similar to one another than the respective sub-communities in Crohn's disease patients, as if following the famous "Anna Karenina" principle for microbiomes. This principle states that gut bacterial communities of healthy people are alike, while disease-associated microbiomes are different in their own way. Notably, here, this principle is valid not for the entire microbial communities but rather for the sub-communities of Butyrate-, Lac-, and B12- phenotype carriers.

Remarkably, the level of PBD-dissimilarity between samples within the HC and CD groups is not explicitly associated with the corresponding differences in CPI (**Figure 8** and **Supplementary Table 7**). For Butyrate, the mean CPI is greater in the HC group for all datasets. The converse is true for B12, while for the Lac phenotype, there is no obvious pattern. Moreover, for other phenotypes with significant CPI differences between the HC and CD groups in all datasets (such as for B7, Lipoate, Propionate, Man, Fuc), almost identical levels of intragroup PBD-dissimilarity are observed (**Supplementary Table 6**). These evidences suggest that Phenotype Beta Diversity acts as another complimentary (to CPI) dataset-wise description of microbial communities. PBD accounts for the phylogenetic data and provides a diversity-based approach for the detection of metabolic features that are presumably associated with clinical status.

CONCLUSION

In this study, we further developed our computational approach for the predictive functional profiling of complex microbial communities, which is based on the concept of binary metabolic phenotypes. Phenotype prediction accuracy was assessed using both (i) the *in silico* generated mock bacterial communities representing HGM with defined taxonomic composition and functional profiles, and (ii) two large metagenomic HGM datasets. The sequence-based scheme for ASV mapping to reference genomes demonstrated overall insignificant prediction uncertainties and outperformed the taxonomy-based mapping schemes. However, for phenotypes which are largely conserved at least on the level of species (such as B-vitamin synthesis), even the taxonomy-based predictions were of reasonable accuracy. It suggests the applicability of our approach for the metabolic profiling of samples that lack a 16S sequencing data and that are described by taxonomic profiles (e.g., originating from shotgun metagenomic sequencing or qPCR). In addition to the abundance-based description of functional traits (phenotypes) in terms of their Community Phenotype Indices, we also considered two diversity-based metrics, Phenotype Alpha Diversity and Phenotype Beta Diversity, that describe the diversity of sub-communities of phenotype carriers. Overall, greater variations of CPI were observed for phenotypes with a low PAD and vice versa, phenotypes with large PAD values demonstrated moderate to low variation of CPI. This makes PAD likely accounting for the phenotype stability in the face of environmental perturbations. Being also a useful criterion for the selection of phylogenetically well-diversified phenotypes for classification tasks (Iablokov et al., 2021), PAD metric itself represents a complementary (to CPI) description of microbial communities, and, when used as feature, is expected to improve the performance of classification and provide additional insights based on phenotype diversity. The PBD metric introduced in this study was used in the comparative analysis of HGM samples from healthy vs. Crohn's disease cohorts. Notably, PBD values for a subset of phenotypes (Butyrate, Lac, B12) were much lower for the healthy subjects as compared to Crohn's disease patients. This illustrates a potential diagnostic utility of PBD metric for a diversity-based detection of metabolic features associated with a particular syndrome. Both phenotype diversity metrics (PAD and PBD) can be also adopted to the sub-communities of phenotype non-carriers such as in the analysis of auxotrophy for essential nutrients (e.g., B-vitamins).

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: www.ebi.ac.uk/ena, project IDs PRJEB11419 (AGP), PRJEB13747 (UKT), PRJNA422193 (ESP), and PRJEB22028 (CHN); ega-archive.org, project IDs EGAS00001002702 (NLD IBD) and EGAS00001001704 (NLD Healthy, LifeLines DEEP).

AUTHOR CONTRIBUTIONS

DR, PN, and SI conceived and designed the research project. SI performed the primary analysis of the sequencing data. SI and DR performed the phenotype profiling. SI, DR, and AO wrote the manuscript. All authors developed the CPI/MTA/PAD/PBD concepts, read, and approved the final manuscript.

FUNDING

This research was supported by the NIH (RO1 DK030292-35) and the Russian Science Foundation (grant #19-14-00305).

ACKNOWLEDGMENTS

We are grateful to the PhenoBiome Inc. R&D team (Walnut Creek, CA, www.phenobiome.com) for providing us with the access to a development version of the Phenotype Profiler tool. We 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/fmicb.2021.653314/full#supplementary-material

REFERENCES

Blanton, L. V., Charbonneau, M. R., Salih, T., Barratt, M. J., Venkatesh, S., Ilkaveya, O., et al. (2016). Gut bacteria that prevent growth impairments transmitted by microbiota from malnourished children. *Science* 351:aad3311. doi: 10.1126/science.aad3311

Bode, L. (2009). Human milk oligosaccharides: prebiotics and beyond. *Nutr. Rev.* 67(Suppl. 2), S183–S191. doi: 10.1111/j.1753-4887.2009.00239.x

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857. doi: 10.1038/s41587-019-0209-9

Borre, Y. E., O'Keeffe, G. W., Clarke, G., Stanton, C., Dinan, T. G., and Cryan, J. F. (2014). Microbiota and neurodevelopmental windows: implications for brain disorders. *Trends Mol. Med.* 20, 509–518. doi: 10.1016/j.molmed.2014.05.002

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bansal, P., Bridge, A. J., et al. (2016). UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. *Methods Mol. Biol.* 1374, 23–54. doi: 10.1007/978-1-4939-3167-5_2

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421

Cani, P. D., and Van Hul, M. (2020). Gut microbiota and obesity: causally linked? *Expert Rev. Gastroenterol. Hepatol.* 14, 401–403. doi: 10.1080/17474124.2020.1758064

Caruso, R., Lo, B. C., and Nunez, G. (2020). Host-microbiota interactions in inflammatory bowel disease. *Nat. Rev. Immunol.* 20, 411–426. doi: 10.1038/s41577-019-0268-7

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

Douglas, G. M., Maffei, V. J., Zaneveld, J. R., Yurgel, S. N., Brown, J. R., Taylor, C. M., et al. (2020). PICRUSt2 for prediction of metagenome functions. *Nat. Biotechnol.* 38, 685–688. doi: 10.1038/s41587-020-0548-6

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

Elbourne, L. D., Tetu, S. G., Hassan, K. A., and Paulsen, I. T. (2017). TransportDB 2.0: a database for exploring membrane transporters in sequenced genomes from all domains of life. *Nucleic Acids Res.* 45, D320–D324. doi: 10.1093/nar/gkw1068

Elmen, 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

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biol. Conserv.* 61, 1–10. doi: 10.1016/0006-3207(92)91201-3

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

Gehrig, J. L., Venkatesh, S., Chang, H. W., Hibberd, M. C., Kung, V. L., Cheng, J., et al. (2019). Effects of microbiota-directed foods in gnotobiotic animals and undernourished children. *Science* 365:aau4732. doi: 10.1126/science.aau4732

Goodrich, J. K., Davenport, E. R., Beaumont, M., Jackson, M. A., Knight, R., Ober, C., et al. (2016). Genetic determinants of the gut microbiome in UK twins. *Cell Host Microbe* 19, 731–743. doi: 10.1016/j.chom.2016.04.017

Grochowska, M., Laskus, T., and Radkowski, M. (2019). Gut microbiota in neurological disorders. *Arch. Immunol. Ther. Exp.* 67, 375–383. doi: 10.1007/s00005-019-00561-6

Gurung, M., Li, Z., You, H., Rodrigues, R., Jump, D. B., Morgun, A., et al. (2020). Role of gut microbiota in type 2 diabetes pathophysiology. *EBioMedicine* 51:102590. doi: 10.1016/j.ebiom.2019.11.051

Guven, D. C., Aktas, B. Y., Simsek, C., and Aksoy, S. (2020). Gut microbiota and cancer immunotherapy: prognostic and therapeutic implications. *Future Oncol.* 16, 497–506. doi: 10.2217/fon-2019-0783

Human Microbiome Project Consortium (2012). A framework for human microbiome research. *Nature* 486, 215–221. doi: 10.1038/nature11209

Iablokov, S. N., Klimenko, N. S., Efimova, D. A., Shashkova, T., Novichkov, P. S., Rodionov, D. A., et al. (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

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. B., 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* 12:1813534. doi: 10.1080/19490976.2020.1813534

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 40, D109–D114. doi: 10.1093/nar/gkr988

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

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., and Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* 42, D490–D495. doi: 10.1093/nar/gkt1178

Lozupone, C., and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* 71, 8228–8235. doi: 10.1128/AEM.71.12.8228-8235.2005

Magnusdottir, S., Heinken, A., Kutt, L., Ravcheev, D. A., Bauer, E., Noronha, A., et al. (2017). Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat. Biotechnol.* 35, 81–89. doi: 10.1038/nbt.3703

Marques, F. Z., Mackay, C. R., and Kaye, D. M. (2018). Beyond gut feelings: how the gut microbiota regulates blood pressure. *Nat. Rev. Cardiol.* 15, 20–32. doi: 10.1038/nrcardio.2017.120

McDonald, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., Ackermann, G., et al. (2018). American gut: an open platform for citizen science microbiome research. *mSystems* 3:e00031-18. doi: 10.1128/mSystems.00031-18

Novichkov, P. S., Kazakov, A. E., Ravcheev, D. A., Leyn, S. A., Kovaleva, G. Y., Sutormin, R. A., et al. (2013). RegPrecise 3.0–a resource for genome-scale exploration of transcriptional regulation in bacteria. *BMC Genomics* 14:745. doi: 10.1186/1471-2164-14-745

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–D745. doi: 10.1093/nar/gkv1189

Osterman, A. L., Overbeek, R., and Rodionov, D. A. (2010). "The use of subsystems to encode biosynthesis of vitamins and cofactors," in *Comprehensive Natural Products II: Chemistry and Biology*, eds L. N. Mander and H. Liu (Kidlington: Elsevier Ltd), 141–159.

Overbeek, R., Bartels, D., Vonstein, V., and Meyer, F. (2007). Annotation of bacterial and archaeal genomes: improving accuracy and consistency. *Chem. Rev.* 107, 3431–3447. doi: 10.1021/cr068308h

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

Pan, S., and Reed, J. L. (2018). Advances in gap-filling genome-scale metabolic models and model-driven experiments lead to novel metabolic discoveries. *Curr. Opin. Biotechnol.* 51, 103–108. doi: 10.1016/j.copbio.2017.12.012

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

Poyet, M., Groussin, M., Gibbons, S. M., Avila-Pacheco, J., Jiang, X., Kearney, S. M., et al. (2019). A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nat. Med.* 25, 1442–1452. doi: 10.1038/s41591-019-0559-3

Price, M. N., and Arkin, A. P. (2017). PaperBLAST: text mining papers for information about homologs. *mSystems* 2:e00039-17. doi: 10.1128/mSystems.00039-17

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

Raman, A. S., Gehrig, J. L., Venkatesh, S., Chang, H. W., Hibberd, M. C., Subramanian, S., et al. (2019). A sparse covarying unit that describes healthy and impaired human gut microbiota development. *Science* 365:aau4735. doi: 10.1126/science.aau4735

Ramirez-Perez, O., Cruz-Ramon, V., Chinchilla-Lopez, P., and Mendez-Sanchez, N. (2017). The role of the gut microbiota in bile acid metabolism. *Ann. Hepatol.* 16(Suppl. 1), S21–S26. doi: 10.5604/01.3001.0010.5672

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., Yang, C., Li, X., Rodionova, I. A., Wang, Y., Obraztsova, A. Y., et al. (2010). Genomic encyclopedia of sugar utilization pathways in the *Shewanella* genus. *BMC Genomics* 11:494. doi: 10.1186/1471-2164-11-494

Romine, M. F., Rodionov, D. A., Maezato, Y., Osterman, A. L., and Nelson, W. C. (2017). Underlying mechanisms for syntrophic metabolism of essential enzyme cofactors in microbial communities. *ISME J.* 11, 1434–1446. doi: 10.1038/ismej.2017.2

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

Stoddard, S. F., Smith, B. J., Hein, R., Roller, B. R., and Schmidt, T. M. (2015). rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Res.* 43, D593–D598. doi: 10.1093/nar/gku1201

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

Tigchelaar, E. F., Zhernakova, A., Dekens, J. A., 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

Wang, G., Huang, S., Wang, Y., Cai, S., Yu, H., Liu, H., et al. (2019). Bridging intestinal immunity and gut microbiota by metabolites. *Cell Mol. Life Sci.* 76, 3917–3937. doi: 10.1007/s00018-019-03190-6

Wemheuer, F., Taylor, J. A., Daniel, R., Johnston, E., Meinicke, P., Thomas, T., et al. (2020). Tax4Fun2: prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences. *Environ. Microb.* 15:11. doi: 10.1186/s40793-020-00358-7

Zhang, X., and Reed, J. L. (2014). Adaptive evolution of synthetic cooperating communities improves growth performance. *PLoS One* 9:e108297. doi: 10.1371/journal.pone.0108297

Zhang, Y., Thiele, I., Weekes, D., Li, Z., Jaroszewski, L., Ginalski, K., et al. (2009). Three-dimensional structural view of the central metabolic network of *Thermotoga maritima*. *Science* 325, 1544–1549. doi: 10.1126/science.1174671

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:e00188-17. doi: 10.1128/mSystems.00188-17

Zou, Y., Xue, W., Luo, G., Deng, Z., Qin, P., Guo, R., et al. (2019). 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat. Biotechnol.* 37, 179–185. doi: 10.1038/s41587-018-0008-8

Conflict of Interest: PN, DR, and AO are co-founders of PhenoBiome Inc., a company pursuing development and biomedical applications of computational tools for predictive phenotype profiling of microbial communities.

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.