Healthy microbiome - moving towards functional interpretation

Kinga Zielińska[1], Klas I. Udekwu[2,3],Witold Rudnicki[4,5,6], , Alina Frolova[7,8], Tomasz Kosciolek[1,9], Paweł P Łabaj[1]

1 - Małopolska Centre of Biotechnology, Jagiellonian University, Krakow, Poland

2 - Department of Biological Sciences, University of Idaho, Moscow, ID 83843, U.S.A.

3 - Swedish Environmental Epidemiology Centre, Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, SE75007, Sweden

4 - Institute of Informatics, University of Białystok, Białystok, Poland

5 - Computational Centre, University of Białystok, Białystok, Poland

6 - Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Warsaw, Poland

7 - Institute of Molecular Biology and Genetics of National Academy of Sciences of Ukraine, Kyiv Academic University, Kyiv 03143, Ukraine

8 - Kyiv Academic University, Kyiv 03142, Ukraine

9 - Department of Data Science and Engineering, Silesian University of Technology, Gliwice, Poland

Abstract

Microbiome-based disease prediction has significant potential as an early, non-invasive marker of gut dysbiosis-mediated conditions, thanks in part to decreasing sequencing and analysis costs. Existing tools, or microbiome health indexes, are based on a microbiome's species richness and dependent on taxonomic classification. However, a developing understanding of microbiome metabolic and phenotypic complexity reveals substantial restrictions of such approaches. In this study, we introduce a new health microbiome index created as an answer to updated microbiome definitions. The novelty of our approach is a shift from a Linnean approach towards microbiome function, while incorporating inter-species interactions to distinguish between healthy and diseased states. We present this significant improvement on the taxonomy-based Gut Microbiome Health Index (GMHI), the most comprehensive index to date, using data obtained from the Human Microbiome Project 2 and American Gut Project cohorts, a study comparison of healthy and Inflammatory Bowel Disease individuals. We validate its performance on a range of diseases, demonstrating its robustness to sequencing depth. Overall, we emphasize the potential and the need to shift towards functional approaches to better understand microbiome health, and provide directions for future index enhancements. Our method, **`q2-predict-dysbiosis`**, is freely available as a **`QIIME 2`** plugin (https://github.com/bioinf-mcb/q2-predict-dysbiosis).

Introduction

The global prevalence of a host of diseases peripherally or directly linked to microbiome health such as cancer, diabetes, obesity or Inflammatory Bowel Disease (IBD) is continuously on the rise, and substantial funds are spent on diagnosis and treatment (M'koma, 2013, Hong et al., 2019). A correlation of the gut microbiome composition to health is now widely acknowledged (Vijay and Valdes, 2022). Accordingly, the ability to evaluate patient health status based on a gut microbiome snapshot is of potentially high clinical value. Stool-based methods are particularly promising because they can be collected non-invasively and frequently, accompanied by a short analysis time. Decreasing costs of next-generation sequencing-based stool sample characterization makes microbiome-based characterization and diagnostics a strong competitor in the field (Bajaj et al., 2020).

Dysbiosis, defined as a reduction of gut homeostasis, is usually accompanied by the reduction of microbiota diversity and an overgrowth of harmful bacteria (DeGruttola et al., 2016, Hrncir, 2022). It can be caused by a wide range of factors including infection, diet, exercise, antibiotics, stress or poor sleep (Martinez et al., 2021). The simplest interventions applied today in order to prevent or alleviate mild microbiome dysbioses include dietary modifications (including prebiotics), lifestyle changes and probiotics. More severe cases, usually failing to respond to the above recommendations, may qualify for fecal microbiota transplants, FMTs, which are increasingly gaining traction clinically (Bull and Plummer, 2015). However, host-microbiota and intra-microbiota interactions are both very complex and highly individual. Despite FMT's being used clinically for a number of years now, scientists still have no real understanding of why or how they work.

There are a number of approaches used to evaluate gut microbiome health from stool. Alpha diversity is a frequent choice as it is closely related to dysbiosis, and microbiome richness is described as a key component of microbiome health and robustness (Li et al., 2022, Gong et al., 2016). Beta diversity has also been applied to identify non-dysbiotic samples based on their time-resolved proximity to other healthy samples in a number of longitudinal studies, albeit less commonly (Zouiouich et al., 2021). The most robust index to date, outperforming diversity indices, is the Gut Microbiome Health Index, or the **GMHI** (Gupta et al., 2020), recently renamed by the authors to Gut Microbiome Wellness Index (**GMWI**). It is based on the ratio of 50 microbial species associated with healthy or unhealthy gut ecosystems and is reported to exceed 73% accuracy in determining disease state. Based on the results, the authors suggest that gut taxonomic signatures can predict health status. Also the improved yet a pre-print version, the **GMWI2**, is purely based on gut taxonomic signatures (Chang et al., 2023). Another index expanding on the **GMHI** is the hiPCA (Zhu et al., 2023), which was introduced as a monitoring framework for personalized health status. The personalized approach is achieved by analyzing the contribution of each bacterium to the index, which allows for the identification of high-influence (ostensibly keystone) species in different patient groups. Importantly, such studies are often defined by datasets limited in scope to industrialized nations and a full consideration of diet-environment-microbiome interactions is not met.

However, a recently re-visited definition of the microbiome emphasizes the importance of not just the microbiota (a community of microorganisms), but the whole "Theatre of activity", ToA (Berg et al., 2020). This also includes structural elements (proteins, lipids, poly-saccharides), metabolites and environmental conditions. The ToA is tightly bound to its corresponding ecological niche, and it is the synergistic relations of species that provide all the necessary components of the ToA defining the whole community. Based on this definition, we maintain that an index constructed from taxonomy

alone is insufficient to accurately capture the biological phenomena occurring within the gut environment – and which are key to the understanding of dysbiotic states. Instead, we hypothesize that **to effectively determine health:** i) a functional profile is required (microbiome phenotype); and, ii) species co-occurrence and not presence should be considered (microbial interaction networks).

We introduce an approach that is based on features beyond alpha diversity and taxonomic information. We derive them based on an exploratory analysis of healthy samples from the Human Microbiome Project 2 (HMP2, Lloyd-Price et al., 2019) and two American Gut Project cohorts (AGP, McDonald et al, 2018), PRJNA389280 and PRJEB1220, further referred to as AGP_1 and AGP_2. Our goal is not only to distinguish between healthy and diseased, but also to quantify the degree of dysbiosis in each sample. We validate our approach on IBD individuals from the same cohorts, and outperform the **`GMHI`** on an unseen set of diseases that the **`GMHI`** was originally developed on. Furthermore, we demonstrate the effectiveness of our method in identifying longitudinal microbiome changes in COVID-19 patients, which the **`GMHI`** is unable to capture. Finally, we show that our approach is robust to varying sequencing depths . Our method is freely available as a **`QIIME 2`** plugin **`q2-predict-dysbiosis`** (https://github.com/bioinf-mcb/q2-predict-dysbiosis).

Results

High prevalence of “core functions” in health

In order to develop a way to assess the degree of dysbiosis in a microbiome sample, we must define microbiome healthy states. We based our initial analysis (described in the Methods section) on 384 healthy individuals from the HMP2 project and searched for the most prevalent species, regardless of abundance (Figure 1a,c). We observed that 60% of species were present in less than 10% of samples, and hardly any species were shared by all individuals. On the other hand, the prevalence of functions within the healthy population had an opposite trend - and 60% of functions were shared by over 75% of individuals (Figure 1b, d), a functional redundancy unaccounted for in the **`GMHI`**. This result convinced us further about the unsuitability of basing an index purely on the presence of “core taxa”, and encouraged a shift of focus towards more prevalent functions instead.

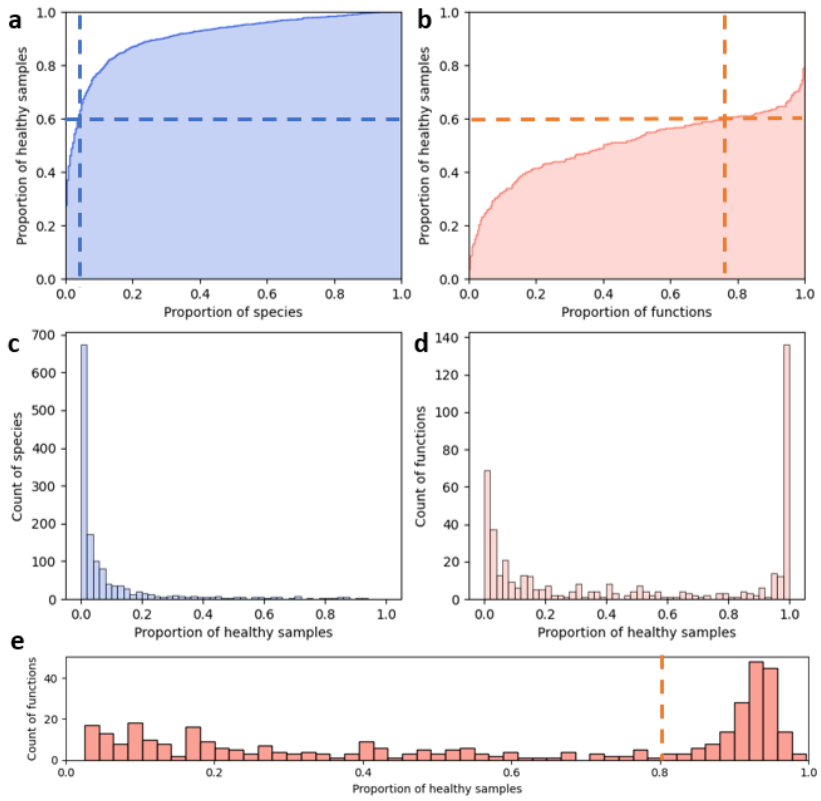

Figure 1. Empirical cumulative distribution functions of species (**a**) and functions (**b**), and matching histograms of species (**c**) and functions (**d**) present in healthy individuals from the HMP2.
2. Distribution of functions across healthy individuals from the HMP2, AGP_1 and AGP_2.

The addition of healthy samples from two AGP cohorts maintained the function distribution profile obtained purely with HMP2 samples (Figure 1e). In order to test whether the functions were universal or cohort-specific, we calculated the distributions separately for functions present in 1, 2 or all 3 cohorts. We found that all functions missing from at least 1 cohort were present in less than 10% of samples, which indicated the occurrence of high-prevalence functions in all three cohorts. Based on the increased frequency of functions present in over 80% of samples (dotted line in Figure 1e), we defined this set as “core functions” (refer to Supplementary Table 1 for full list). According to MetaCyc classification, 73.5% of “core functions” were assigned as “Biosynthesis” pathways, followed by 18.8% (“Degradation/Utilization/Assimilation”) and 7.6% (“Generation of Precursor Metabolites and Energy”), a few of these additionally classified as “Superpathways”. The classification aligned well with a previously reported high prevalence of carbohydrate and amino acid metabolism-related pathways, potentially forming the functional microbiome core (Zou et al., 2019). A detailed analysis of the core functions identified in our study, however, is out of the scope of this manuscript.

Diversity analysis revealed that the number of functions per sample remains similar, or even increases, during microbiome transitions from healthy state to dysbiosis (Supplementary Figure 1). We asked whether “core functions” are maintained or replaced by other functions in IBD samples, investigating the presence of “core functions” in different groups. While the differences were not significant in most cases, we noticed a visibly higher fraction of "core functions" among all functions (Figure 2a) and found a higher percentage of all "core functions" (Figure 2b) in healthy compared to

disease samples. We carried out differential enrichment analysis using LEfSe (Segata et al., 2011), performed separately for each cohort, and identified 100 pathways that were more abundant in healthy as compared to disease cohorts. Over 90% of functions enriched in healthy samples were "core functions", while they constituted less than 5% of functions enriched in the unhealthy class (Supplementary Figure 2).

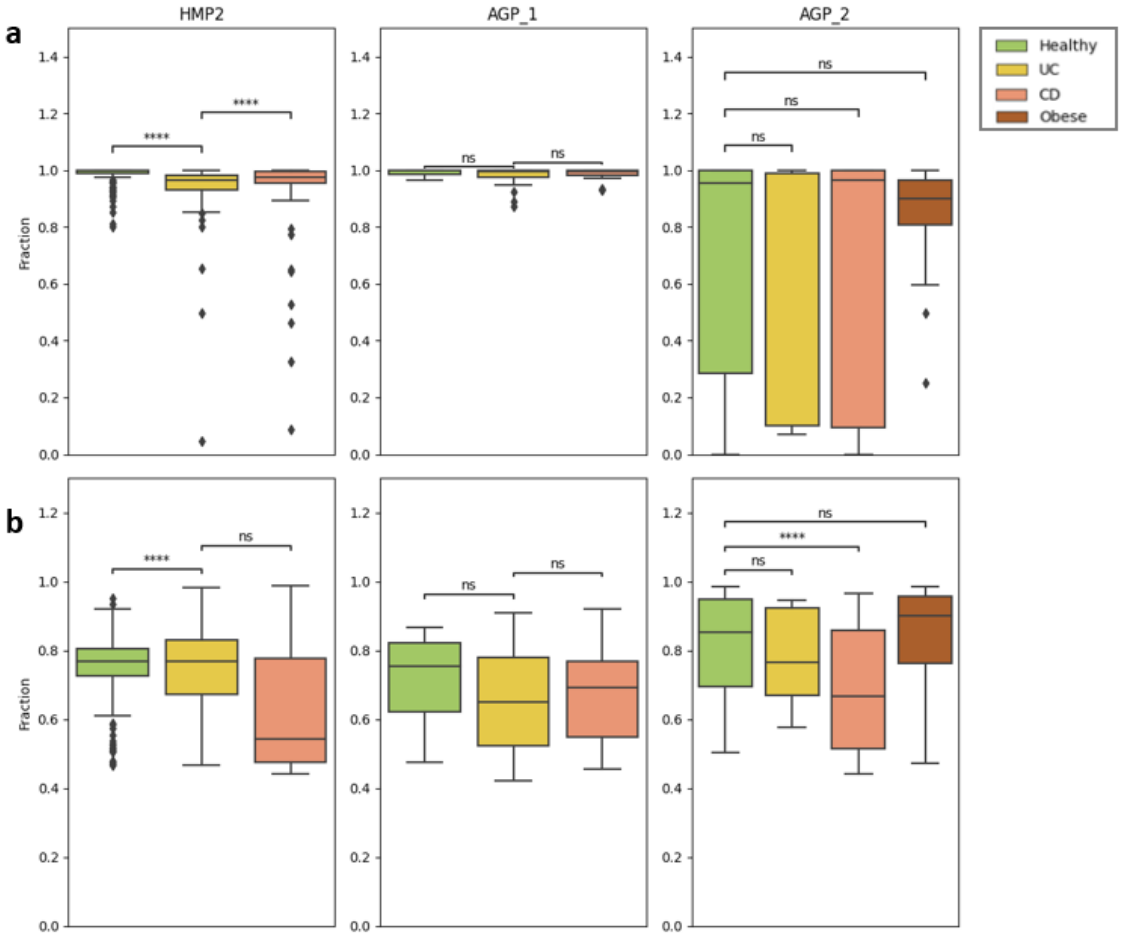

Figure 2. Fraction of "core functions" found per sample (**a**) and fraction of "core functions" among all functions per sample (**b**), grouped by project and health group.

Species interactions and function contributions in health

In accordance with past studies, we observed a decrease in the number of species in dysbiotic samples (Supplementary Figure 3, Mirsepasi-Lauridsen et al., 2018, Mosca et al., 2016). Having previously noted an increase in the number of functions (above), we speculated that the remaining species may contribute to core or new functions, forming new connections with one another. Due to the substantial number of initial connections to analyze (170 "core functions" and 1490 species present in at least two projects), we restricted the number of species to those most informative in

the context of healthy- / disease-state prediction. We chose Multi-Dimensional Feature Selection (MDFS) algorithm, as it was the only feature selection method accounting for inter-feature interactions that we were aware of at the time of the manuscript submission. This approach reduced the number of species to 587 allowing us to place focus on the most important interactions (see Methods for details about the feature selection procedure).

We then used the SparCC algorithm, designed specifically for compositional data, to investigate correlations between the resulting species in health and disease (Friedman and Alm, 2012). We did not observe any trends in the number of correlations, or in the fraction of positive correlations per sample, that would indicate differences between the two. However, we identified opposite relationships of some species in different groups (Figure 3). A number of species we found to be negatively correlated with diseased state and generally considered beneficial *(e.g. Eubacterium rectale*, *Faecalibacterium prausnitzii,* and a number of *Bacteroides* species), and those relationships would be disrupted in dysbiotic groups. We therefore added the co-occurrence of such species that were negatively correlated with the diseased state as another feature of interest potentially determining microbiome health.

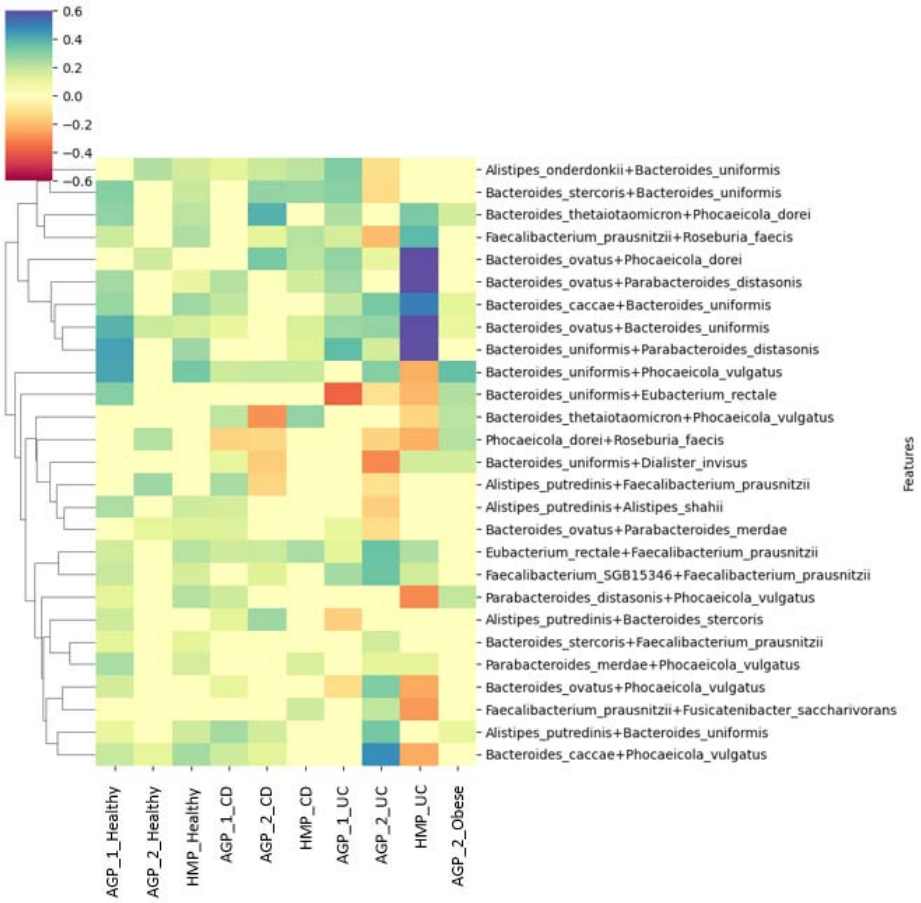

Figure 3. `SparCC` correlations between species, restricted to pairs non-negatively correlated in health.

Based on our previous results, we hypothesized that the contributions of each species to functions would be relatively stable in the healthy state and less predictable in the diseased. To test this, we compared the contributions of MDFS-identified species to "core functions" in different groups (Supplementary Figure 4). We did not observe any differences between health and disease, despite a relatively tight clustering of the healthy groups. Despite this, we included the average number of function contributions per species in our downstream analysis. If the feature later emerged as informative for disease prediction, it would be congruent with our earlier suspicions of functional plasticity; modulation of function and thus altered connectivity in the interaction network, shifting towards less abundant, non-core functions upon perturbation of homeostasis.

Testing `q2-predict-dysbiosis` accuracy of prediction for healthy and IBD individuals

Our final list of potentially health-defining microbiome features included 4 parameters: i) the fraction of "core functions" found, ii) the fraction of "core functions" among all functions, iii) the fraction of co-occurent species pairs in healthy samples, and iv) the average number of function contributions per species (details of how each feature was calculated can be found in Methods). In addition, we also included two parameters derived from the GMHI method – the number of “good” and “bad” GMHI species identified in a sample, which would enable us to compare between approaches. We then fed these described parameters into a machine learning model (See Figure 4 for the experimental design flowchart) .

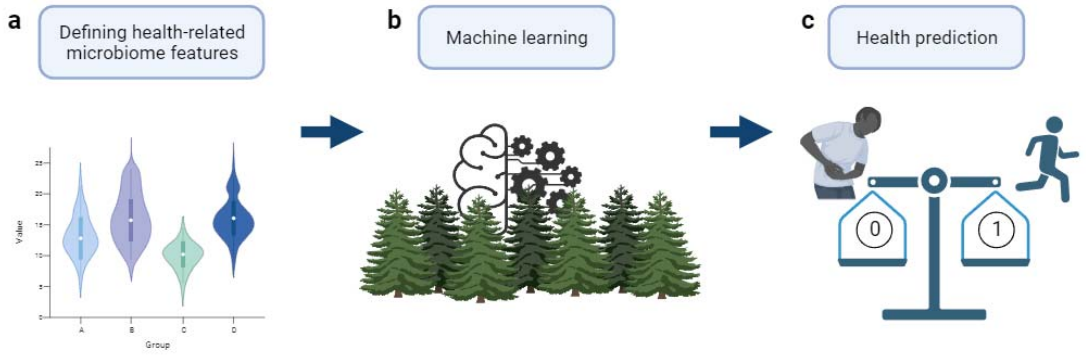

Figure 4. Schematic representation of q2-predict-dysbiosis.

We opted for a Random Forest model due to its intrinsic ability to identify informative features and performed leave-one-out cross-validation. Collecting information on feature importance at every iteration of the leave-one-out cross-validation procedure model, we consistently identified the two GMHI parameters as the least important (Supplementary Figure 5). On the other hand, the parameters defined by us merely switched positions among the top 4 places depending on the feature being excluded. We identified species co-occurrence and the fraction of core functions among all functions as the most important features and the importance of each was individually significantly greater than that of the **`GMHI`** features combined. To confirm this observation, we also applied the MDFS 1D and 2D (pure feature selection and the inter-feature interaction-aware mode), both of which identified all parameters as important (Supplementary Table 2).

Next we compared the **GMHI** values to our predictions (**q2-predict-dysbiosis**) for two IBD subtypes: ulcerative colitis (UC) and Crohn's disease (CD) from the HMP2 and the two AGP cohorts (Figure 5). In this comparison, increasingly higher scores were indicative of quantitative improvement in health status. In the case of the **GMHI**, a binary interpretation is used where any score greater than 0 is considered healthy. While both methods effectively distinguished healthy and diseased individuals from the HMP2 and AGP_1 cohorts, the **GMHI** failed to make a correct distinction between healthy, UC and Obese samples from the AGP_2. In addition, **q2-predict-dysbiosis** produced visually higher scores for all healthy than unhealthy cohorts, despite currently being designed for within-cohort analyses.

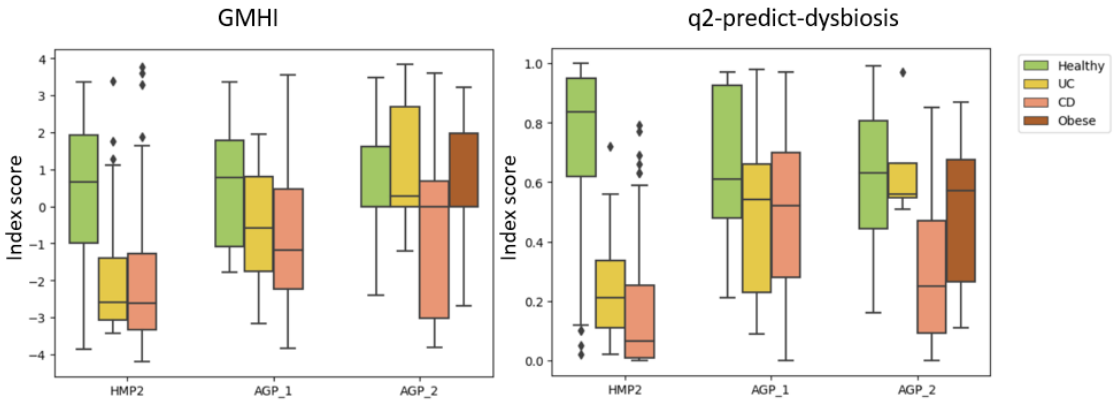

Figure 5. GMHI (a) and **q2-predict-dysbiosis (b)** scores for HMP2 and AGP healthy, IBD and Obese individuals.

In addition to the results obtained on data processed with our pipeline (see Methods), we tested the robustness of the approaches against differential sample preparation, known to impact sequencing results, and older software versions. We calculated the scores for the precalculated **Metaphlan3** and **Humann3** profiles, released with the original HMP2 publication. Both methods distinguished between dysbiotic and eubiotic (a term coined for healthy living individuals' microbiomes) samples from healthy and CD groups, even though they were less concordant for UC patients (Supplementary Figure 6). While **q2-predict-dysbiosis** did not identify significant differences between eubiosis and dysbiosis in UC individuals, it positively correlated with 37 metabolites and GMHI with 3 in this group. Among the metabolites positively correlating only with our approach was the artificial sweetener acesulfame, recently shown to induce proinflammatory cytokines and metabolic syndrome in mice (Hanawa et al., 2021). Correlations between the indexes and metabolites were similar in healthy groups, as was clustering based on index value correlations with metabolites resulting in the clustering of eubiotic groups with both the **GMHI** and the **q2-predict-dysbiosis** (Supplementary Figure 7). A complete list of correlations can be found in Supplementary Table 3.

Performance comparison between methods on select disease datasets

While only healthy and IBD individuals had been included in the development and validation of the approach so far, we investigated the applicability of the **`q2-predict-dysbiosis`** to other diseases. In order to test the performance of **`q2-predict-dysbiosis`** more extensively, we used a number of datasets as different from our previous cohorts as possible. These were datasets upon which the **`GMHI`** method had been developed and thus should perform better than our method (Figure 6). While both approaches could separate healthy samples from colon cancer (Figure 6a) or acute diarrhoea (Figure 6b) samples, the **`q2-predict-dysbiosis`** produced a better separation for both (**`GMHI`** colon cancer $p < 0.05$, **`GMHI`** acute diarrhoea $p < 0.05$, **`q2-predict-dysbiosis`** colon cancer $p < 0.001$, **`q2-predict-dysbiosis`** acute diarrhoea $p < 0.01$). The **`GMHI`** successfully distinguished between healthy and atherosclerotic cardiovascular disease, ACVD, individuals (Figure 6c, $p < 0.0001$), however, it produced negative scores for non-diseased individuals for both ACVD and metabolic syndrome (Figure 6d) cohorts - hence failing to identify healthy samples (GMHI metabolic syndrome $p > 0.05$). Interestingly, the **`GMHI`** predicted lower scores for healthy individuals than those with metabolic syndrome. While the results produced by the **`q2-predict-dysbiosis`** for ACVD and metabolic syndrome cohorts were not significant ($p > 0.05$ for both), the trend was of the expected direction.

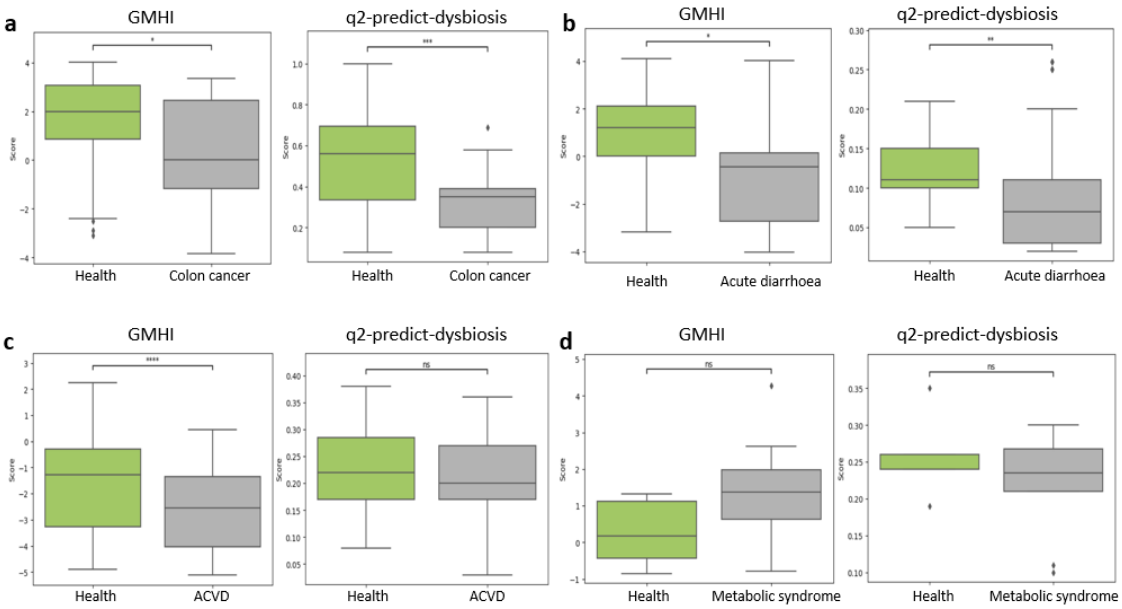

Figure 6. Comparison of the **`GMHI`** and the **`q2-predict-dysbiosis`** predictions for colon cancer (**a**), acute diarrhoea (**b**), artherosclerotic cardiovascular disease (**c**) and metabolic syndrome (**d**).

Score robustness

To evaluate the ability of both approaches to compare longitudinal and sparse data, we applied them to a shallowly sequenced COVID-19 dataset. The set consisted of three groups - controls (healthy hospital staff), and COVID-19 patients who were either transferred to the intensive care unit (ICU) or recovered (noICU). For every individual, two timepoints were selected - “1” that was

collected at hospital admission or early in the pandemic in the case of staff, and "2" which was the final sample taken from each individual. Both methods determined similar health levels for controls and noICU patients, though only the **`q2-predict-dysbiosis`** indicated a noticeable, albeit non-significant decrease in health of the control staff as the pandemic progressed (Figure 7a). More importantly, while **`q2-predict-dysbiosis`** determined a predictable decrease in the health of ICU-transferred patients, the **`GMHI`** produced the opposite trend, wrongly indicating patient recovery. In addition, the **`GMHI`** failed to identify differences between the 2_ICU and both noICU groups, whereas the separation of the two cohorts was evidently produced by the **`q2-predict-dysbiosis`**. Recalculating scores on deeply sequenced paired samples confirmed the ability of the **`q2-predict-dysbiosis`** to distinguish between healthy and ICU samples processed with different sequencing types (Figure 7b). However, in this case, the difference between samples collected upon admission and upon patient transfer to the ICU disappeared. Importantly, the **`GMHI`** produced the highest scores for the patients at the ICU and even failed to identify controls.

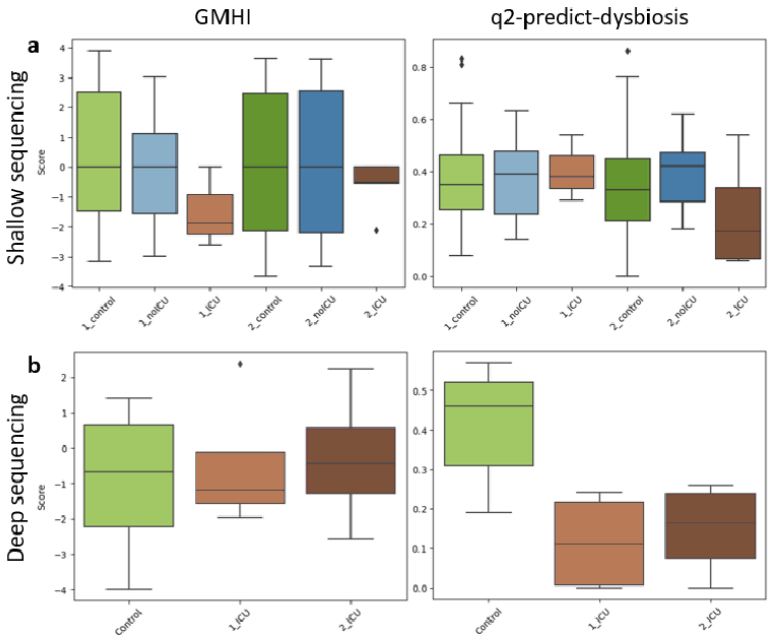

Figure 7. Comparison of the **`GMHI`** and **`q2-predict-dysbiosis`** predictions on COVID-19 samples processed using shallow (**a**) and deep (**b**) sequencing.

The separation of 1_ICU and 2_ICU patients varied dependent upon sequencing method, so we explored the robustness of each method to variation in sequencing depth. In order to investigate that, we applied various degrees of rarefaction to deeply sequenced control samples expecting to see similar importance scores despite changing sequencing depth. The scores appeared to increase with increasing degree of rarefaction, and although negligible, this change occurred similarly for both methods (Figure 8 a, b). The results corresponding to the lowest coverage considered (2,000,000 reads reported a reasonable consensus between cost and resolution (Hillmann et al., 2018)) would be visibly further from the remaining values, indicating this as a coverage threshold for method stability. Plotting values for each feature separately revealed that at this level certain

species (Figure 8c,d) or "core functions" (Figure 8 e, f) are undetectable. "The fraction of core features among other features", and, "the mean number of function contributions per species" would vary with coverage the most, which was anticipated considering their dependence on low-abundance functions (Figure 8e-h). Interestingly, the numbers of "good" and "bad" species identified at any sequencing depth would cover only 32% of the complete **GMHI** list (2 and 14, versus 7 and 43 for good and bad species, respectively). This overlap also argues against taxonomy-based indexes. Overall, we show that the **q2-predict-dysbiosis** is surprisingly robust considering the variation of its components, being nearly as stable as the **GMHI** across sequencing depths. However, consideration for depth is required when comparing health between samples with significantly different coverage and milder dysbiotic symptoms.

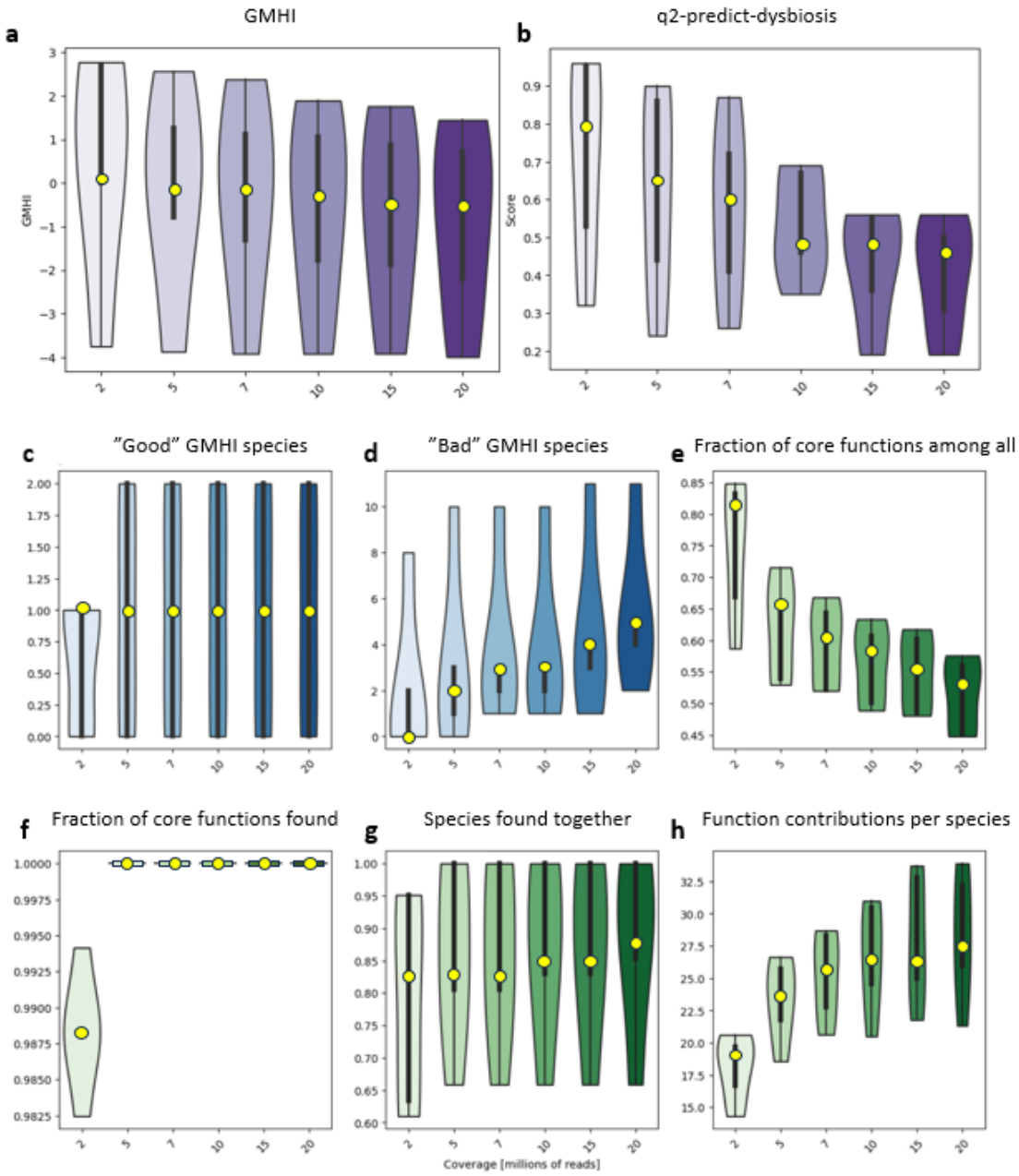

Figure 8. Robustness of health predictions to sequencing coverage. **GMHI** (**a**) and **q2-predict-dysbiosis** (**b**) scores for deeply sequenced healthy samples from the COVID-19 cohort, rarefied to corresponding depths. Taxonomic **GMHI**-inspired **q2-predict-dysbiosis** features (**c**,**d**) remain relatively stable above the coverage of 5 million reads. Functional and species interaction-related features are much more sensitive (**e**-**h**).

Discussion

A connection between the human gut microbiome and gut health is now well established, and a number of approaches have been employed to try to identify gut dysbiosis from sequence-based analysis of stool samples. Those methods are based on taxonomy and rely either on measures of microbiome richness (alpha / beta diversity) or on the presence or absence of "good" and "bad" bacteria, proposing the health-indexes **GMHI** (Gupta et al., 2020) and **hiPCA** (Zhu et al., 2023). Hampering such approaches, however, are ecological considerations of redundancy in complex environments. Currently, re-defining the microbiome to include functionality bolstered by recent findings highlights the importance of interactions between microbiome components and functional aspects thereof (Berg et al., 2020). To address the inadequacy of this Linnéan approach to evaluating change in microbiomes, we developed a novel method that incorporates function in bioinformatics-based assessment of microbiome dysbioses. We show that features based on microbiome functions and interactions define a healthy microbiome more accurately. There exists a set of "core functions", which are consistently identified as present in healthy gut microbiomes, and which disappear in the advent of dysbiosis. We compare our results to the **GMHI**, rather than to the newer **GMWI2**, to benchmark our work against a version that is extensively validated by a wide range of users. By applying feature selection, we demonstrate that all parameters defined by us are relevant for healthy versus non-healthy sample classification. As a result, the **q2-predict-dysbiosis** not only outperforms the **GMHI** in a range of cases, including datasets that the **GMHI** had been developed on, but is also nearly as robust to sequencing depth despite having coverage-sensitive features (such as "function contribution per species", dependent both on the number of species and functions identified).

There remain challenges that need to be addressed in subsequent developments of the method. We propose that a deeper understanding of how the identified function- and interaction-based microbiome features respond to variations in sequencing depth and quality is needed. We found longitudinal data to be of particular interest as it could provide more insight into personalized "core features" of a microbiome, and hence indicate individual deviations from the normal. This would provide guidelines for index rescaling and unification across cohorts, ultimately allowing for inter-study analysis.

Currently the only widely accepted functional annotation software is HUMAnN (any version), which requires data to be processed in a particular way. In future work we would like to explore other possibilities and ultimately migrate towards a more universal solution, such as gene content. Only then can recent developments in augmented functional annotation be efficiently utilized, as we already have shown in other applications (Maranga et al., 2023, Koehler Leman et al., 2023, Szydlowski et al., 2023).

We also consider expanding to a multi-omics approach in our predictions. With decreasing costs of metagenomics, metabolomics and proteomics data acquisition, predicting patient health status by

the incorporation of all three -omics methods might be a reasonable option that carries significant accuracy improvements in the not-too-distant future.

Finally, we would highlight the high levels of accuracy obtained by **`q2-predict-dysbiosis`**, based only on a minimal number of parameters. . Motivated by this success we are convinced that our methodology and approach are accurately directed towards a better understanding of the human gut microbiome and its health, based on fundamental ecological principles defining microbial community structure.

Methods

Metagenomic Data Processing

The overview of datasets used in this study can be found in Supplementary Table 4. All raw shotgun sequencing fastq files were processed with **`Trim Galore 0.6.10`** (Krueger, 2023) to ensure sufficient read quality. The taxonomic profiles were calculated with **`MetaPhlAn 4.0.6`** (Blanco-Míguez et al., 2023) and functional profiles with **`HUMAnN 3.7`** (Beghini et al., 2021). Species selection was performed using MultiDimentional Feature Selection, or the **`MDFS`** (Piliszek et al., 2018, version 1.5.2), with a Benjamini-Hochberg p-value correction in a 2D mode. Selection was performed per cohort (healthy versus each of the non-healthy groups in separate **`MDFS`** runs), and the final list of species was the union of the MDFS results for all cohorts (corrected p-values < 0.05). Rarefaction of deep COVID-19 samples, originally published by us before (Kopera et al., 2023), was performed with **`seqtk 1.4`** (Li, 2023).

Random Forest training

Metagenomic features eventually passed to the random forest were constructed based on healthy samples from the HMP2 and two AGP cohorts (AGP_1 and AGP_2). The features were as follows:

- the fraction of "core functions" found (number of “core functions” in a sample / number of all “core functions” in our list)
- the fraction of "core functions" among all functions (number of “core functions” in a sample / number of all functions in a sample)
- the fraction of species pairs commonly occurring together in healthy samples (SparCC correlation of species > 0.1, only nonnegative correlations in healthy cohorts)
- the average number of function contributions per species (number of all species to functions contributions based on stratified output / number of all species in a sample)
- the numbers of “good” and “bad” GMHI species.

During random forest training, healthy individuals were labeled as “1” and the remaining IBD + obese individuals from the same projects were added as disease samples and labeled as “0”. A leave-one-out cross-validation (613 iterations) procedure was applied to predict health scores for each sample in order to avoid model overfitting or the need to split the data into training and testing sets.

The final health score is the output of the random forest's *predict_proba* method, which expresses the probability of each sample belonging to the healthy (or unhealthy) group.

The model was trained and validated on healthy versus IBD / Overweight cohorts from the HMP and AGP. It was then tested on cohorts with other diseases.

Downstream analyses

Metagenomic table formatting and alpha diversity plots were done using **`QIIME 2`** (Bolyen et al., 2019). The GMHI scores were calculated with the **`QIIME 2 q2-health-index`** plugin ("q2-health-index," 2023) and **`SparCC`** correlations with the **`SCNIC`** plugin (https://github.com/lozuponelab/q2-SCNIC). Plots were made using custom python scripts. Any statistical tests were calculated with the independent t-test, unless stated otherwise. We initially intended on comparing our method also to the **`hiPCA`** (Zhu et al., 2023), however, we were not able to reproduce the code.

Data and code availability

The HMP2 data can be found in SRA with the accession code PRJNA398089. The two AGP projects can be located under accessions PRJNA389280 and PRJEB1220. The four cohorts from the curatedMetagenomicsData were WirbelJ_2018 (colon cancer), DavidLA_2015 (acute diarrhoea), JieZ_2017 (ACVD) and LiSS_2016 (metabolic syndrome). The COVID-19 cohort can be found under the accession PRJEB64515 in the European Nucleotide Archive. Our method is deposited on GitHub and can be accessed here: https://github.com/bioinf-mcb/q2-predict-dysbiosis.

Acknowledgements

This research was conducted as a part of the NCN Sonata BIS grant number 2020/38/E/NZ2/00598. We gratefully acknowledge Poland's high-performance Infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH, PCSS, CI TASK, WCSS) for providing computer facilities. We would like to thank Dr Krzysztof Mnich, Dr Sajad Shahbazi, Dr Balakrishnan Subramanian and Piotr Stomma from the University of Białystok for their expertise in the Multi-Dimensional Feature Selection algorithm and insightful contributions to our work, and members of Bioinformatics Research Group at MCB JU for their comments and support.

Ethics declarations

All authors declare no competing interests.

Bibliography

American Gut: an Open Platform for Citizen Science Microbiome Research | mSystems [WWW Document], n.d. URL https://journals.asm.org/doi/10.1128/msystems.00031-18 (accessed 12.3.23).

Bajaj, J.S., Acharya, C., Sikaroodi, M., Gillevet, P.M., Thacker, L.R., 2020. Cost-effectiveness of integrating gut microbiota analysis into hospitalisation prediction in cirrhosis. GastroHep 2, 79–86. https://doi.org/10.1002/ygh2.390

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., 2021. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. eLife 10, e65088. https://doi.org/10.7554/eLife.65088

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., Souza, R.S.C. de, van Overbeek, L., Singh, B.K., Wagner, M., Walsh, A., Sessitsch, A., Schloter, M., 2020. Microbiome definition re-visited: old concepts and new challenges. Microbiome 8, 103. https://doi.org/10.1186/s40168-020-00875-0

Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L.J., Thompson, K.N., Zolfo, M., Manghi, P., Dubois, L., Huang, K.D., Thomas, A.M., Nickols, W.A., Piccinno, G., Piperni, E., Punčochář, M., Valles-Colomer, M., Tett, A., Giordano, F., Davies, R., Wolf, J., Berry, S.E., Spector, T.D., Franzosa, E.A., Pasolli, E., Asnicar, F., Huttenhower, C., Segata, N., 2023. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. Nat. Biotechnol. 41, 1633–1644. https://doi.org/10.1038/s41587-023-01688-w

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., Da Silva, R., 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., vander 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., 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. 37, 852–857. https://doi.org/10.1038/s41587-019-0209-9

Bull, M.J., Plummer, N.T., 2015. Part 2: Treatments for Chronic Gastrointestinal Disease and Gut Dysbiosis. Integr. Med. Clin. J. 14, 25–33.

DeGruttola, A.K., Low, D., Mizoguchi, A., Mizoguchi, E., 2016. Current understanding of dysbiosis in disease in human and animal models. Inflamm. Bowel Dis. 22, 1137–1150. https://doi.org/10.1097/MIB.0000000000000750

Friedman, J., Alm, E.J., 2012. Inferring Correlation Networks from Genomic Survey Data. PLOS Comput. Biol. 8, e1002687. https://doi.org/10.1371/journal.pcbi.1002687

Gong, D., Gong, X., Wang, L., Yu, X., Dong, Q., 2016. Involvement of Reduced Microbial Diversity in Inflammatory Bowel Disease. Gastroenterol. Res. Pract. 2016, 6951091. https://doi.org/10.1155/2016/6951091

Gupta, V.K., Kim, M., Bakshi, U., Cunningham, K.Y., Davis, J.M., Lazaridis, K.N., Nelson, H., Chia, N., Sung, J., 2020. A predictive index for health status using species-level gut microbiome profiling. Nat. Commun. 11, 4635. https://doi.org/10.1038/s41467-020-18476-8

Gut Microbiome Wellness Index 2 for Enhanced Health Status Prediction from Gut Microbiome Taxonomic Profiles | bioRxiv [WWW Document], n.d. URL https://www.biorxiv.org/content/10.1101/2023.09.30.560294v1 (accessed 12.3.23).

Hanawa, Y., Higashiyama, M., Kurihara, C., Tanemoto, R., Ito, S., Mizoguchi, A., Nishii, S., Wada, A., Inaba, K., Sugihara, N., Horiuchi, K., Okada, Y., Narimatsu, K., Komoto, S., Tomita, K., Hokari, R., 2021. Acesulfame potassium induces dysbiosis and intestinal injury with enhanced lymphocyte migration to intestinal mucosa. J. Gastroenterol. Hepatol. 36, 3140–3148. https://doi.org/10.1111/jgh.15654

Hillmann, B., Al-Ghalith, G.A., Shields-Cutler, R.R., Zhu, Q., Gohl, D.M., Beckman, K.B., Knight, R., Knights, D., 2018. Evaluating the Information Content of Shallow Shotgun Metagenomics. mSystems 3, 10.1128/msystems.00069-18. https://doi.org/10.1128/msystems.00069-18

Hong, Y.-R., Huo, J., Desai, R., Cardel, M., Deshmukh, A.A., 2019. Excess Costs and Economic Burden of Obesity-Related Cancers in the United States. Value Health J. Int. Soc. Pharmacoeconomics Outcomes Res. 22, 1378–1386. https://doi.org/10.1016/j.jval.2019.07.004

Hrncir, T., 2022. Gut Microbiota Dysbiosis: Triggers, Consequences, Diagnostic and Therapeutic Options. Microorganisms 10, 578. https://doi.org/10.3390/microorganisms10030578

Koehler Leman, J., Szczerbiak, P., Renfrew, P.D., Gligorijevic, V., Berenberg, D., Vatanen, T., Taylor, B.C., Chandler, C., Janssen, S., Pataki, A., Carriero, N., Fisk, I., Xavier, R.J., Knight, R., Bonneau, R., Kosciolek, T., 2023. Sequence-structure-function relationships in the microbial protein universe. Nat. Commun. 14, 2351. https://doi.org/10.1038/s41467-023-37896-w

Kopera, K., Gromowski, T., Wydmański, W., Skonieczna-Żydecka, K., Muszyńska, A., Zielińska, K., Wierzbicka-Woś, A., Kaczmarczyk, M., Kadaj-Lipka, R., Cembrowska-Lech, D., Januszkiewicz, K., Kotfis, K., Witkiewicz, W., Nalewajska, M., Feret, W., Marlicz, W., Łoniewski, I., Łabaj, P.P., Rydzewska, G., Kosciolek, T., 2023. Gut Microbiome Dynamics and Predictive Value in Hospitalized COVID-19 Patients: A Comparative Analysis of Shallow and Deep Shotgun Sequencing. https://doi.org/10.1101/2023.11.29.568526

Krueger, F., 2023. Trim Galore.

Li, H., 2023. lh3/seqtk.

Li, Z., Zhou, J., Liang, H., Ye, L., Lan, L., Lu, F., Wang, Q., Lei, T., Yang, X., Cui, P., Huang, J., 2022. Differences in Alpha Diversity of Gut Microbiota in Neurological Diseases. Front. Neurosci. 16.

Lloyd-Price, J., Arze, C., Ananthakrishnan, A.N., Schirmer, M., Avila-Pacheco, J., Poon, T.W., Andrews, E., Ajami, N.J., Bonham, K.S., Brislawn, C.J., Casero, D., Courtney, H., Gonzalez, A., Graeber, T.G., Hall, A.B., Lake, K., Landers, C.J., Mallick, H., Plichta, D.R., Prasad, M., Rahnavard, G., Sauk, J., Shungin, D., Vázquez-Baeza, Y., White, R.A., Braun, J., Denson, L.A., Jansson, J.K., Knight, R., Kugathasan, S., McGovern, D.P.B., Petrosino, J.F., Stappenbeck, T.S., Winter, H.S., Clish, C.B., Franzosa, E.A., Vlamakis, H., Xavier, R.J., Huttenhower, C., 2019. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature 569, 655–662. https://doi.org/10.1038/s41586-019-1237-9

Maranga, M., Szczerbiak, P., Bezshapkin, V., Gligorijevic, V., Chandler, C., Bonneau, R., Xavier, R.J., Vatanen, T., Kosciolek, T., 2023. Comprehensive Functional Annotation of Metagenomes and Microbial Genomes Using a Deep Learning-Based Method. mSystems 8, e01178-22.

https://doi.org/10.1128/msystems.01178-22
Martinez, J.E., Kahana, D.D., Ghuman, S., Wilson, H.P., Wilson, J., Kim, S.C.J., Lagishetty, V., Jacobs, J.P., Sinha-Hikim, A.P., Friedman, T.C., 2021. Unhealthy Lifestyle and Gut Dysbiosis: A Better Understanding of the Effects of Poor Diet and Nicotine on the Intestinal Microbiome. Front. Endocrinol. 12.
Mirsepasi-Lauridsen, H.C., Vrankx, K., Engberg, J., Friis-Møller, A., Brynskov, J., Nordgaard-Lassen, I., Petersen, A.M., Krogfelt, K.A., 2018. Disease-Specific Enteric Microbiome Dysbiosis in Inflammatory Bowel Disease. Front. Med. 5.
M'koma, A.E., 2013. Inflammatory Bowel Disease: An Expanding Global Health Problem. Clin. Med. Insights Gastroenterol. 6, CGast.S12731. https://doi.org/10.4137/CGast.S12731
Mosca, A., Leclerc, M., Hugot, J.P., 2016. Gut Microbiota Diversity and Human Diseases: Should We Reintroduce Key Predators in Our Ecosystem? Front. Microbiol. 7.
Piliszek, R., Mnich, K., Migacz, S., Tabaszewski, P., Sułecki, A., Polewko-Klim, A., Rudnicki, W., 2018. MDFS - MultiDimensional Feature Selection.
q2-health-index, 2023.
Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., Huttenhower, C., 2011. Metagenomic biomarker discovery and explanation. Genome Biol. 12, R60. https://doi.org/10.1186/gb-2011-12-6-r60
Szydlowski, L.M., Bulbul, A.A., Simpson, A.C., Kaya, D.E., Singh, N.K., Sezerman, U.O., Łabaj, P.P., Kosciolek, T., Venkateswaran, K.J., 2023. Adaptation to space conditions of novel bacterial species isolated from the International Space Station revealed by functional gene annotations and comparative genome analysis. https://doi.org/10.1101/2023.09.28.559980
Vijay, A., Valdes, A.M., 2022. Role of the gut microbiome in chronic diseases: a narrative review. Eur. J. Clin. Nutr. 76, 489–501. https://doi.org/10.1038/s41430-021-00991-6
Zhu, J., Xie, H., Yang, Z., Chen, J., Yin, J., Tian, P., Wang, H., Zhao, J., Zhang, H., Lu, W., Chen, W., 2023. Statistical modeling of gut microbiota for personalized health status monitoring. Microbiome 11, 184. https://doi.org/10.1186/s40168-023-01614-x
Zou, Y., Xue, W., Luo, G., Deng, Z., Qin, P., Guo, R., Sun, H., Xia, Y., Liang, S., Dai, Y., Wan, D., Jiang, R., Su, L., Feng, Q., Jie, Z., Guo, T., Xia, Z., Liu, C., Yu, J., Lin, Y., Tang, S., Huo, G., Xu, X., Hou, Y., Liu, X., Wang, J., Yang, H., Kristiansen, K., Li, J., Jia, H., Xiao, L., 2019. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. Nat. Biotechnol. 37, 179–185. https://doi.org/10.1038/s41587-018-0008-8
Zouiouich, S., Loftfield, E., Huybrechts, I., Viallon, V., Louca, P., Vogtmann, E., Wells, P.M., Steves, C.J., Herzig, K.-H., Menni, C., Jarvelin, M.-R., Sinha, R., Gunter, M.J., 2021. Markers of metabolic health and gut microbiome diversity: findings from two population-based cohort studies. Diabetologia 64, 1749–1759. https://doi.org/10.1007/s00125-021-05464-w