RESEARCH

Evaluating trait-based sets for taxonomic enrichment analysis applied to human microbiome data sets

Quang P. Nguyen[1,2*??], Anne G. Hoen[1,2??] and H. Robert Frost[1??]

[*]Correspondence:
quangpmnguyen@gmail.com
[1]Department of Biomedical Data Science, Dartmouth College, Hanover, NH, USA
Full list of author information is available at the end of the article
[†]Co-corresponding authors

Abstract

Background: Set-based pathway analysis is a powerful tool that allows researchers to summarize complex genomic variables in the form of biologically interpretable sets. Since the microbiome is characterized by a high degree of inter-individual variability in taxonomic compositions, applying enrichment methods using on functionally driven taxon sets can increase both the reproducibility and interpretability of microbiome association studies. However, there is still an open question of which knowledge base to utilize for set construction. Here, we evaluate microbial trait databases, which aggregates experimentally determined microbial phenotypes, as a potential avenue for meaningful construction of taxon sets.

Methods: Using publicly available microbiome sequencing data sets (both 16S rRNA gene metabarcoding and whole-genome metagenomics), we assessed these trait-based sets on three aspects: first, do they cover the diversity of microbes obtained from a typical data set, and second, do they confer additional predictive power on disease prediction tasks when assessed against measured pathway abundances and PICRUSt2 prediction.

Results: Trait annotations are well annotated to a small number but most abundant taxa within the community, concordant with the concept of the core-peripheral microbiome. This pattern is consistent across all categories of traits and body-sites for whole genome sequencing data, but much more heterogenous and inconsistent in 16S rRNA metabarcoding data due to difficulties in assigning species-level traits to genus. However, trait-set features are well predictive of disease outcomes compared against predicted and measured pathway abundances. Most important trait-set features are more interpreable and reveal interesting insights on the relationship between microbiome, its function, and health outcomes.

Conclusion: We demonstrated that trait-based taxonomic sets are useful resources to explore mechanistically driven hypotheses in microbiome data analysis.

Keywords: microbiome; enrichment analysis; trait-based analysis

Introduction

Advancements in high-throughout sequencing technologies have allowed researchers to characterize the identity and functional potential of a large proportion of microorganisms in human-associated microbiomes. This has enabled efficient study of the link between health outcomes and the microbiota without reliance on currently

limited culture-based approaches [1]. As such, there has been an increase in microbiome profiling studies, primarily aiming towards identifying specific microbes that are differentially abundant between groups of individuals defined by an exposure or disease state vs a control population [2]. However, such analyses face unique computational and statistical challenges [3], which includes addressing the burden of multiple testing and providing meaningful biological interpretations.

This challenge of understanding the results of microbiome analyses in the broader context of biological systems mirrors that of other high-throughput data sets. One approach that has proven to be fruitful in human genomic studies is gene set testing (or pathway analysis) which focuses on analyzing the coordinated expression of groups of genes (termed gene sets or pathways) [4]. From a statistical perspective, set-based statistics are are more reproducible and have greater power compared to their gene-level counter parts [5]. The true benefit of set-based approaches, however, is the ability to incorporate *a priori* knowledge of specific cellular processes [6]. Microbiome differential abundance analyses can also benefit from set-based approaches instead of a microbe-centric approach. In addition to statistical benefits such as reduced dimensionality and sparsity [7, 8], set-based approaches are also more reflective of the underlying biology. Like genes, microbes act in concert with co-abundant partners to drive biochemical processes that interact with the host, thereby impacting health outcomes [9]. For example, when comparing patients with inflammatory bowel disease against healthy subjects, microbes thought to be disease-causing for inflammatory bowel disease were also strongly co-occurring [10], suggesting that they might jointly contribute to the microbiome-disease causal pathway instead of acting as independent factors. This is also represented in the development of therapies, where products often contain multiple strains of bacteria [11, 12]. Furthermore, organizing microbes into functionally-driven groups (also termed "guilds" [9]) is also congruent with the perspective that human microbiomes are complex ecosystems whose properties emerge from localized interactions between microbial communities representing individuals that exploit and contribute to their environment in similar ways [13].

Unfortunately, there is currently limited research in curating and evaluating appropriate microbe annotations similar to the transcriptomic literature. Repositories like the Molecular Signatures Database (MSigDB) [6] aggregate information about gene function across multiple sources, incorporating both laboratory results and computational inferences. Even though similar databases such as Disbiome [14] and MSEA [8] exist, they are usually human-centric and define microbial groups based on their potential to be pathological rather than through common biochemical roles. As such, these databases are limited in generating meaningful hypotheses linking taxonomic changes to ecosystem function especially in novel disease conditions. Trait-based analysis [15, 16, 17], with its long history in traditional macroecological studies [13, 18], is a promising approach to address this gap. Traits directly represent microbial physiological characteristics and metabolic phenotypes (for example, sulfur reduction, nitrate utilization, or gram positivity) and therefore can serve as annotations for potential ecosystem function. For 16S rRNA gene sequencing data sets, where one can only obtain taxonomic abundances, performing enrichment analysis on trait-based sets can elucidate the taxa-function relationship and

identify microbial processes that are differentially active between healthy and diseased patients. For whole genome metagenomic data sets, traits still offer unique perspectives. First, traits are often sourced from the long history of laboratory experiments such as journal articles and *Bergey's Manual of Systematic Archaea and Bacteria* [19] which is different from homology-based sequence queries typically performed to profile gene family abundances. Second, traits are complex phenotypes that represent multiple molecular pathways, which means that they are more comparable to higher-order pathway annotations in hierarchical databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [20] and MetaCyc [21]. As such, utilizing traits as the source to group microbes into functional and phenotypical categories can assist in interpreting microbiome profiling studies, and generating mechanistically meaningful hypotheses that link ecosystem function and its taxa.

Even though trait-based approaches have been utilized in various studies [15, 16, 22, 18], to our knowledge there is currently no effort to formalize trait-based databases in terms of microbial sets and evaluate their utility in a typical enrichment analysis of 16s rRNA metabarcoding or metagenomic data. Here, we constructed taxon sets from pre-existing trait databases at both the species and genus level. Then, we computed the coverage of these traits across different human-associated environments and sequencing approaches. Finally, we evaluated whether trait-based set features confer predictive capacity for diseased individuals compared to measured (from whole genome sequencing data) and predicted (from PICRUSt2 [23]) pathway abundances. Finally, we identified the most important features for prediction and assessed whether they matched existing literature on the microbiome-disease relationship of interest.

Material and methods

All analyses were performed in the R programming language (version 4.1.2) [24] and the Python programming language (version 3.10.4). All graphics were generated using `ggplot2`, `ggsci`, `patchwork`. Tabular data manipulation was performed using `pandas` for python, and `tidyverse` suite of packages for R. Additional packages utilized include: `BiocSet`, `taxizedb`, `phyloseq`, `TreeSummarizedExperiment`. For enrichment analyses, we leveraged the CBEA [7] method (version 1.0.1) developed previously by our lab. All analyses were performed using the `snakemake` workflow [25]. All reproducible code and intermediate analysis products can be found on GitHub (qpmnguyen/microbe_set_trait).

Generating taxonomic sets from trait databases

We utilized pre-compiled trait databases from previous publications: Madin et al. 2020 [17] and Weissman et al. 2021 [15]. The former was chosen due to the fact that it is the most comprehensive compilation of microbial (bacteria and archaea) physiological traits based on existing sources to date. The latter is a newer database that hand curates traits specifically for human microbiomes based on Bergey's manual. Both of these databases source their trait assignments primarily from biochemical and microbiological laboratory experiments over genomic-based annotation. We focused our analyses on categorical traits, namely metabolism, gram stain, enzymatic pathways, sporulation, motility, cellular shape, and substrate utilization. We

are particularly interested in traits belonging to the class of enzymatic pathways and substrate utilization as they represent functions that most directly impact the microbe-host relationship [26].

We combined both databases into a joint knowledge base and constructed sets for each available categorical trait. Additionally for the Madin et al. database, we updated data entries sourced from Genomes Online Database (GOLD) [27] due to the fact that compared to other compiled sources, GOLD is continuously updated via community submissions. We grouped all traits belonging to the same National Center for Biotechnology Information (NCBI) species-level identifier. When there are conflicts in assigning traits, we prioritized Weissman et al. over Madin et al. and GOLD due to its hand curated nature. If there are ambiguities in taxonomic assignment in the Weissman et al. source, we considered that trait to be missing. The exceptions to the above logic are enzymatic pathways and substrate utilization categories where trait values across sources for the same species are concatenated instead of reconciled. For example, if a species A has entries from multiple databases suggesting the presence of "nitrogen degredation" and "ammonia degredation", then instead of attempting to chose the best annotation based on source we assumed that species A has the capacity to metabolize both nitrogen and ammonia.

All traits are defined at the species level via NCBI identifiers, however, due to restrictions for 16S rRNA gene sequencing data sets to resolve beyond the genus level [28], we also assigned traits to each genus based on a two-step process for each major trait category:

- A hypergeometric test is used to ascertain whether the genus is underrepresented in the database based on the total number of species assigned to that genus in NCBI Taxonomy [29] compared to our trait database. If a genus is underrepresented in our database (i.e. the proportion of species number of genera in the database is significantly less than what one would expect if one were to randomly draw species from the NCBI database), then the trait is not assigned to that genus since we do not have enough information. Specifically, we assessed $P(X \leq x)$ at $\alpha = 0.05$ where $X \sim Hypergeometric(k, N, K)$, with x as the total number of species assigned to that genus in the database with an assigned value for the trait category of interest, k as the total number of species in the database with an assigned value for the trait category, N as the total number of species in NCBI Taxonomy, and K as the total number of species assigned to the genus in NCBI Taxonomy.
- For all genera that are well represented in the database, we then assessed the proportion of species under that genus that have the trait. If over 95% of species of a given genus have the trait, then the trait is assigned to the genus.

We then defined trait-based sets using the aforementioned assignments. Each trait value with a category, e.g. "obligate anaerobic" from the category "metabolism", is defined as a set with elements representing the species (or genus) annotated to that trait value. In the analysis stage, each identified taxon within a data set is matched to a trait based on their NCBI identifier. For 16S rRNA gene metabarcoding data sets, we matched all amplicon sequence variants (ASV) with traits belonging to the genus level NCBI identifier matched to the ASV sequence. All processed databases and resulting taxonomic sets can be found on GitHub in the analysis repository.

Evaluation data sets

We evaluated trait-based sets on publicly available 16S rRNA gene metabarcoding and whole-genome metagenomic data sets. For study-specific metabarcoding data sets, we obtained data directly from associated European Nucleotide Archive (ENA) repositories and re-processed raw sequence files into ASV tables using the `dada2` QIIME 2 (version 2022.2) plugin [30, 31]. Taxonomic classification was performed using a pre-trained weighted naive bayes model [32, 33] using the SILVA NR 99 database version 138 [34] available via QIIME 2. For all our metagenomic data sets, we downloaded taxonomic and pathway abundance tables directly from the `curatedMetagenomicData` R package [35] (2021-10-19 snapshot), which processed the data via the `bioBakery` [36] metagenomic data processing pipeline by the package authors. Data from the Human Microbiome Project (HMP) was obtained using the `HMP16SData` [37] (for metabarcoding data) and `curatedMetagenomicData` [35] (for metagenomic data) R packages.

To assess trait annotation coverage, we utilized data from both Phase I and II of the HMP [38] as it contains surveys for multiple human-associated environments from healthy subjects. For predictive and concordance analyses, we focused on colorectal cancer (CRC) and inflammatory bowel disease (IBD) as study conditions. Both CRC and IBD are well represented across both metabarcoding and metagenomic data sets, allowing comparisons across sequencing methodology. Furthermore, these conditions are also under active study within the microbiome literature, which improves the ability to interpret the biological significance of the results. For CRC, we utilized data from Zeller et al. [39], Feng et al. [40], Gupta et al. [41], Hannigan et al. [42], Thomas et al. [43], Vogtmann et al. [44], Wirbel et al. [45], Yachida et al. [46], and Yu et al. [47]. For IBD, we utilized data from the integrative HMP [48], Gevers et al. [10], Hall et al. [49], Ijaz et al. [50], Li et al. [51], Nielsen et al. [52], and Vich Vila et al. [53]. A detailed description of each data set and data-processing procedures is available in the Supplementary Materials.

Coverage analysis

In this analysis, we sought to identify how well trait databases cover the taxonomic diversity of different human-associated environments. We leveraged healthy samples from multiple body sites from Phase I and II of the HMP [38]. We quantified coverage as a per-sample measure considering both taxa absence/presence and its abundance.

- For each sample, we computed the proportion of taxa that is present (non-zero counts) assigned to at least one trait (a sample-level trait-specific richness).
- For each sample, we computed the proportion of reads assigned to taxa that is present and annotated to at least one trait (a sample-level trait-specific evenness).

In addition to coverage stratified by trait categories and body sites, we also generated category-specific and site-specific coverage values by averaging across all sites or categories, respectively.

Prediction analysis

We also aimed to evaluate whether trait-based features can add information for microbiome-based disease prediction compared to other data inputs. Here, we generated sample-level enrichment scores for each trait using `CBEA` and utilized them

as inputs to a standard random forest model [54]. Model fitting was done using `scikit-learn` [55] where all parameters were set to default values with the exception of the total number of trees per ensemble (500) and the total number of features considered per split (equal to the square root of the total number of features). We compared model performance using trait enrichment scores against measured and PICRUSt2 predicted pathway abundances (for metagenomic and metabarcoding data sets, respectively).

Model performance was measured using the area under the receiver operating characteristic curve (AUROC) and Brier scores [56]. These metrics and associated confidence intervals were obtained by fitting and evaluating the model via a 10-fold cross-validation procedure. To obtain calibrated predictive probabilities for Brier scores, we applied Platt's method (using `CalibratedClassifierCV`) with 5-fold cross validation nested within the training fold and used the ensemble model to generate test set probabilities [57].

In order to identify which features are important to the disease prediction process, for each input type we re-split the entire data set into train/test splits (80% training data). We then refitted our calibrated random forest model on the training set as described above. Since our final model is an ensemble of calibrated random forest classifiers, we obtained feature importance values as the average across all calibrated cross validation folds ($N = 10$). Feature importance per random forest model is defined based on the implementation in `scikit-learn` as the decrease in Gini impurity when the feature is split averaged across all decision trees in a forest.

Results

Trait-based taxonomic sets

Our combined database contains a total of 6,795 unique traits across 5 different categories as detailed in Table 1. A total of 102,432 unique species and 9,373 unique genus was assigned to at least one trait. A unique species or genus is defined as those with a unique NCBI taxonomic identifier at the associated level.

Table 1 Overview of trait database

Trait Category	Unique Sub-traits	Unique Species[1]	Unique Genus[1]
Sporulation	2	12,400	1,929
Gram Stain	2	41,265	2,713
Motility	6	14,896	2,185
Metabolism	8	19,477	2,694
Cell Shape	17	15,635	2,198
Pathways	1,838	6,242	1,777
Substrate	4,922	5,596	1,652

[1] Number of unique species/genus represents the number of species/genus assigned to at least one trait in the category

Database coverage

We computed the coverage for each trait category across each body site in the HMP data set. Fig 1 illustrates results for species-level trait assignment for samples profiled via whole genome metagenomics. In panel A, coverage is evaluated as the total number of taxa present per sample annotated to a trait (a measure of cross-trait richness), while in panel B coverage is the total number of reads assigned to taxa annotated to a trait (a measure of cross-trait evenness). Richness provides

a general overview on how many members of a community is assigned to a trait, while evenness accounts for their relative abundances by up-weighting species that have high abundance across all samples. Overall, for any body site, at most 25% of taxa are assigned to a trait, but when considering the proportion of reads, coverage increased to more than 80%. This shows that traits are usually well annotated to the most abundant taxa. This pattern holds for samples profiled with 16S rRNA gene sequencing (Fig S1), even though the proportions were much lower due to difficulties in aggregating species level traits to genus. For many body sites and trait category combinations, traits could not be assigned to any taxa.

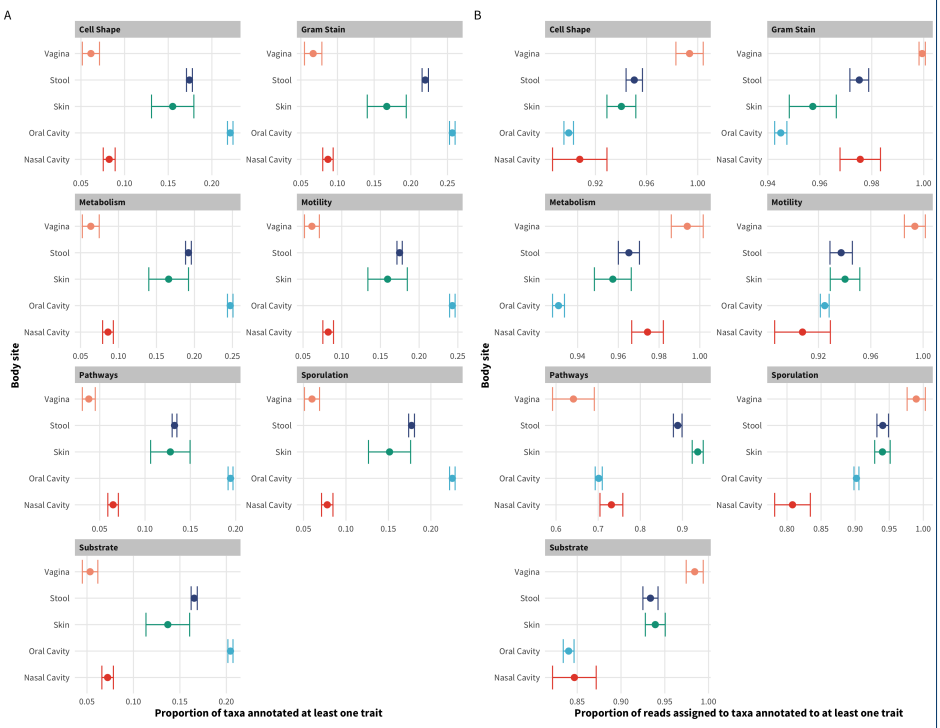

Figure 1 Trait annotation coverage across different body sites for the HMP data set profiled using whole genome shotgun sequencing. Panel **(A)** illustrates the proportion of present taxa per sample annotated to at least one trait. Panel **(B)** illustrates the proportion of reads assigned to taxa annotated to at least one trait which accounts for taxa relative abundances. Each plot facet represents different trait categories that were evaluated. Error bar represents the standard error of the evaluation statistic of interest across the total number of samples evaluated per body site.

We also observed heterogeneity in the annotation coverage across different body sites and trait categories. For richness, nasal cavity and vaginal body sites were the lowest in coverage, with less than 5% of taxa annotated with at least one trait across all trait categories while conversely, oral cavity sites consistently had the highest coverage under this metric. This pattern was reversed when considering coverage as the proportion of assigned reads per sample, but overall values were consistently high. Averaging coverage across body sites (Fig 2) also supports this observation, showing overall that the proportion of reads covered are similar across all body sites despite differences in the proportion of present taxa covered by trait annotations. Similar results were observed for sites profiled with 16S rRNA gene sequencing (Fig

S1), where oral sub-sites have the highest coverage across both richness and evenness metrics but, on average, all sites were similar in coverage statistics. Surprisingly, stool samples were low in coverage across multiple categories despite being one of the well studied systems.

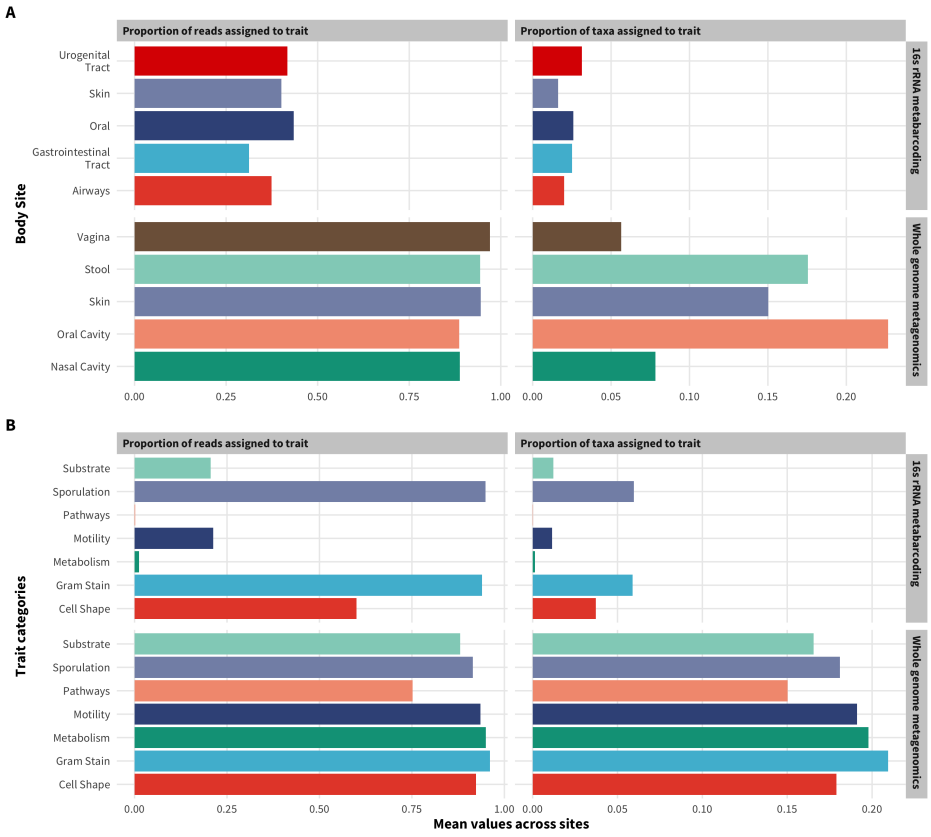

Figure 2 Trait annotation coverage statistics for HMP data across samples profiled with both 16S rRNA gene metabarcoding and whole genome metagenomics. Panel **(A)** illustrates coverage statistics for each body site averaged across evaluated samples and trait categories. Panel **(B)** illustrates statistics for each trait category averaged across evaluated samples and body sites.

We also stratified our coverage analyses by trait categories (Fig 1, Fig 2, Fig S1). For samples profiled with whole genome sequencing, all trait categories are evenly covered, with about 15% - 20% of taxa were annotated to a trait of any category. However, these taxa comprise around 75% to 100% of the total reads per sample suggesting that the overall read level coverage is very high. However, in samples profiled with 16S rRNA gene sequencing, the overall coverage value across categories is low. Sporulation, substrate utilization and motility are the most covered category while pathways and metabolism has no coverage at all.

Predictive analysis

To determine the utility of trait-based sets, we generated enrichment scores for covered traits using CBEA [7] for evaluated data sets and compared the predictive performance of using trait-set enrichment scores as inputs compared to alternative

functional-based predictors. We evaluated two disease conditions, CRC and IBD, with data sets drawn from both 16S rRNA gene metabarcoding and whole genome metagenomic profiling techniques. We fitted a calibrated random forest model to each input type and computed predictive performance as AUROC (discriminatory power) and Brier scores (probability estimates) using 10-fold cross-validation.

For the CRC prediction task, traits covered 2.7% of taxa and 27.3% of reads for the 16S rRNA gene metabarcoding data set, while for the whole genome sequencing data set, traits covered 9.1% of taxa and 87.2% of reads. For the IBD prediction task, traits covered 1.61% of taxa and 26.7% of reads for 16S rRNA gene metabarcoding data set, while for the whole genome sequencing data set, traits covered 6.6% of taxa and 91.2% of reads.

Fig 3 illustrates results of our model evaluations. Overall, enrichment scores for trait-sets are as good as other alternate function-based predictors at discriminating between case and control patients across both CRC and IBD conditions. Aside from pure discrimination power, models fitted on CBEA trait-set scores are also equivalent in approximating predicted probabilities. This is surprising especially for the 16S rRNA gene metabarcoding data sets, where the trait coverage is low. Even though the differences in performance is not significant, there are instances where trait-set scores perform slightly better than their pathway abundance counterparts. Since trait-features are also more descriptive, utilizing them can increase interpretation while also not scarificing performance.

In addition to predicted performance, we also identified the top 10 features that are most important for model fitting. Since our model involves a 10-fold cross-validation procedure within the training set to calibrate predicted probabilities, top features are identified using the mean feature importance value across the 10 folds. Fig 4 illustrates results for whole genome metagenomic data sets while Fig S2 illustrates results for the 16S rRNA gene metabarcoding data sets. Even though these are the top 10 features, the observed mean feature importance statistics are low, suggesting that no individual features were definitively the most important in discriminating between patient classes.

Discussion

Trait are annotated with high coverage at the species-level

Traits are annotated with high coverage at the species-level

We computed the coverage of trait annotation on a typical dataset to understand the extent in which community function is captured, thereby serving as a proxy for expected confidence for an enrichment analysis performed using trait-based taxon sets. Low coverage in this case indicates that the database does not adequately capture the diversity of microbes found in the target data. This is because there might not be enough taxa present in the data set to serve as evidence for the trait. Alternately, this could also mean that the analysis is missing a majority of underlying community traits, many of which might be core to the health outcome association of interest but simply missing in the analysis. We computed coverage based on two metrics: first, a richness-like metric which computes coverage as the proportion of taxa present per sample annotated to a trait (for a given category and data set); second, an evenness-like metric that accounts for relative abundances of

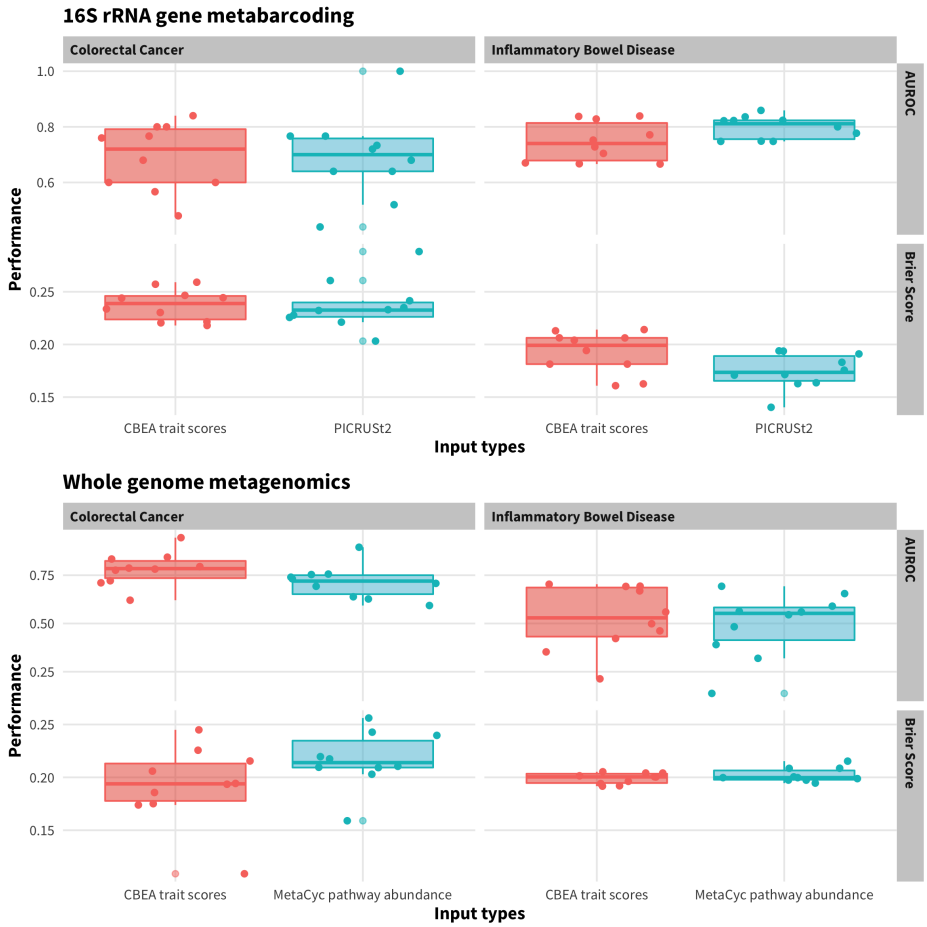

Figure 3 Predictive performance of a calibrated random forest model across different disease conditions, profiling technique, and performance metrics. Scores were obtained via 10-fold nested cross validation where within each training fold there is a 10-fold cross validation procedure to calibrate predicted probabilities. For each condition and data type, CBEA trait-set scores were compared against MetaCyc pathway abundances from relevant sources (measured abundances for whole genome sequencing data sets and PICRUSt2 predicted abundances for 16S rRNA gene metabarcoding data sets)

each annotated taxa by computing coverage as the proportion of reads per sample annotated to a trait.

When evaluated on the HMP data set (Fig 1), we can see that the overall richness coverage is low (less than 25% of identified species) across all sites and data sets, particularly for nasal cavity and vaginal sub-sites. However, when considering evenness of coverage, almost all of reads were annotated to a trait. This is consistent with the observation that relative abundances of human-associated microbiomes are highly skewed [38], where a small number of species usually dominate the community. As such, even though traits might only cover a small number of taxa, they might represent the majority of community abundance. For example, Ravel et al. [58] observed that *Lactobacillus* species dominate the vaginal microbiome and, in some phylotypes, almost all reads are assigned to a single species. This shows that our trait-database has high degree of coverage across the most abundant taxon

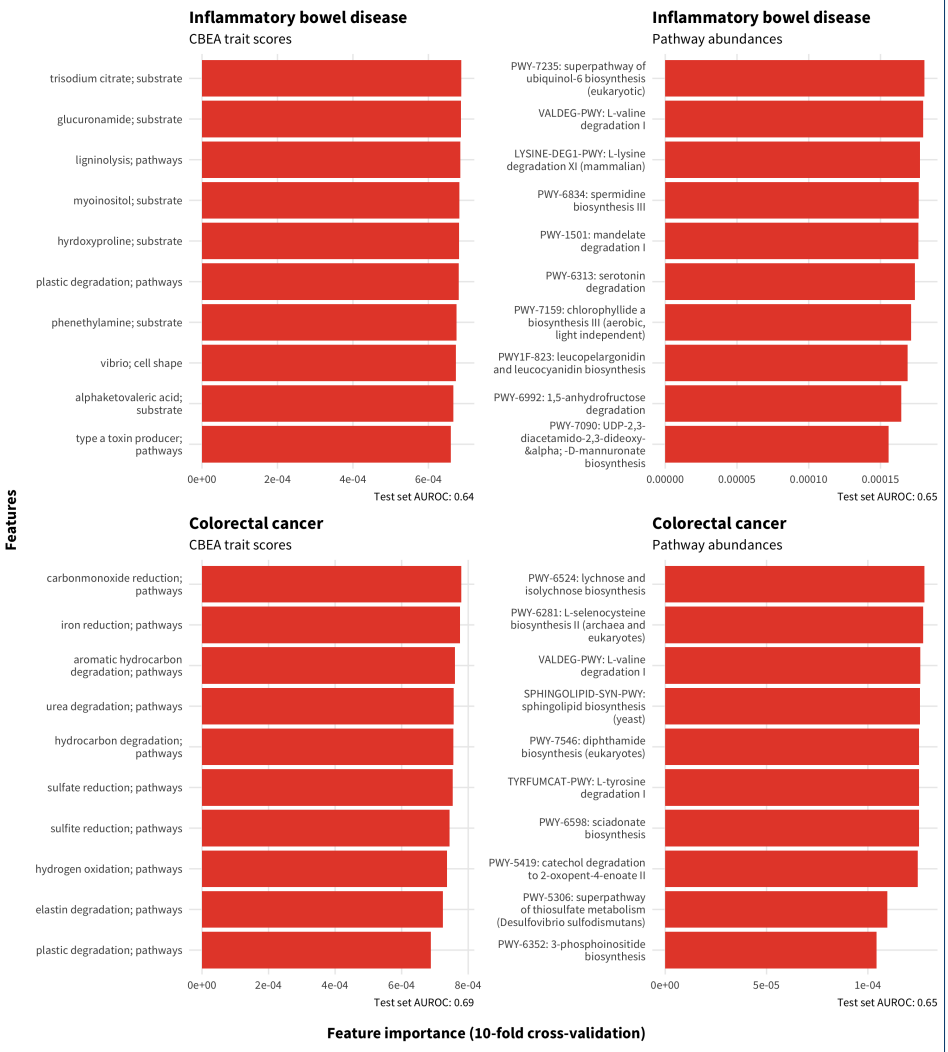

Figure 4 Top 10 important features based on random forest model fitted different inputs from data sets profiled with whole genome metagenomics. Features were selected from mean decrease in Gini impurity averaged across 500 decision trees and 10-fold cross-validation (nested with the training set) as implemented in `scikit-learn`. AUROC scored on a held-out test set is also presented for each input type and disease condition.

within a community, which supports utilizing these sets to perform exploratory analyses. However, low richness coverage also indicates that our database might not capture traits associated with rare taxa, which can play an important role in regulating host health [59].

Unfortunately, coverage is significantly lower for samples profiled using 16S rRNA gene metabarcoding (Fig S1). For some trait categories, such as pathways, no traits were assigned to any taxa (Fig 2). We hypothesized that this is due to two issues. First, metabarcoding data sets (especially those using a fragment of the 16S rRNA gene) can only resolve taxonomies at the genus level [28], while traits are usually defined at the species and strain levels. Aggregating consensus traits to the genus is difficult due to the high degree of strain and species level diversity within the microbiome [60]. Second, taxonomic assignments for metabarcoding data sets are

often based amplification of a specific hyper-variable region for a marker gene (most often the 16S rRNA gene). This means that taxonomic assignment can be sensitive to the choice of region, and can be inaccurate. Furthermore, the choice of taxonomic database (e.g. Ribosomal Database Project [61], SILVA [34]) can also play a part in reducing the ability for trait annotation coverage. Differences between taxonomic paths [62] can result in certain taxa not being able to be matched to traits, which usually assign traits based on NCBI identifiers.

Trait-set features are predictive of disease outcomes

We assessed the predictive performance of models fit on trait-set enrichment scores compared to other function-based inputs. For whole genome sequencing data sets, measured pathway abundances were utilized as a comparison point while for 16S rRNA gene sequencing data sets, predicted pathway abundances via PICRUSt2 were utilized instead. Fig 2 shows that across all conditions and profiling techniques, trait-set features are competitive in producing well performing models and were able to discriminate between cases and controls. Surprisingly, performance was also comparable in the 16S rRNA gene sequencing data set despite overall low coverage across both richness and evenness metrics. This demonstrates that trait-set abundances can still provide an informative approximation to functional potential similar to PICRUSt2 that can be used for exploratory and hypothesis generating purposes.

To determine which features are important for overall model performance, we extracted the top 10 features based on the mean decrease in Gini impurity. However, the overall feature importance values are not high, suggesting that no individual feature was dominant in classifying patient status. This is further supported by the fact that some nonsense features show up in the top 10 list for models fit using pathway abundances such as PWY-7235 and LYSINE-DEG-PWY, which are mammalian and eukaryotic pathways, respectively. However, the models still show respectable discriminatory power when evaluated on the test set (AUROC ~ 0.7). Since random forest models can capture interactions between predictors [63], we hypothesized that the interaction between features contribute to test set performance rather than marginal effects. As such, we did not observe a high degree of feature importance scores since these measures are not designed to capture interaction effects [64].

However, despite such limitations, we were still able to recover existing knowledge about the condition of interest. For example, "sulfide reduction;pathways" was shown to be an important feature in discriminating subjects with CRC vs control subjects in Fig 4. This is supported by previous research showing that an increase in abundance of sulfate reducing bacteria is associated with the condition [46]. Mechanistically, this process, when using methionine or cysteine as substrates [65], generates H_2S as a product, which can stimulate CRC by inhibiting butyrate oxidation (which helps prevent the breakdown of the gut barrier) as well as promoting the generation of reactive oxygen species [66]. Another trait feature is "urea degredation;pathways", which suggests the importance of bacterial-driven urea hydrolysis process, which is one of the main sources of ammonia in the human gut [67].

Sustained exposure of colonocytes to free ammonia may contribute to the development of CRC [68], which is supported by animal experiments showing histological damage in the distal colon after long-term ammonium exposure [69].

Limitations and future directions

Even though our results demonstrate that utilizing trait-based sets can provide meaningful insight to microbiome data sets, there are several major challenges to widespread adoption. Although trait databases do not suffer from the same types of biases that exist in genomic reference databases [70], the reliance on curated experimental data means that traits are usually only annotated for species that are well studied and culturable. While using predictive models can help in assigning traits to a broader category of taxa [71], such automated approaches can result in misclassification of traits and increased noise in downstream analyses. Additionally, high-quality trait annotations require a time-consuming, manual curation process [15]. A source that is based on user submission such as GOLD [27] can cover a larger number of taxa and traits, but unfortunately can have erroneous and duplicated assignments due to the lack of a standardized nomenclature. There is currently a gap in producing a high-quality and diverse trait databases that are maintained and continuously updated.

In addition to issues with trait database quality, there are also problems matching the identity of taxa in a given trait database with identifiers found in references for sequence-based taxonomic profiling such as SILVA [34]. For whole genome metagenomic data sets, standard tools (such as MetaPhlan [36]) can provide NCBI identifiers at the species or strain levels. However, it is currently unclear how to aggregate or disaggregate traits if the taxonomic resolution of the observed data set is higher or lower than that of the trait database in use. This is even more difficult with metabarcoding datasets, where low taxonomic resolution makes trait-to-taxa assignments sparse and less confident.

Finally, there are also hurdles in being able to properly validate traits that are found to be significantly enriched due to a lack of ground truth data sets. While some traits can be matched to pathways directly, others involve complex coordination of multiple genetic pathways. As such, further investigation into ways to identify biological concordance between obtained results and external measurements can help improve confidence in utilizing traits for microbiome analyses.

Conclusion

Set-based enrichment analysis is a useful approach for analyzing microbiome data sets since it not only reflects underlying biology but can also provide more unique perspectives of function that is linked to ecosystem services. Microbial trait databases are a promising resource to construct taxon-sets as traits represent physiological phenotypes. We demonstrated that trait-based sets have high coverage across body sites, especially for samples profiled using whole genome metagenomics. Furthermore, enrichment scores computed on such sets are also competitive in predicting case/control status compared to pathway abundances. As such, trait features found to be important in model fitting can be used to define interesting mechanistic hypotheses.

Appendix

S1 Fig. **Trait coverage statistics for samples profiled with 16S rRNA gene metabarcoding.** Panel **(A)** illustrates the proportion of present taxa per sample annotated to at least one trait. Panel **(B)** illustrates the proportion of reads assigned to taxa annotated to at least one trait which accounts for taxa relative abundances. Each plot facet represents different trait categories that were evaluated. Error bar represents the standard error of the evaluation statistic of interest across the the total number of samples evaluated per body site.

S2 Fig. **Top 10 important features based on random forest model fitted different inputs from data sets profiled with 16S rRNA gene sequencing**. Features were selected from mean decrease in Gini impurity averaged across 500 decision trees and 10-fold cross-validation (nested with the training set) as implemented in `scikit-learn`. AUROC scored on a held-out test set is also presented for each input type and disease condition.

Acknowledgements
We would like to acknowledge Dr. Becky Lebeaux, Dr. Levi Waldron, Dr. Margaret Karagas, and Dr. Brock Christensen for their comments on this manuscript.

Funding
We would like to acknowledge funding provided by National Institutes of Health grants R21CA253408, P20GM130454, R01LM012723 and P30CA023108. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abbreviations
CRC: Colorectal Cancer
HMP: Human Microbiome Project
IBD: Inflammatory Bowel Disease
CBEA: Competitive Balances for Taxonomic Enrichment Analysis
ENA: European Nucleotide Archive

Availability of data and materials
All data for this manuscript is publicly available either via ENA projects () or the `curatedMetagenomicData` and `HMP16SData` R packages. Reproducible scripts are also available on GitHub at `https://www.github.com/qpmnguyen/microbe_set_trait`

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
QPN, AGH, and HRF conceived of the manuscript and planned the experiments. QPN performed the analyses and wrote up the manuscript. QPN, AGH, and HRF contributed to improving and revising the manuscript.

Authors' information
Author details
[1]Department of Biomedical Data Science, Dartmouth College, Hanover, NH, USA. [2]Department of Epidemiology, Dartmouth College, Hanover, NH, USA.

References
1. Lagier, J.-C., Khelaifia, S., Alou, M.T., Ndongo, S., Dione, N., Hugon, P., Caputo, A., Cadoret, F., Traore, S.I., Seck, E.H., Dubourg, G., Durand, G., Mourembou, G., Guilhot, E., Togo, A., Bellali, S., Bachar, D., Cassir, N., Bittar, F., Delerce, J., Mailhe, M., Ricaboni, D., Bilen, M., Dangui Nieko, N.P.M., Dia Badiane, N.M., Valles, C., Mouelhi, D., Diop, K., Million, M., Musso, D., Abrahão, J., Azhar, E.I., Bibi, F., Yasir, M., Diallo, A., Sokhna, C., Djossou, F., Vitton, V., Robert, C., Rolain, J.M., La Scola, B., Fournier, P.-E., Levasseur, A., Raoult, D.: Culture of previously uncultured members of the human gut microbiota by culturomics. Nat Microbiol **1**(12), 16203 (2016). doi:10.1038/nmicrobiol.2016.203
2. Zhang, X., Li, L., Butcher, J., Stintzi, A., Figeys, D.: Advancing functional and translational microbiome research using meta-omics approaches. Microbiome **7**(1), 154 (2019). doi:10.1186/s40168-019-0767-6
3. Li, H.: Statistical and Computational Methods in Microbiome and Metagenomics. In: Balding, D., Moltke, I., Marioni, J. (eds.) Handbook of Statistical Genomics, 1st edn., pp. 977–550. Wiley, ??? (2019). doi:10.1002/9781119487845.ch35
4. Maleki, F., Ovens, K., Hogan, D.J., Kusalik, A.J.: Gene Set Analysis: Challenges, Opportunities, and Future Research. Front. Genet. **11**, 654 (2020). doi:10.3389/fgene.2020.00654
5. Goeman, J.J., Bühlmann, P.: Analyzing gene expression data in terms of gene sets: Methodological issues. Bioinformatics **23**(8), 980–987 (2007). doi:10.1093/bioinformatics/btm051
6. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., Tamayo, P.: The Molecular Signatures Database Hallmark Gene Set Collection. Cell Systems **1**(6), 417–425 (2015). doi:10.1016/j.cels.2015.12.004
7. Nguyen, Q.P., Hoen, A.G., Frost, H.R.: Cbea: Competitive balances for taxonomic enrichment analysis. PLoS Computational Biology **18**(5), 1010091 (2022)
8. Kou, Y., Xu, X., Zhu, Z., Dai, L., Tan, Y.: Microbe-set enrichment analysis facilitates functional interpretation of microbiome profiling data. Sci Rep **10**(1), 21466 (2020). doi:10.1038/s41598-020-78511-y
9. Wu, G., Zhao, N., Zhang, C., Lam, Y.Y., Zhao, L.: Guild-based analysis for understanding gut microbiome in human health and diseases. Genome Med **13**(1), 22 (2021). doi:10.1186/s13073-021-00840-y
10. Gevers, D., Kugathasan, S., Denson, L.A., Vázquez-Baeza, Y., Van Treuren, W., Ren, B., Schwager, E., Knights, D., Song, S.J., Yassour, M., Morgan, X.C., Kostic, A.D., Luo, C., González, A., McDonald, D., Haberman, Y., Walters, T., Baker, S., Rosh, J., Stephens, M., Heyman, M., Markowitz, J., Baldassano, R., Griffiths, A., Sylvester, F., Mack, D., Kim, S., Crandall, W., Hyams, J., Huttenhower, C., Knight, R., Xavier, R.J.: The Treatment-Naive Microbiome in New-Onset Crohn's Disease. Cell Host & Microbe **15**(3), 382–392 (2014). doi:10.1016/j.chom.2014.02.005
11. Berg, G., Rybakova, D., Fischer, D., Cernava, T., Vergès, M.-C.C., Charles, T., Chen, X., Cocolin, L., Eversole, K., Corral, G.H., Kazou, M., Kinkel, L., Lange, L., Lima, N., Loy, A., Macklin, J.A., Maguin, E., Mauchline, T., McClure, R., Mitter, B., Ryan, M., Sarand, I., Smidt, H., Schelkle, B., Roume, H., Kiran, G.S., Selvin, J., de Souza, R.S.C., van Overbeek, L., Singh, B.K., Wagner, M., Walsh, A., Sessitsch, A., Schloter, M.: Microbiome definition re-visited: Old concepts and new challenges. Microbiome **8**(1), 103 (2020). doi:10.1186/s40168-020-00875-0
12. Durack, J., Lynch, S.V.: The gut microbiome: Relationships with disease and opportunities for therapy. J Exp Med **216**(1), 20–40 (2019). doi:10.1084/jem.20180448
13. Faust, K., Raes, J.: Microbial interactions: From networks to models. Nat Rev Microbiol **10**(8), 538–550 (2012). doi:10.1038/nrmicro2832

14. Janssens, Y., Nielandt, J., Bronselaer, A., Debunne, N., Verbeke, F., Wynendaele, E., Van Immerseel, F., Vandewynckel, Y.-P., De Tré, G., De Spiegeleer, B.: Disbiome database: Linking the microbiome to disease. BMC Microbiol **18**(1), 50 (2018). doi:10.1186/s12866-018-1197-5
15. Weissman, J.L., Dogra, S., Javadi, K., Bolten, S., Flint, R., Davati, C., Beattie, J., Dixit, K., Peesay, T., Awan, S., Thielen, P., Breitwieser, F., Johnson, P.L.F., Karig, D., Fagan, W.F., Bewick, S.: Exploring the functional composition of the human microbiome using a hand-curated microbial trait database. BMC Bioinformatics **22**(1), 306 (2021). doi:10.1186/s12859-021-04216-2
16. Bewick, S., Gurarie, E., Weissman, J.L., Beattie, J., Davati, C., Flint, R., Thielen, P., Breitwieser, F., Karig, D., Fagan, W.F.: Trait-based analysis of the human skin microbiome. Microbiome **7**(1), 101 (2019). doi:10.1186/s40168-019-0698-2
17. Madin, J.S., Nielsen, D.A., Brbic, M., Corkrey, R., Danko, D., Edwards, K., Engqvist, M.K.M., Fierer, N., Geoghegan, J.L., Gillings, M., Kyrpides, N.C., Litchman, E., Mason, C.E., Moore, L., Nielsen, S.L., Paulsen, I.T., Price, N.D., Reddy, T.B.K., Richards, M.A., Rocha, E.P.C., Schmidt, T.M., Shaaban, H., Shukla, M., Supek, F., Tetu, S.G., Vieira-Silva, S., Wattam, A.R., Westfall, D.A., Westoby, M.: A synthesis of bacterial and archaeal phenotypic trait data. Sci Data **7**(1), 170 (2020). doi:10.1038/s41597-020-0497-4
18. Krause, S., Le Roux, X., Niklaus, P.A., Bodegom, V., M, P., Lennon, J.T., Bertilsson, S., Grossart, H.-P., Philippot, L., Bodelier, P.L.E.: Trait-based approaches for understanding microbial biodiversity and ecosystem functioning. Front. Microbiol. **5** (2014). doi:10.3389/fmicb.2014.00251
19. Trujillo, M.E., Dedysh, S., DeVos, P., Hedlund, B., Kämpfer, P., Rainey, F.A., Whitman, W.B. (eds.): Bergey's Manual of Systematics of Archaea and Bacteria, 1st edn. Wiley, ??? (2015). doi:10.1002/9781118960608
20. Kanehisa, M.: Toward understanding the origin and evolution of cellular organisms. Protein Sci. **28**(11), 1947–1951 (2019). doi:10.1002/pro.3715
21. Caspi, R., Billington, R., Keseler, I.M., Kothari, A., Krummenacker, M., Midford, P.E., Ong, W.K., Paley, S., Subhraveti, P., Karp, P.D.: The MetaCyc database of metabolic pathways and enzymes - a 2019 update. Nucleic Acids Research **48**(D1), 445–453 (2020). doi:10.1093/nar/gkz862
22. Guittar, J., Shade, A., Litchman, E.: Trait-based community assembly and succession of the infant gut microbiome. Nat Commun **10**(1), 512 (2019). doi:10.1038/s41467-019-08377-w
23. Douglas, G.M., Maffei, V.J., Zaneveld, J.R., Yurgel, S.N., Brown, J.R., Taylor, C.M., Huttenhower, C., Langille, M.G.I.: PICRUSt2 for prediction of metagenome functions. Nature Biotechnology **38**(6), 685–688 (2020). doi:10.1038/s41587-020-0548-6
24. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2021). R Foundation for Statistical Computing
25. Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., Köster, J.: Sustainable data analysis with Snakemake. F1000Res **10**, 33 (2021). doi:10.12688/f1000research.29032.2
26. Vieira-Silva, S., Falony, G., Darzi, Y., Lima-Mendez, G., Yunta, R.G., Okuda, S., Vandeputte, D., Valles-Colomer, M., Hildebrand, F., Chaffron, S., Raes, J.: Species–function relationships shape ecological properties of the human gut microbiome. Nature Microbiology **1**(8), 16088 (2016). doi:10.1038/nmicrobiol.2016.88
27. Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Sundaramurthi, J.C., Lee, J., Kandimalla, M., Chen, I.-M.A., Kyrpides, N.C., Reddy, T.B.K.: Genomes OnLine Database (GOLD) v.8: Overview and updates. Nucleic Acids Research **49**(D1), 723–733 (2021). doi:10.1093/nar/gkaa983
28. Johnson, J.S., Spakowicz, D.J., Hong, B.-Y., Petersen, L.M., Demkowicz, P., Chen, L., Leopold, S.R., Hanson, B.M., Agresta, H.O., Gerstein, M., Sodergren, E., Weinstock, G.M.: Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nature Communications **10**(1), 1–11 (2019). doi:10.1038/s41467-019-13036-1
29. Schoch, C.L., Ciufo, S., Domrachev, M., Hotton, C.L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J.P., Sun, L., Turner, S., Karsch-Mizrachi, I.: NCBI Taxonomy: A comprehensive update on curation, resources and tools. Database (Oxford) **2020**, 062 (2020). doi:10.1093/database/baaa062
30. Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P.: DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods **13**(7), 581–583 (2016). doi:10.1038/nmeth.3869
31. Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J.E., Bittinger, K., Brejnrod, A., Brislawn, C.J., Brown, C.T., Callahan, B.J., Caraballo-Rodríguez, A.M., Chase, J., Cope, E.K., Silva, R.D., Diener, C., Dorrestein, P.C., Douglas, G.M., Durall, D.M., Duvallet, C., Edwardson, C.F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J.M., Gibbons, S.M., Gibson, D.L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G.A., Janssen, S., Jarmusch, A.K., Jiang, L., Kaehler, B.D., Kang, K.B., Keefe, C.R., Keim, P., Kelley, S.T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M.G.I., Lee, J., Ley, R., Liu, Y.-X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B.D., McDonald, D., McIver, L.J., Melnik, A.V., Metcalf, J.L., Morgan, S.C., Morton, J.T., Naimey, A.T., Navas-Molina, J.A., Nothias, L.F., Orchanian, S.B., Pearson, T., Peoples, S.L., Petras, D., Preuss, M.L., Pruesse, E., Rasmussen, L.B., Rivers, A., Robeson, M.S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S.J., Spear, J.R., Swafford, A.D., Thompson, L.R., Torres, P.J., Trinh, P., Tripathi, A., Turnbaugh, P.J., Ul-Hasan, S., van der Hooft, J.J.J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K.C., Williamson, C.H.D., Willis, A.D., Xu, Z.Z., Zaneveld, J.R., Zhang, Y., Zhu, Q., Knight, R., Caporaso, J.G.: Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology **37**(8), 852–857 (2019). doi:10.1038/s41587-019-0209-9
32. Bokulich, N.A., Kaehler, B.D., Rideout, J.R., Dillon, M., Bolyen, E., Knight, R., Huttley, G.A., Gregory Caporaso, J.: Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome **6**(1), 90 (2018). doi:10.1186/s40168-018-0470-z

33. Kaehler, B.D., Bokulich, N.A., McDonald, D., Knight, R., Caporaso, J.G., Huttley, G.A.: Species abundance information improves sequence taxonomy classification accuracy. Nat Commun **10**(1), 4643 (2019). doi:10.1038/s41467-019-12669-6
34. Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O.: The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res **41**(D1), 590–596 (2013). doi:10.1093/nar/gks1219
35. Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D.T., Beghini, F., Malik, F., Ramos, M., Dowd, J.B., Huttenhower, C., Morgan, M., Segata, N., Waldron, L.: Accessible, curated metagenomic data through ExperimentHub. Nature Methods **14**(11), 1023–1024 (2017). doi:10.1038/nmeth.4468
36. Beghini, F., McIver, L.J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A.M., Valles-Colomer, M., Weingart, G., Zhang, Y., Zolfo, M., Huttenhower, C., Franzosa, E.A., Segata, N.: Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. eLife **10**, 65088 (2021). doi:10.7554/eLife.65088
37. Schiffer, L., Azhar, R., Shepherd, L., Ramos, M., Geistlinger, L., Huttenhower, C., Dowd, J.B., Segata, N., Waldron, L.: HMP16SData: Efficient access to the human microbiome project through bioconductor. American Journal of Epidemiology (2019). doi:10.1093/aje/kwz006
38. Consortium, T.H.M.P., Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J.H., Chinwalla, A.T., Creasy, H.H., Earl, A.M., FitzGerald, M.G., Fulton, R.S., Giglio, M.G., Hallsworth-Pepin, K., Lobos, E.A., Madupu, R., Magrini, V., Martin, J.C., Mitreva, M., Muzny, D.M., Sodergren, E.J., Versalovic, J., Wollam, A.M., Worley, K.C., Wortman, J.R., Young, S.K., Zeng, Q., Aagaard, K.M., Abolude, O.O., Allen-Vercoe, E., Alm, E.J., Alvarado, L., Andersen, G.L., Anderson, S., Appelbaum, E., Arachchi, H.M., Armitage, G., Arze, C.A., Ayvaz, T., Baker, C.C., Begg, L., Belachew, T., Bhonagiri, V., Bihan, M., Blaser, M.J., Bloom, T., Bonazzi, V., Brooks, J.P., Buck, G.A., Buhay, C.J., Busam, D.A., Campbell, J.L., Canon, S.R., Cantarel, B.L., Chain, P.S.G., Chen, I.-M.A., Chen, L., Chhibba, S., Chu, K., Ciulla, D.M., Clemente, J.C., Clifton, S.W., Conlan, S., Crabtree, J., Cutting, M.A., Davidovics, N.J., Davis, C.C., DeSantis, T.Z., Deal, C., Delehaunty, K.D., Dewhirst, F.E., Deych, E., Ding, Y., Dooling, D.J., Dugan, S.P., Dunne, W.M., Durkin, A.S., Edgar, R.C., Erlich, R.L., Farmer, C.N., Farrell, R.M., Faust, K., Feldgarden, M., Felix, V.M., Fisher, S., Fodor, A.A., Forney, L.J., Foster, L., Francesco, V.D., Friedman, J., Friedrich, D.C., Fronick, C.C., Fulton, L.L., Gao, H., Garcia, N., Giannoukos, G., Giblin, C., Giovanni, M.Y., Goldberg, J.M., Goll, J., Gonzalez, A., Griggs, A., Gujja, S., Haake, S.K., Haas, B.J., Hamilton, H.A., Harris, E.L., Hepburn, T.A., Herter, B., Hoffmann, D.E., Holder, M.E., Howarth, C., Huang, K.H., Huse, S.M., Izard, J., Jansson, J.K., Jiang, H., Jordan, C., Joshi, V., Katancik, J.A., Keitel, W.A., Kelley, S.T., Kells, C., King, N.B., Knights, D., Kong, H.H., Koren, O., Koren, S., Kota, K.C., Kovar, C.L., Kyrpides, N.C., Rosa, P.S.L., Lee, S.L., Lemon, K.P., Lennon, N., Lewis, C.M., Lewis, L., Ley, R.E., Li, K., Liolios, K., Liu, B., Liu, Y., Lo, C.-C., Lozupone, C.A., Lunsford, R.D., Madden, T., Mahurkar, A.A., Mannon, P.J., Mardis, E.R., Markowitz, V.M., Mavromatis, K., McCorrison, J.M., McDonald, D., McEwen, J., McGuire, A.L., McInnes, P., Mehta, T., Mihindukulasuriya, K.A., Miller, J.R., Minx, P.J., Newsham, I., Nusbaum, C., O'Laughlin, M., Orvis, J., Pagani, I., Palaniappan, K., Patel, S.M., Pearson, M., Peterson, J., Podar, M., Pohl, C., Pollard, K.S., Pop, M., Priest, M.E., Proctor, L.M., Qin, X., Raes, J., Ravel, J., Reid, J.G., Rho, M., Rhodes, R., Riehle, K.P., Rivera, M.C., Rodriguez-Mueller, B., Rogers, Y.-H., Ross, M.C., Russ, C., Sanka, R.K., Sankar, P., Sathirapongsasuti, J.F., Schloss, J.A., Schloss, P.D., Schmidt, T.M., Scholz, M., Schriml, L., Schubert, A.M., Segata, N., Segre, J.A., Shannon, W.D., Sharp, R.R., Sharpton, T.J., Shenoy, N., Sheth, N.U., Simone, G.A., Singh, I., Smillie, C.S., Sobel, J.D., Sommer, D.D., Spicer, P., Sutton, G.G., Sykes, S.M., Tabbaa, D.G., Thiagarajan, M., Tomlinson, C.M., Torralba, M., Treangen, T.J., Truty, R.M., Vishnivetskaya, T.A., Walker, J., Wang, L., Wang, Z., Ward, D.V., Warren, W., Watson, M.A., Wellington, C., Wetterstrand, K.A., White, J.R., Wilczek-Boney, K., Wu, Y., Wylie, K.M., Wylie, T., Yandava, C., Ye, L., Ye, Y., Yooseph, S., Youmans, B.P., Zhang, L., Zhou, Y., Zhu, Y., Zoloth, L., Zucker, J.D., Birren, B.W., Gibbs, R.A., Highlander, S.K., Methé, B.A., Nelson, K.E., Petrosino, J.F., Weinstock, G.M., Wilson, R.K., White, O.: Structure, function and diversity of the healthy human microbiome. Nature **486**(7402), 207–214 (2012). doi:10.1038/nature11234
39. Zeller, G., Tap, J., Voigt, A.Y., Sunagawa, S., Kultima, J.R., Costea, P.I., Amiot, A., Böhm, J., Brunetti, F., Habermann, N., Hercog, R., Koch, M., Luciani, A., Mende, D.R., Schneider, M.A., Schrotz-King, P., Tournigand, C., Tran Van Nhieu, J., Yamada, T., Zimmermann, J., Benes, V., Kloor, M., Ulrich, C.M., Knebel Doeberitz, M., Sobhani, I., Bork, P.: Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol **10**(11), 766 (2014). doi:10.15252/msb.20145645
40. Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., Zhang, D., Xia, H., Xu, X., Jie, Z., Su, L., Li, X., Li, X., Li, J., Xiao, L., Huber-Schönauer, U., Niederseer, D., Xu, X., Al-Aama, J.Y., Yang, H., Wang, J., Kristiansen, K., Arumugam, M., Tilg, H., Datz, C., Wang, J.: Gut microbiome development along the colorectal adenoma-carcinoma sequence. Nat Commun **6**, 6528 (2015). doi:10.1038/ncomms7528
41. Gupta, A., Dhakan, D.B., Maji, A., Saxena, R., P K, V.P., Mahajan, S., Pulikkan, J., Kurian, J., Gomez, A.M., Scaria, J., Amato, K.R., Sharma, A.K., Sharma, V.K.: Association of Flavonifractor plautii, a Flavonoid-Degrading Bacterium, with the Gut Microbiome of Colorectal Cancer Patients in India. mSystems **4**(6), 00438–19 (2019). doi:10.1128/mSystems.00438-19
42. Hannigan, G.D., Duhaime, M.B., Ruffin, M.T., Koumpouras, C.C., Schloss, P.D.: Diagnostic Potential and Interactive Dynamics of the Colorectal Cancer Virome. mBio **9**(6), 02248–18 (2018). doi:10.1128/mBio.02248-18
43. Thomas, A.M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., Beghini, F., Manara, S., Karcher, N., Pozzi, C., Gandini, S., Serrano, D., Tarallo, S., Francavilla, A., Gallo, G., Trompetto, M., Ferrero, G., Mizutani, S., Shiroma, H., Shiba, S., Shibata, T., Yachida, S., Yamada, T., Wirbel, J., Schrotz-King, P., Ulrich, C.M., Brenner, H., Arumugam, M., Bork, P., Zeller, G., Cordero, F., Dias-Neto, E., Setubal, J.C., Tett, A., Pardini, B., Rescigno, M., Waldron, L., Naccarati, A., Segata, N.: Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat Med **25**(4), 667–678 (2019). doi:10.1038/s41591-019-0405-7

44. Vogtmann, E., Hua, X., Zeller, G., Sunagawa, S., Voigt, A.Y., Hercog, R., Goedert, J.J., Shi, J., Bork, P., Sinha, R.: Colorectal Cancer and the Human Gut Microbiome: Reproducibility with Whole-Genome Shotgun Sequencing. PLoS One **11**(5), 0155362 (2016). doi:10.1371/journal.pone.0155362
45. Wirbel, J., Pyl, P.T., Kartal, E., Zych, K., Kashani, A., Milanese, A., Fleck, J.S., Voigt, A.Y., Palleja, A., Ponnudurai, R., Sunagawa, S., Coelho, L.P., Schrotz-King, P., Vogtmann, E., Habermann, N., Niméus, E., Thomas, A.M., Manghi, P., Gandini, S., Serrano, D., Mizutani, S., Shiroma, H., Shiba, S., Shibata, T., Yachida, S., Yamada, T., Waldron, L., Naccarati, A., Segata, N., Sinha, R., Ulrich, C.M., Brenner, H., Arumugam, M., Bork, P., Zeller, G.: Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat Med **25**(4), 679–689 (2019). doi:10.1038/s41591-019-0406-6
46. Yachida, S., Mizutani, S., Shiroma, H., Shiba, S., Nakajima, T., Sakamoto, T., Watanabe, H., Masuda, K., Nishimoto, Y., Kubo, M., Hosoda, F., Rokutan, H., Matsumoto, M., Takamaru, H., Yamada, M., Matsuda, T., Iwasaki, M., Yamaji, T., Yachida, T., Soga, T., Kurokawa, K., Toyoda, A., Ogura, Y., Hayashi, T., Hatakeyama, M., Nakagama, H., Saito, Y., Fukuda, S., Shibata, T., Yamada, T.: Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat Med **25**(6), 968–976 (2019). doi:10.1038/s41591-019-0458-7
47. Yu, J., Feng, Q., Wong, S.H., Zhang, D., Liang, Q.Y., Qin, Y., Tang, L., Zhao, H., Stenvang, J., Li, Y., Wang, X., Xu, X., Chen, N., Wu, W.K.K., Al-Aama, J., Nielsen, H.J., Kiilerich, P., Jensen, B.A.H., Yau, T.O., Lan, Z., Jia, H., Li, J., Xiao, L., Lam, T.Y.T., Ng, S.C., Cheng, A.S.-L., Wong, V.W.-S., Chan, F.K.L., Xu, X., Yang, H., Madsen, L., Datz, C., Tilg, H., Wang, J., Brünner, N., Kristiansen, K., Arumugam, M., Sung, J.J.-Y., Wang, J.: Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut **66**(1), 70–78 (2017). doi:10.1136/gutjnl-2015-309800
48. Proctor, L.M., Creasy, H.H., Fettweis, J.M., Lloyd-Price, J., Mahurkar, A., Zhou, W., Buck, G.A., Snyder, M.P., Strauss, J.F., Weinstock, G.M., White, O., Huttenhower, C., The Integrative HMP (iHMP) Research Network Consortium: The Integrative Human Microbiome Project. Nature **569**(7758), 641–648 (2019). doi:10.1038/s41586-019-1238-8
49. Hall, A.B., Yassour, M., Sauk, J., Garner, A., Jiang, X., Arthur, T., Lagoudas, G.K., Vatanen, T., Fornelos, N., Wilson, R., Bertha, M., Cohen, M., Garber, J., Khalili, H., Gevers, D., Ananthakrishnan, A.N., Kugathasan, S., Lander, E.S., Blainey, P., Vlamakis, H., Xavier, R.J., Huttenhower, C.: A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients. Genome Med **9**(1), 103 (2017). doi:10.1186/s13073-017-0490-5
50. Ijaz, U.Z., Quince, C., Hanske, L., Loman, N., Calus, S.T., Bertz, M., Edwards, C.A., Gaya, D.R., Hansen, R., McGrogan, P., Russell, R.K., Gerasimidis, K.: The distinct features of microbial 'dysbiosis' of Crohn's disease do not occur to the same extent in their unaffected, genetically-linked kindred. PLoS One **12**(2), 0172605 (2017). doi:10.1371/journal.pone.0172605
51. Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., Arumugam, M., Kultima, J.R., Prifti, E., Nielsen, T., Juncker, A.S., Manichanh, C., Chen, B., Zhang, W., Levenez, F., Wang, J., Xu, X., Xiao, L., Liang, S., Zhang, D., Zhang, Z., Chen, W., Zhao, H., Al-Aama, J.Y., Edris, S., Yang, H., Wang, J., Hansen, T., Nielsen, H.B., Brunak, S., Kristiansen, K., Guarner, F., Pedersen, O., Doré, J., Ehrlich, S.D., MetaHIT Consortium, Bork, P., Wang, J., MetaHIT Consortium: An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol **32**(8), 834–841 (2014). doi:10.1038/nbt.2942
52. Nielsen, H.B., Almeida, M., Juncker, A.S., Rasmussen, S., Li, J., Sunagawa, S., Plichta, D.R., Gautier, L., Pedersen, A.G., Le Chatelier, E., Pelletier, E., Bonde, I., Nielsen, T., Manichanh, C., Arumugam, M., Batto, J.-M., Quintanilha Dos Santos, M.B., Blom, N., Borruel, N., Burgdorf, K.S., Boumezbeur, F., Casellas, F., Doré, J., Dworzynski, P., Guarner, F., Hansen, T., Hildebrand, F., Kaas, R.S., Kennedy, S., Kristiansen, K., Kultima, J.R., Léonard, P., Levenez, F., Lund, O., Moumen, B., Le Paslier, D., Pons, N., Pedersen, O., Prifti, E., Qin, J., Raes, J., Sørensen, S., Tap, J., Tims, S., Ussery, D.W., Yamada, T., MetaHIT Consortium, Renault, P., Sicheritz-Ponten, T., Bork, P., Wang, J., Brunak, S., Ehrlich, S.D., MetaHIT Consortium: Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol **32**(8), 822–828 (2014). doi:10.1038/nbt.2939
53. Vich Vila, A., Imhann, F., Collij, V., Jankipersadsing, S.A., Gurry, T., Mujagic, Z., Kurilshikov, A., Bonder, M.J., Jiang, X., Tigchelaar, E.F., Dekens, J., Peters, V., Voskuil, M.D., Visschedijk, M.C., van Dullemen, H.M., Keszthelyi, D., Swertz, M.A., Franke, L., Alberts, R., Festen, E.A.M., Dijkstra, G., Masclee, A.A.M., Hofker, M.H., Xavier, R.J., Alm, E.J., Fu, J., Wijmenga, C., Jonkers, D.M.A.E., Zhernakova, A., Weersma, R.K.: Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. Sci Transl Med **10**(472), 8914 (2018). doi:10.1126/scitranslmed.aap8914
54. Breiman, L.: Random Forests. Machine Learning **45**(1), 5–32 (2001). doi:10.1023/A:1010933404324
55. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
56. Brier, G.W.: VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. Mon. Wea. Rev. **78**(1), 1–3 (1950). doi:10.1175/1520-0493(1950)078¡0001:VOFEIT¿2.0.CO;2
57. Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: ADVANCES IN LARGE MARGIN CLASSIFIERS, pp. 61–74. MIT Press, ??? (1999)
58. Ravel, J., Gajer, P., Abdo, Z., Schneider, G.M., Koenig, S.S.K., McCulle, S.L., Karlebach, S., Gorle, R., Russell, J., Tacket, C.O., Brotman, R.M., Davis, C.C., Ault, K., Peralta, L., Forney, L.J.: Vaginal microbiome of reproductive-age women. Proc. Natl. Acad. Sci. U.S.A. **108**(supplement_1), 4680–4687 (2011). doi:10.1073/pnas.1002611107
59. Velazquez, E.M., Nguyen, H., Heasley, K.T., Saechao, C.H., Gil, L.M., Rogers, A.W.L., Miller, B.M., Rolston, M.R., Lopez, C.A., Litvak, Y., Liou, M.J., Faber, F., Bronner, D.N., Tiffany, C.R., Byndloss, M.X., Byndloss, A.J., Bäumler, A.J.: Endogenous Enterobacteriaceae underlie variation in susceptibility to Salmonella infection. Nat Microbiol **4**(6), 1057–1064 (2019). doi:10.1038/s41564-019-0407-8
60. Carrow, H.C., Batachari, L.E., Chu, H.: Strain diversity in the microbiome: Lessons from Bacteroides fragilis.

PLoS Pathog **16**(12), 1009056 (2020). doi:10.1371/journal.ppat.1009056
61. Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R., Tiedje, J.M.: Ribosomal Database Project: Data and tools for high throughput rRNA analysis. Nucl. Acids Res. **42**(D1), 633–642 (2014). doi:10.1093/nar/gkt1244
62. Balvočiūtė, M., Huson, D.H.: SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? BMC Genomics **18**(S2), 114 (2017). doi:10.1186/s12864-017-3501-4
63. Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed edn. Springer Series in Statistics. Springer, New York, NY (2009)
64. Wright, M.N., Ziegler, A., König, I.R.: Do little interactions get lost in dark random forests? BMC Bioinformatics **17**(1), 145 (2016). doi:10.1186/s12859-016-0995-8
65. Cheng, Y., Ling, Z., Li, L.: The Intestinal Microbiota and Colorectal Cancer. Front. Immunol. **11** (2020). doi:10.3389/fimmu.2020.615056
66. Marquet, P., Duncan, S.H., Chassard, C., Bernalier-Donadille, A., Flint, H.J.: Lactate has the potential to promote hydrogen sulphide formation in the human colon. FEMS Microbiology Letters **299**(2), 128–134 (2009). doi:10.1111/j.1574-6968.2009.01750.x
67. Blachier, F., Mariotti, F., Huneau, J.F., Tomé, D.: Effects of amino acid-derived luminal metabolites on the colonic epithelium and physiopathological consequences. Amino Acids **33**(4), 547–562 (2007). doi:10.1007/s00726-006-0477-9
68. Clausen, M.R., Mortensen, P.B.: Fecal ammonia in patients with adenomatous polyps and cancer of the colon. Nutrition and Cancer **18**(2), 175–180 (1992). doi:10.1080/01635589209514217
69. Lin, H.-C., Visek, W.J.: Colon Mucosal Cell Damage by Ammonia in Rats. The Journal of Nutrition **121**(6), 887–893 (1991). doi:10.1093/jn/121.6.887
70. Xu, Z., Malmer, D., Langille, M.G.I., Way, S.F., Knight, R.: Which is more important for classifying microbial communities: Who's there or what they can do? ISME J **8**(12), 2357–2359 (2014). doi:10.1038/ismej.2014.157
71. Weimann, A., Mooren, K., Frank, J., Pope, P.B., Bremges, A., McHardy, A.C.: From Genomes to Phenotypes: Traitar, the Microbial Trait Analyzer. mSystems **1**(6), 00101–16 (2016). doi:10.1128/mSystems.00101-16

Additional Files

Fig S1 — Trait coverage statistics for samples profiled with 16S rRNA gene metabarcoding

Fig S2 — Top 10 important features based on random forest model fitted on different inputs from data sets profiled with 16S rRNA gene sequencing