Healthy microbiome - moving towards functional interpretation

Kinga Zielińska[1], Witold Rudnicki[2,3,4], Tomasz Kosciolek[1,5], Paweł P Łabaj[1]

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

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

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

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

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

Abstract

Microbiome-based disease detection approaches carry a potential for becoming practicable diagnostic options, thanks to decreasing costs of stool sample characterization. Existing tools, or microbiome health indexes, are based on microbiome richness and taxonomic information. However, a developing understanding of the microbiome 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 towards healthy cohort-defined microbiome function, as well as microbial community-driven inter-species interactions. We present the superiority of our approach over the taxonomy-based Gut Microbiome Health Index (GMHI), the most comprehensive index to date, in discrimination between healthy and Inflammatory Bowel Disease individuals from the Human Microbiome Project 2 and American Gut Project cohorts. We validate its performance further on a range of diseases, and demonstrate its robustness to sequencing depth. Overall, we emphasize the potential and the need to shift towards functions to understand microbiome health better, and provide directions for future index enhancements.

Introduction

The global prevalence of diseases 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). The connection of the gut microbiome to health is now universally acknowledged (Vijay and Valdes, 2022). Therefore, the ability to evaluate one's health status based on a gut microbiome snapshot is of high clinical value. Stool-based methods are particularly promising because they can be collected noninvasively and frequently. Decreasing costs of next-generation sequencing-based stool sample characterization is making microbiome-based characterization and diagnostics a strong competitor in the field (Bajaj et al., 2020).

Dysbiosis, or a loss of homeostatic balance, 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 or poor sleep

(Martinez et al., 2021). The simplest interventions applied today in order to prevent or heal mild microbiome dysbiosis include dietary modifications, probiotics and lifestyle changes. More severe cases, usually failing to respond to the simple treatments, may be qualified for fecal microbiota transplants (Bull and Plummer, 2015). However, host-microbiota and intra-microbiota interactions are very complex and highly individual. Despite the interventions being used for a number of years now, the understanding of why or how they work (or fail) is still not reached.

There are a number of approaches available to evaluate gut microbiome health from stool. Alpha diversity is a frequent choice as it is closely related to dysbiosis, and microbiome richness is a key component required for microbiome stability (Li et al., 2022, Gong et al., 2016). Beta diversity has also been applied to identify non-dysbiotic samples based on their proximity to other healthy samples in a number of studies, however, 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 50 microbial species associated with healthy gut ecosystems and is reported to exceed 73% accuracy in determining disease presence. Based on the results, the authors suggest that gut taxonomic signatures can predict health status. Also the improved version, the **GMWI2**, is purely based on gut taxonomic signatures (Chang et al., 2023). Another index expanding on the **GMHI** is the (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 species in different patient groups.

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 includes also 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 to secure the existence of the whole community. Based on this definition, we believe 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 its normal and abnormal behaviour. Instead, we hypothesize that **to effectively determine health:** i) a functional profile is required (how the microbiome works); ii) species co-occurrence rather than presence should be considered (how its components interact).

In this study, 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. Despite developing our ideas purely on healthy samples and testing them against IBD individuals from the same cohorts, we outperform the `GMHI` when testing our previously trained model on datasets with other 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 firstly needed to learn more about microbiome health. We based our initial analysis (described in the Methods section) on 384 healthy individuals from the HMP2 project and searched for the most prevalent, regardless of abundance, species (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). This result convinced us further about the unsuitability of basing an index purely on the presence of “core taxa”, and encouraged the shift of focus towards more prevalent functions instead.

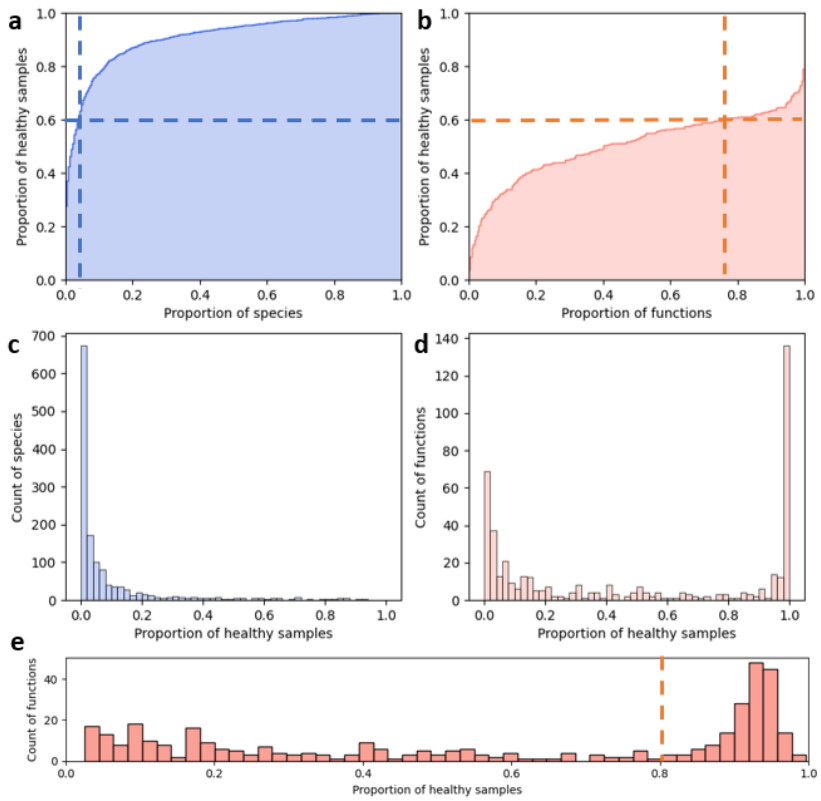

Figure 1. Empirical cumulative distribution functions of species (**a**) and functions (**b**), and matching histograms of species (**c**) and functions (**d**) present in healthy individuals from the HMP2.
2. Distribution of functions across healthy individuals from the HMP2, AGP_1 and AGP_2.

An 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 from the other cohorts, what indicated the occurrence of high-prevalence functions in all three cohorts. Based on the increased frequency of functions present in over 80% of samples (dashed line in Figure 1e), we defined this set as important and further referred to as “core

functions” (the full list of "core functions" can be found in Supplementary Table 1). According to MetaCyc classification, 73.5% of the “core functions” were assigned to the “Biosynthesis” class, followed by 18.8% in “Degradation/Utilization/Assimilation” and 7.6% in “Generation of Precursor Metabolites and Energy”, with a few 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, as a microbiome transitions to dysbiosis (Supplementary Figure 1). We wondered whether “core functions” are maintained or replaced by other functions in IBD samples, therefore, we investigated 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 a higher percentage of all "core functions" found (Figure 2b) in healthy than in some disease samples. Differential enrichment analysis with LEfSe, performed separately per cohort, identified 77 pathways more abundant in healthy individuals in both HMP2 and one AGP cohort (only 7 pathways passed significance thresholds for AGP_1 and none of them overlapped between all projects). Over 90% of functions enriched in healthy samples were "core functions", while they constituted less than 5% of functions enriched in unhealthy individuals (Supplementary Figure 2).

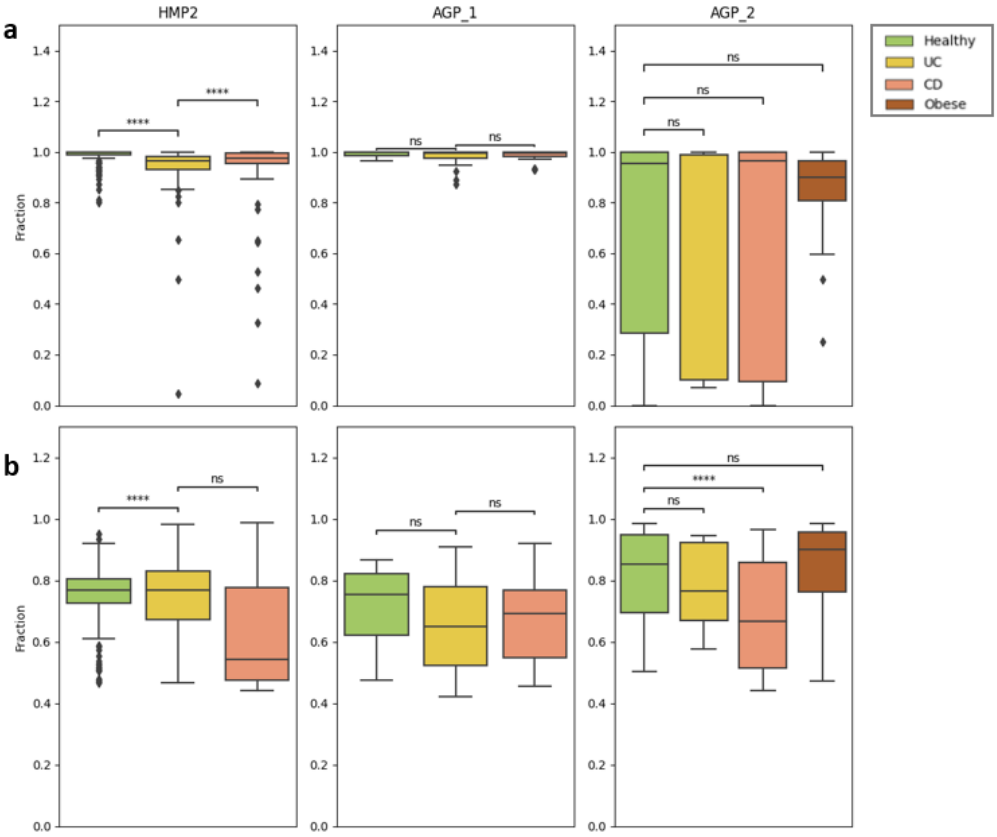

Figure 2. Fraction of "core functions" found per sample (**a**) and fraction of "core functions" among all functions per sample (**b**), grouped by project and health group.

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). The fact that we also noted an aforementioned increase in the number of functions made us wonder about the possibility of the remaining species contributing to core or new functions and forming new connections with one another. Due to a substantial number of initial connections to investigate (170 "core functions" and 1490 species present in at least two projects), we decided to restrict the number of species to those most informative in the context of healthy/disease prediction. We chose MultiDimensional 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 (Piliszek et al., 2018). The approach reduced the number of species to 587 and allowed us for a more targeted focus on the most important interactions (see Methods for details about the feature selection procedure).

We further used the **`SparCC`** algorithm to investigate correlations between the resulting species in health and disease. 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 non-negatively correlated in healthy individuals are generally considered beneficial *(such as Eubacterium rectale*, *Faecalibacterium prausnitzii,* and a number of *Bacteroides* species), and those relationships would be distorted in dysbiotic groups. We therefore added the co-occurrence of species non-negatively correlated in health as another feature of interest potentially determining microbiome health.

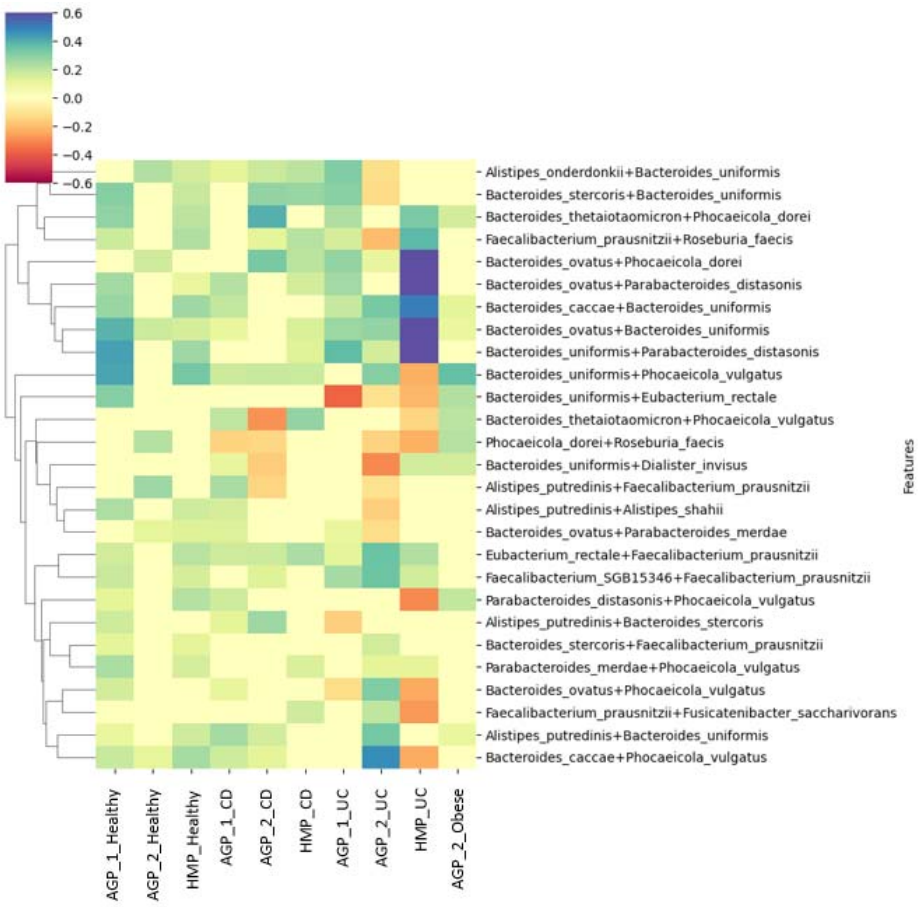

Figure 3. **SparCC** correlations between species, restricted to pairs nonnegatively correlated in health.

Based on our previous results, we hypothesized that the contributions of species to functions would be relatively stable in health and become less predictable in disease. In order to test that, we compared the connections of **MDFS** species to "core functions" in different groups (Supplementary Figure 4). We did not observe any visible differences between health and disease, despite healthy groups clustering together. However, we still decided to include information about the average number of function contributions per species. If this feature turned out to be informative in the context of disease prediction, this would align well with our observations that species might have a tendency to shift their activity towards less abundant non-core functions in the event of lost homeostasis.

Superiority of our method in separation of healthy and IBD individuals

Our final list of potentially health-defining microbiome features included 4 positions: the fraction of "core functions" found, the fraction of "core functions" among all functions, the fraction of species pairs commonly occurring together in healthy samples and the average number of function contributions per species (details of how each feature was calculated can be found in Methods). We also decided to add two parameters based on the **GMHI** – the number of “good” and “bad” **GMHI**

species identified in a sample in order to compare their importance to the parameters defined by us. Those were then fed to a machine learning model - an overview of our complete approach is shown in Figure 4.

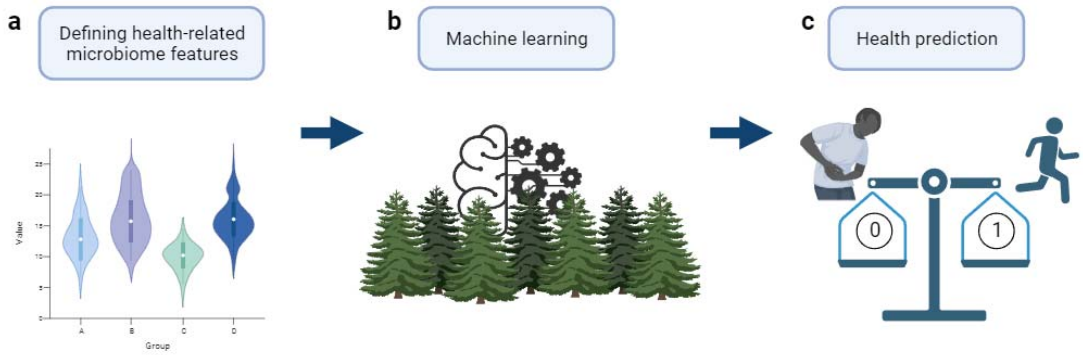

Figure 4. Schematic representation of q2-predict-dysbiosis.

We chose the Random Forest model due to its ability to identify informative features and performed leave-one-out cross-validation. Feature importance information was collected at every iteration of the leave-one-out cross-validation procedure and consistently identified the two **`GMHI`** parameters as the least important (Supplementary Figure 5). On the other hand, the parameters defined by us would switch their positions within the top 4 places. The most important features were species co-occurrence and the fraction of core functions among all functions, with the importance of any of them being much greater than the importance of the **`GMHI`** features combined. We also applied the **`MDFS 1D`** and **`2D`** (pure feature selection and the inter-feature interaction-aware mode) as an additional way to estimate parameter importance. Both **`MDFS`** modes identified all parameters as relevant (Supplementary Table 2).

We compared the **`GMHI`** values to our predictions (**`q2-predict-dysbiosis`**) on the HMP2 and the two AGP cohorts (Figure 5). Higher scores indicate better health (in the case of the **`GMHI`**, 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.

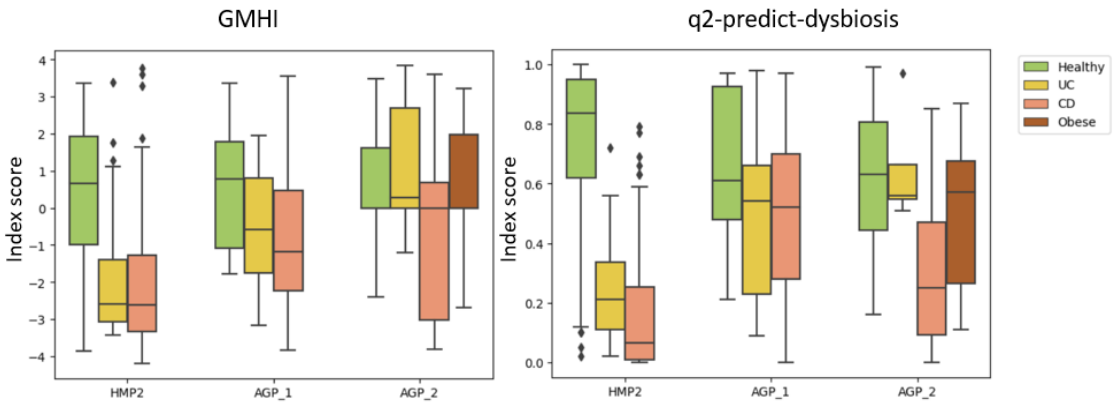

Figure 5. `GMHI` (a) and `q2-predict-dysbiosis` (b) scores for HMP2 and AGP healthy, IBD and obese individuals.

In addition to the results obtained on data processed with our pipeline (see Methods), we tested the robustness of the approaches against different sample preparation and older software versions. We calculated the scores for the precalculated Metaphlan3 and Humann3 profiles, supplied with the original HMP2 publication. Both methods distinguished between dysbiotic and eubiotic individuals from healthy and CD groups, though they were less concordant in the case of 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 metabolites positively correlating only with our approach was acesulfame, known to have implications in dysbiosis and intestinal injury (Hanawa et al., 2021). Correlations between the indexes and metabolites were similar in healthy groups, and so was clustering based on index value correlations with metabolites that resulted in eubiotic groups clustering together for both the **`GMHI`** and the **`q2-predict-dysbiosis`** (Supplementary Figure 7). A complete list of correlations can be found in Supplementary Table 3.

Superiority / similar performance to GMHI on datasets with other diseases

While only healthy and IBD individuals had been included in the development and validation of the approach so far, we wondered about the applicability of the **`q2-predict-dysbiosis`** to other diseases. In order to test that, we selected a number of datasets as different from our previous cohorts as possible and which the **`GMHI`** had been developed on and compared its performance to the **`q2-predict-dysbiosis`** (Figure 6). While both approaches separated healthy from colon cancer (Figure 6a) or acute diarrhoea (Figure 6b) individuals, the **`q2-predict-dysbiosis`** produced a better separation for both. The **`GMHI`** successfully distinguished between healthy and ACVD individuals (Figure 6c), however, it produced negative scores for non-diseased individuals for both ACVD and metabolic syndrome (Figure 6d) cohorts - hence failed to identify healthy samples. 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, the trend was of the expected direction.

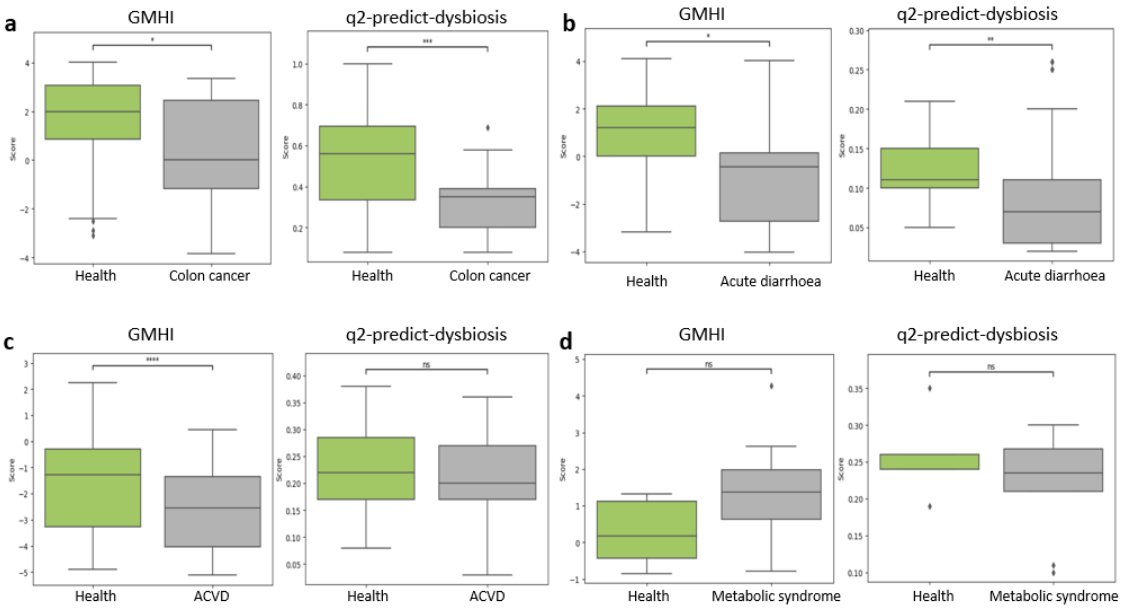

Figure 6. Comparison of the GMHI and the q2-predict-dysbiosis predictions for colon cancer (**a**), acute diarrhoea (**b**), artherosclerotic cardiovascular disease (**c**) and metabolic syndrome (**d**).

Score robustness

To evaluate the ability of both approaches to compare longitudinal 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, still non-significant decrease in health of the control staff as the pandemic progressed (Figure 7a). Interestingly, while the **`q2-predict-dysbiosis`** determined a predictable decrease in the health of ICU patients as they proceeded to the ICU, the **`GMHI`** produced the opposite trend, wrongly indicating patient recovery. In addition, the **`GMHI`** failed to identify differences between the last ICU and noICU cases, whereas the separation of the two groups 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, this time the difference between samples collected at admission and when the patients were transferred to the ICU disappeared. On the other hand, the **`GMHI`** produced the highest scores for the patients at the ICU and failed to identify controls.

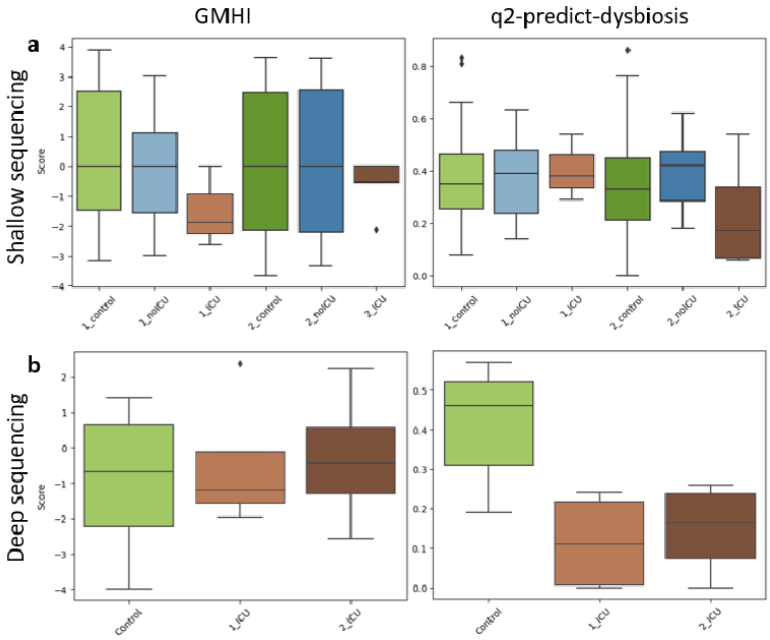

Figure 7. Comparison of the GMHI and q2-predict-dysbiosis predictions on COVID-19 samples processed using shallow (**a**) and deep (**b**) sequencing.

The separation of 1_ICU and 2_ICU patients varied between the sequencing types, so we wondered about the stability of both methods against sequencing depth. In order to investigate that, we applied different degrees of rarefaction to deeply sequenced control samples and expected to observe similar scores produced despite changing depths. The scores appeared to climb with the increased degree of rarefaction, however, the increase was slight and occurred for both methods (Figure 8a,b). The results corresponding to the lowest coverage considered (2,000,000 reads chosen as a reasonable consensus between cost and resolution (Hillmann et al., 2018)) would be visibly further from the remaining values, indicating 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 8e,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 not surprising considering how dependent on low-abundance functions they are (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 of good and bad species, respectively). This overlap indicated further issues with taxonomy-based indexes. Overall, we show that the **`q2-predict-dysbiosis`** is nearly as stable as the **`GMHI`** across sequencing depths and is surprisingly robust considering the variation of its components. However, consideration for depth is required when comparing health between samples with significantly different coverage.

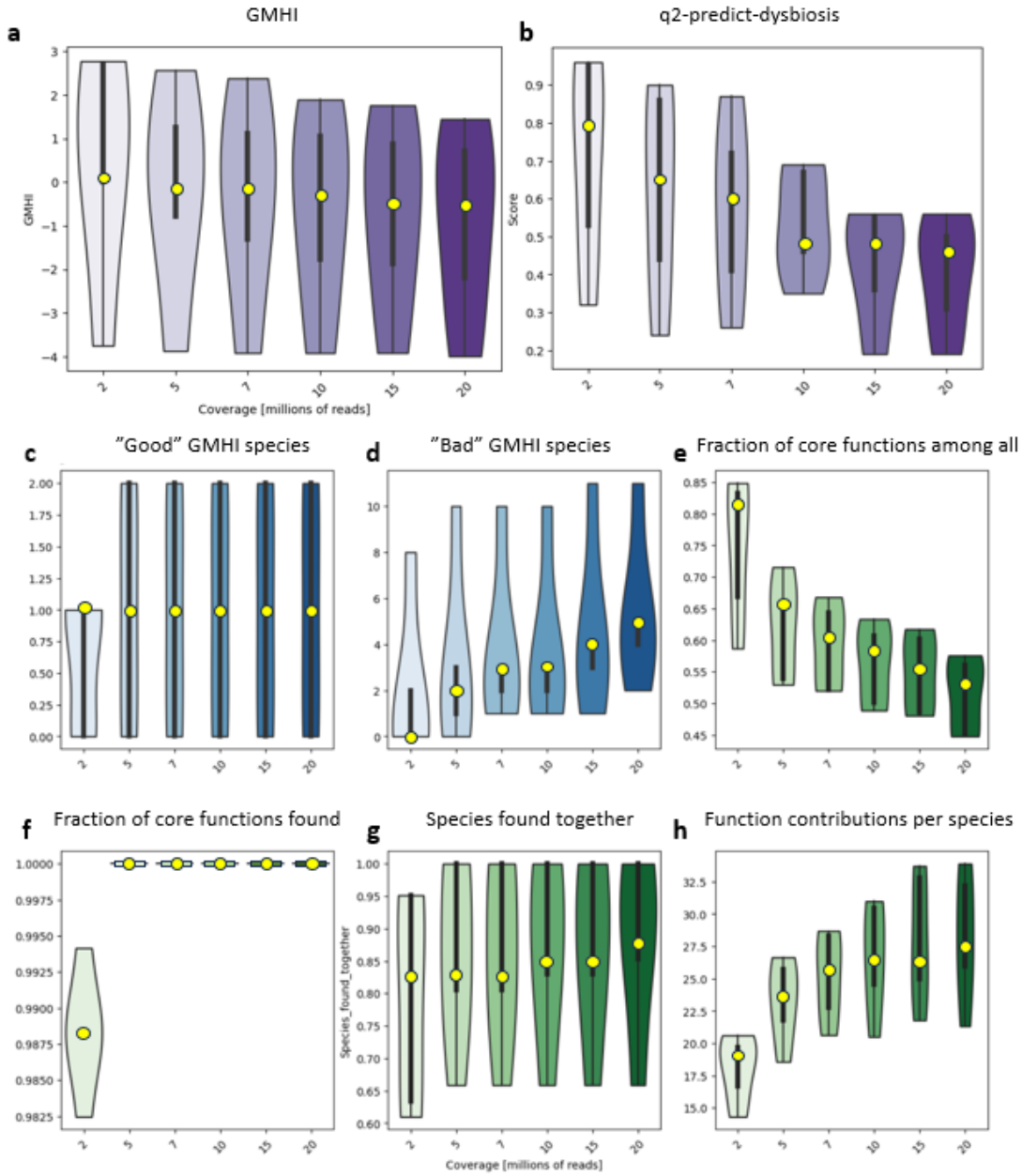

Figure 8. Robustness of health predictions to sequencing coverage. GMHI (**a**) and q2-predict-dysbiosis (**b**) scores for deeply sequenced healthy samples from the COVID-19 cohort, rarefied to corresponding depths. Taxonomic GMHI-inspired q2-predict-dysbiosis features (**c**,**d**) remain relatively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

Discussion

The connection of the human gut microbiome to health is now well established, and there have been a number of approaches aiming to identify dysbiosis from stool samples. Those methods are

based on taxonomy and rely either on measures of microbiome richness (alpha / beta diversity) or on the presence of "good" and "bad" bacteria, and include indexes such as the **GMHI** or **hiPCA** (Gupta et al., 2020, Zhu et al., 2023). However, the re-defined microbiome definition and recent findings highlight the importance of interactions between microbiome components as well as its functional aspects (Berg et al., 2020). Therefore, focusing solely on taxonomy is insufficient. In order to address those issues, we developed a novel approach to microbiome dysbiosis assessment. 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. 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).

Future work

There are still some challenges that need to be addressed in subsequent developments of the method. We believe that a deeper understanding of how the identified function- and interaction-based microbiome features respond to different sequencing depths and quality is needed. We found longitudinal data of particular interest as it could provide more insights 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 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 genes as then one can take advantage of recent developments in augmented functional annotation, 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 the decreasing costs of metagenomics, metabolomics and proteomics data, predicting health status by the incorporation of all three might be a reasonable option that carries significant accuracy improvements in the not-too-distant future.

In the end, we would like to highlight that the q2-predict-dysbiosis is only based on a few parameters and already reaching good levels of accuracy. Motivated with this success we are convinced that we are heading in the right direction towards a better understanding of the human gut microbiome and its health.

Methods

Metagenomic Data Processing

The overview of datasets used in this study can be found in Supplementary Table 4. All raw 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 MultiDimensional 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

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.

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