Healthy microbiome - moving towards functional interpretation

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

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

2 - Department of Biological Sciences, Bioinformatics and Computational Biology Program, 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 – Faculty of Computer Science, University of Białystok, Białystok, Poland

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

6 - Institute of Molecular Biology and Genetics of National Academy of Sciences of Ukraine, Kyiv, Ukraine

7 - Kyiv Academic University, Kyiv, Ukraine

Abstract

Microbiome-based disease prediction has significant potential as an early, non-invasive marker of multiple health conditions linked to dysbiosis of the human gut microbiota, thanks in part to decreasing sequencing and analysis costs. Microbiome health indices and other computational tools currently proposed in the field often are based on a microbiome's species richness and are completely reliant on taxonomic classification. A resurgent interest in a metabolism-centric, ecological approach has led to an increased understanding of microbiome metabolic and phenotypic complexity revealing substantial restrictions of taxonomy-reliant approaches. In this study, we introduce a new metagenomic health index developed as an answer to recent developments in microbiome definitions, in an effort to distinguish between healthy and unhealthy microbiomes, here in focus, inflammatory bowel disease (IBD). The novelty of our approach is a shift from a traditional Linnean phylogenetic classification towards a more holistic consideration of the metabolic functional potential underlining ecological interactions between species. Based on well-explored data cohorts, we compare our method and its performance with the most comprehensive indices to date, the taxonomy-based Gut Microbiome Health Index (**GMHI**), and the high dimensional principal component analysis (**hiPCA)** methods, as well as to the standard taxon-, and function-based Shannon entropy scoring. After demonstrating better performance on the initially targeted IBD cohorts, in comparison with other methods, we retrain our index on an additional 27 datasets obtained from different clinical conditions and validate our index's ability to distinguish between healthy and disease states using a variety of complementary benchmarking approaches. Finally, we demonstrate its superiority over the **GMHI** and the **hiPCA** on a longitudinal COVID-19 cohort and highlight the distinct robustness of our method to sequencing depth. Overall, we emphasize the potential of this metagenomic approach and advocate a shift towards functional approaches in order to better understand and assess microbiome health as well as provide

directions for future index enhancements. Our method, **`q2-predict-dysbiosis (Q2PD),`** is freely available (https://github.com/Kizielins/q2-predict-dysbiosis).

Introduction

The prevalence of a range of diseases and conditions peripherally or directly linked to microbiome health such as Inflammatory Bowel Disease (IBD), diabetes, obesity and even various cancers, continue to increase globally and substantial funds are currently spent on diagnosis and treatment (M'koma, 2013, Hong et al., 2019). While a correlation between gut microbiome composition and human health is widely acknowledged (Vijay and Valdes, 2022), the accurate identification of microbial and host markers of disease states remains elusive. Accordingly, the ability to evaluate patient health status based on a gut microbiome snapshot would be of high clinical value. Stool-based methods are promising because they can be collected non-invasively and frequently, and analysis time is short. Furthermore, decreasing costs of stool-sample analysis via next generation sequencing makes such microbiome characterization a strong competitor as a diagnostic tool (Bajaj et al., 2020).

Dysbiosis, defined as a perturbation of gut homeostasis, is believed to be accompanied by reduced microbiota diversity and increased prevalence of 'harmful' bacteria in adults (DeGruttola et al., 2016, Hrncir, 2022). Eubiosis (opposite of dysbiosis) can be perturbed by a wide range of factors including infection, diet, exercise, antibiotics, stress or poor sleep (Martinez et al., 2021). The simplest interventions currently applied for the prevention or alleviation of mild microbiome dysbioses include dietary modification or prebiotics (often non-digestible food types which promote the growth of beneficial microorganisms), ingested live bacteria or probiotics (beneficial bacteria usually in capsules), and lifestyle changes. More severe cases of gut dysbiosis, failing to respond to the above interventions, may qualify for fecal microbiota transplants, FMTs, which are increasingly gaining traction in clinics worldwide (Bull and Plummer, 2015). However, host-microbiota and intra-

68 microbiota interactions are both extremely complex and highly individual, and despite success with
69 FMT's, we have as yet no real understanding of why or how they work.

70 There are a number of accepted approaches used for the evaluation of a given gut microbiome's
71 health status based on stool composition. Alpha diversity (Shannon entropy, for example) is a
72 frequent choice, as microbiome richness was long believed to be a key driver of microbiome health
73 and robustness (Li et al., 2022, Gong et al., 2016). Beta diversity has also been applied in a number of
74 longitudinal studies, albeit to a lesser degree and mainly to identify eubiotic samples based on a
75 time-resolved proximity to other healthy samples (Zouiouich et al., 2021). The most robust index to
76 date, outperforming diversity indices, is the Gut Microbiome Health Index, or the **GMHI** (Gupta et
77 al., 2020), recently renamed the Gut Microbiome Wellness Index (**GMWI**). An updated version of this
78 index has recently been published (Chang et al., 2024). The GMHI is based on the ratio of 50
79 microbial species associated with healthy or unhealthy gut ecosystems and is reported to exceed
80 73% accuracy in determining disease state, thus the authors suggesting that gut taxonomic
81 signatures can predict health status. Another metagenomic gut-health index expanding on the **GMHI**
82 approach, **hiPCA** (Zhu et al., 2023), was introduced as a monitoring framework for personalized
83 health purposes. The personalized approach is achieved by analyzing the contribution of each
84 bacterium to the index, which allows for the identification of high-influence (ostensibly keystone)
85 species in different patient groups. The **hiPCA** claim of better performance than the **GMHI** is
86 attributed to the authors' application of additional transformation and clustering algorithms.
87 Importantly, such studies are often defined by datasets limited in scope to industrialized nations and
88 thus a less than complete consideration of diet-environment-microbiome interactions.

89 However, a recently re-visited definition of the microbiome emphasizes the importance of not just
90 the microbiota (a community of microorganisms), but the whole "Theatre of Activity", ToA (Berg et
91 al., 2020). This ToA includes structural elements (proteins, lipids, polysaccharides), metabolites and
92 environmental conditions (Lee et al., 2024). It is tightly bound to its corresponding ecological niche,

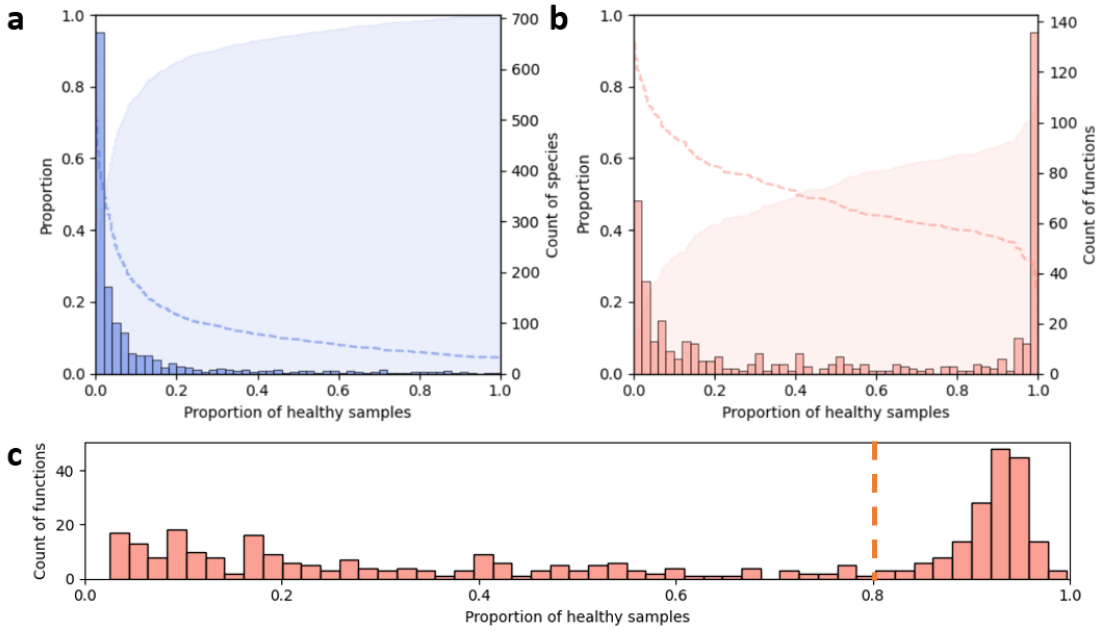
and the synergistic relations between species provide all the necessary, community defining components. Based on this definition, we maintain that an index constructed from taxonomy alone is hardly sufficient to accurately capture biological phenomena occurring within the gut environment – the key to understanding gut dysbiosis. Instead, we hypothesize that **to effectively determine health:** i) a metagenomic functional profile is required (microbiome phenotype); and ii) species interactions (e.g measured as co-occurrence) but not just presence should be considered.

We introduce an approach that is based on identifiable metagenomic features within ecosystems that extend beyond diversity measures and basic taxonomic information. This **function-**centricism is broached in two ways: i) directly by evaluating the functional potential within and between species and ii) indirectly by assessing co-occurrence and synergism between bacterial species. Our goal is not only to distinguish between healthy and diseased, but importantly, to also quantify the degree of dysbiosis in each sample for the given cohort. We derive the health-describing features based on an exploratory analysis of healthy samples from the Human Microbiome Project 2 (HMP2, Lloyd-Price et al., 2019) and outperform **`Shannon entropy,`** the **`GMHI`** and the **`hiPCA`** in healthy versus inflammatory bowel disease (IBD) and obese classification. The robustness of our index is further validated by corroboratively, classifying two additional IBD-focused cohorts. By retraining the IBD-specific parameters on an additional set of 30 diverse cohorts encompassing a range of diseases, we demonstrate the superior performance of our approach. Our findings reveal that function-, rather than taxonomy-based features are more informative for the accurate classification of biological samples. Additionally, our method effectively identifies longitudinal microbiome changes in COVID-19 patients, which the **`GMHI`** and the **`hiPCA`** are unable to capture, and, crucially, it is <u>distinctly robust</u> to sequencing depth. Our method **`q2-predict-dysbiosis (Q2PD)`** is freely available (https://github.com/Kizielins/q2-predict-dysbiosis).

Results

High prevalence of “core functions” in health

118 In order to develop a strategy to assess the degree of dysbiosis in a given microbiome sample, we
119 must define eubiosis, i.e. the healthy microbiome. We based our initial analysis (described in the
120 Methods section) on 384 medically-determined healthy samples from the HMP2 project and
121 identified the most prevalent species, regardless of abundance (Figure 1a). We observed that 50% of
122 species were present in less than 5% of samples, and hardly any species were shared by all
123 individuals. On the other hand, the prevalence of functions within the healthy population had an
124 opposite trend - 50% of functions were already represented by at least 40% of individuals (Figure
125 1b), a functional redundancy unaccounted for in the **`GMHI`** (or the **`hiPCA`** which is based on it).
126 These results convinced us further about the unsuitability of basing an index purely on the presence
127 of "core taxa" and encouraged a shift of focus towards more prevalent functions instead.



128

129 **Figure 1.** Distributions of species **(a)** and functions **(b)** present in healthy samples from the HMP2;
130 absolute values of species and function counts are shown as histograms (with scales on the right-
131 hand side, with shaded cumulative sum in the background and an inverse of the cumulative sum
132 represented with a dashed line (with scales on the left-hand side). **(c)** Distribution of functions in
133 healthy individuals from the HMP2 and two validation cohorts.

134 The addition of healthy samples from two validation cohorts maintained the function distribution
135 profile obtained solely with HMP2 samples (Figure 1c). In order to test whether the functions were
136 universal or cohort-specific, we calculated separately the distributions for functions present in 1, 2,
137 or all 3 cohorts. We found that all functions missing from at least 1 cohort were present in less than
138 10% of samples, which indicated the presence of high-prevalence functions in all three cohorts.
139 Based on the increased occurrence of certain functions in over 80% of samples (dotted line in Figure
140 1c), we defined them as “core functions” (refer to Supplementary Table 1 for full list). According to
141 MetaCyc classification, 73.5% of “core functions” were assigned as “Biosynthesis” pathways, 18.8%
142 “Degradation/Utilization/Assimilation”, and 7.6% “Generation of Precursor Metabolites and Energy”.
143 In addition, a few of the above were additionally classified as “Superpathways”. The classification
144 aligned well with a previously reported high prevalence of carbohydrate and amino acid metabolism-
145 related pathways, potentially forming the functional microbiome core (Zou et al., 2019). A detailed
146 analysis of the core functions identified in our study, however, is out of the scope of this manuscript.

147 Shannon entropy calculated on species and functions allowed for good discrimination between
148 healthy and unhealthy samples in the HMP2 dataset (Figure 2a top). However, the trends were
149 unclear in the two validation cohorts (Figure 2a middle and bottom), reflecting the need for more
150 complex methodology, expanding beyond microbiome richness, in order to classify datasets without
151 obvious separation. Diversity analyses revealed that the number of functions per sample remained
152 similar, or even increased, during microbiome transitions from healthy state to dysbiosis in HMP2
153 (Supplementary Figure 1). While admittedly, this cannot be measured solely using metagenomics
154 data, the similarity could hypothetically be due to altered expression of genes usually silenced in the
155 eubiotic state, although this observation was not reproduced in the validation cohorts. Next, we
156 investigated the presence of “core functions” in different groups, testing whether “core functions”
157 are maintained or replaced by others in dysbiotic, IBD, samples. While the differences were not
158 significant in most cases, we noted a visibly higher percentage of "core functions" and a higher
159 percentage of all "core functions" in healthy as compared to disease samples (Supplementary Figure

160 2). We carried out differential enrichment analysis using LEfSe (Segata et al., 2011), performed
161 separately for each cohort, and identified 100 functions that were more abundant in healthy as
162 compared to disease cohorts (Figure 2b, Supplementary Figure 2). Over 90% of functions enriched in
163 healthy samples were "core functions", while they constituted less than 5% of functions enriched in
164 the unhealthy class of validation cohort 2 and HMP2 (Figure 2c). The LEfSe analysis on validation
165 cohort 1 revealed only 10 significantly enriched functions (7 in health and 3 in disease), all of which
166 were core. This indicated a more heterogeneous functional landscape within this cohort.

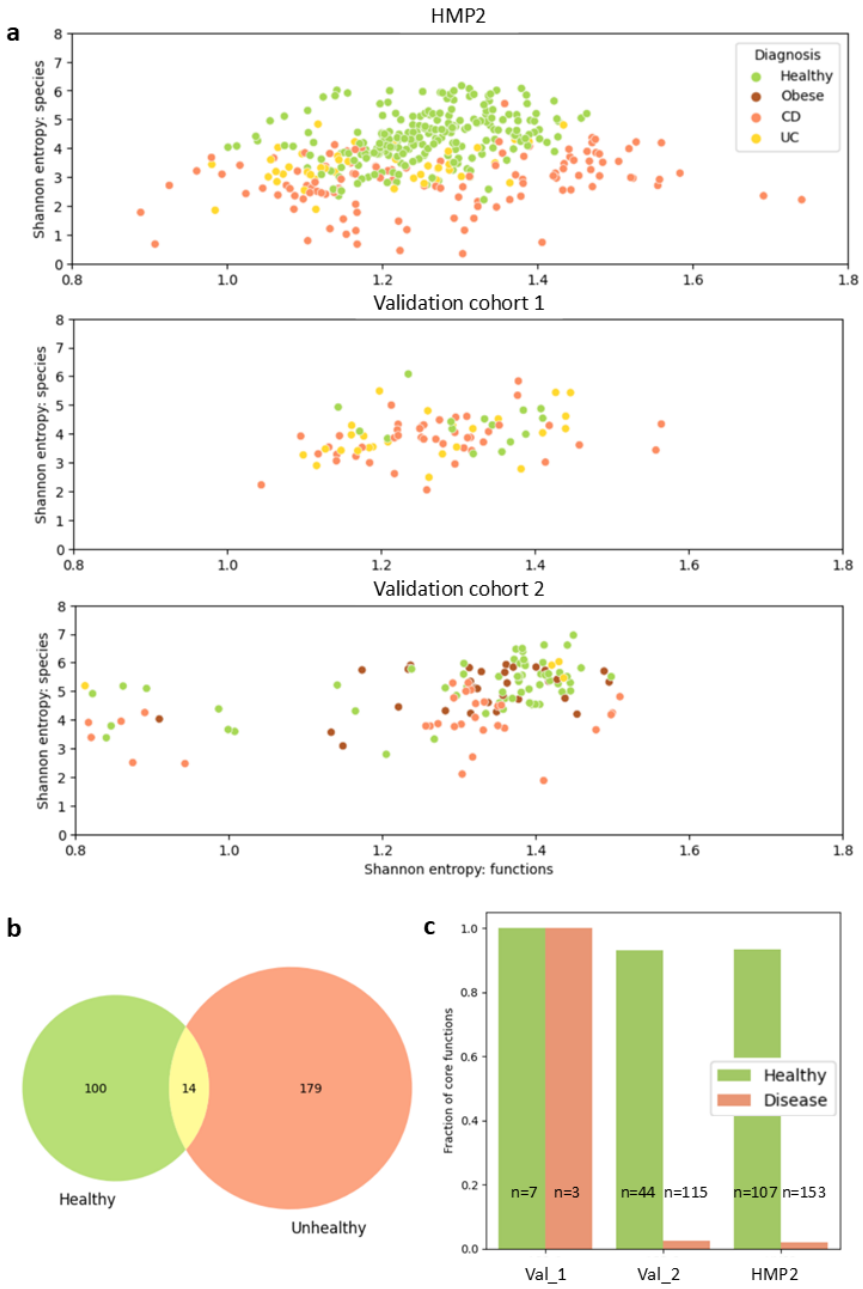


Figure 2. (**a**) Shannon entropy scores for species and functions in healthy and unhealthy samples from the HMP2 and validation cohorts. (**b**) LEfSe differential enrichment analysis: overlap of enriched pathways in healthy and unhealthy individuals in the validation and HMP2 projects. (**c**) Fraction of core functions among differentially enriched functions in healthy and unhealthy individuals in the validation and HMP2 cohorts.

Species interactions and function contributions in health

Corroborating results from past studies, we observed a decrease in the species abundance 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 (Supplementary Figure 1), 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 health- / disease-state separation. We chose the 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 manuscript submission (Piliszek et al., 2019, Mnich & Rudnicki, 2020). This approach reduced the number of relevant species to 587 allowing us to eliminate noise and focus on the most important interactions (see Methods for more details about the feature selection procedure).

We then used the SparCC algorithm, designed specifically for compositional data, to investigate correlations between the MDFS-selected 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 group, that would indicate differences between the two. However, we identified opposite relationships of some species in different groups (Figure 3a). A number of species we found to be positively correlated in eubiosis and are 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 observed that the prevalence of the pairs positively correlated with health was higher than in a number of disease-associated groups (Figure 3b). Due to this, we included the co-occurrence of such species as another feature of interest to aid potentially in the determination of microbiome health.

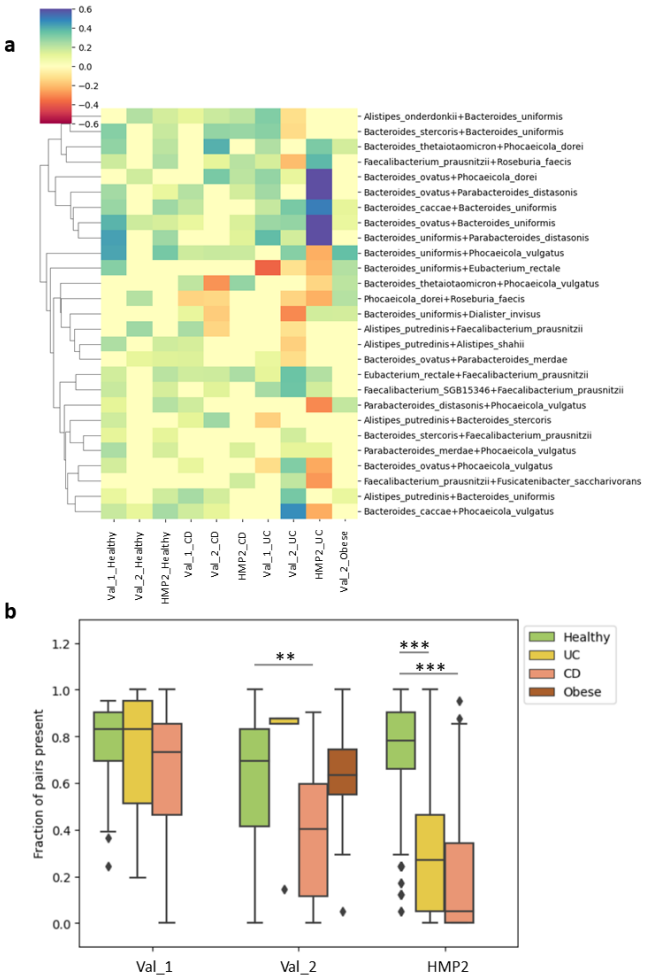


Figure 3. (**a**) **SparCC** correlation strengths between species, restricted to pairs which were not negatively correlated in any healthy cohort. (**b**) Prevalence of the pairs in different cohorts.

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 disease. 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. However, we found stronger results when exploring functional redundancy. While the average number of species per function and the average number of functions per species did not reliably separate healthy from diseased profiles (Supplementary

Figure 5), the latter approach was more informative as described in detail below. This finding was congruent with our earlier suspicions of an inherent functional plasticity of microbiome structure; modulation of function altering connectivity in the interaction network, leading to a shift towards less abundant, non-core functions upon perturbation of homeostasis. It also highlighted the challenge of identifying dysbiosis based on singular features, which were never statistically significant for all cohorts – and the utility of multiple perspectives for microbiome description, in order to meaningfully classify them.

Testing the accuracy of prediction for healthy and IBD individuals

Our final set of health-defining microbiome features included the following parameters (details of how each feature was calculated can be found in Table 1):

i) the fraction of "core functions" found,

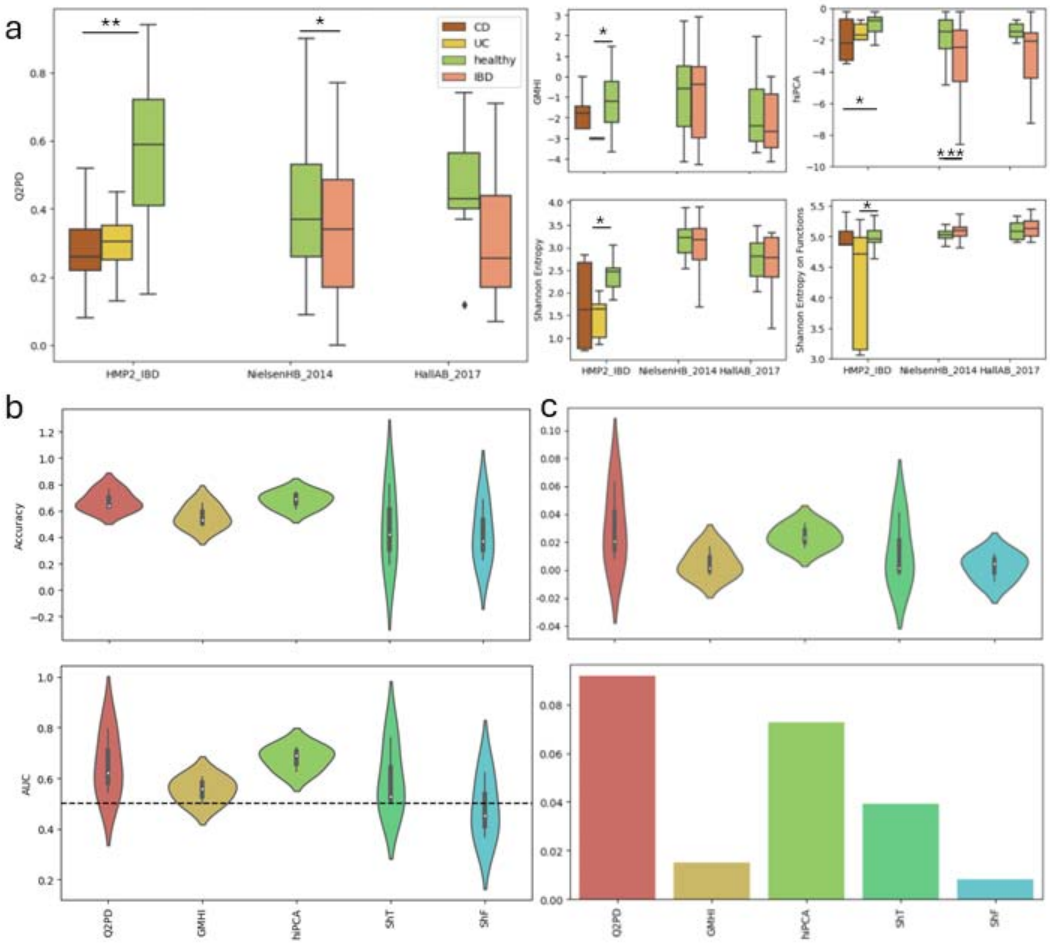
ii) the proportion of "core functions" among all functions,

iii) the proportion of co-occurrent species pairs in healthy samples,

iv) the average number of functional 'contributions' per species.

In addition, we 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. We opted for a Random Forest classification algorithm (Breiman, 2001) due to its robustness to imbalanced data and interpretability, and performed a leave-one-out cross-validation . For the model training, validation and subsequent testing, we used taxonomic and functional profiles from the curated Metagenomics database (Pasolli et al., 2017) to ensure consistency and reproducibility (see Methods).

Our index demonstrated the strongest statistically significant separation between healthy and Crohn's disease or ulcerative colitis individuals across all methods evaluated (Figure 4a). Alongside hiPCA, it was one of only two approaches to achieve a statistically significant distinction in the Nielsen_2014 cohort. Overall, both our index and hiPCA exhibited comparable levels of accuracy and AUC across the three cohorts, outperforming the other methods by a notable margin (Figure 4b). By contrast, the GMHI consistently delivered mediocre results, and Shannon entropy only performed well applied to the HMP2 cohort, yielding low AUC scores (0.37–0.53) in all other cases. When index values were used as features in the Boruta algorithm (Kursa & Rudnicki, 2010), with health status (0/1) as the target variable, our index emerged with the highest mean and summed importance across all cohorts (Figure 4c) . The above highlights the ability of our index to provide the most informative scores for health status prediction as compared to the other methods.



245 **Figure 4. (a) Q2PD, GMHI, hiPCA** and **Shannon entropy (on species and**
246 **functions)** scores for healthy and IBD (Crohn's disease = CD, ulcerative colitis = UC) individuals.
247 (**b**) Accuracy and AUC values for each index, per IBD cohort. ShT = Shannon entropy on taxa, ShF =
248 Shannon entropy on functions. (**c**) Average (top) and summed (bottom) importance of each index in
249 the context of IBD prediction, determined by Boruta.

250

251 Beyond IBD

252 While only healthy and IBD individuals had erstwhile been included in the development and
253 validation of our approach, we wondered about the applicability of the **Q2PD** to dysbiosis attributed
254 to other diseases. We extended our dataset to another 27 additional cohorts from various disease
255 states, equating to 30 datasets used for method validation. We did not change any parameters of
256 the **Q2PD,** which were originally determined based on the healthy samples from the HMP2.
257 Instead, we retrained the model with the new data and performed a leave-one-cohort-out approach
258 to ensure a robust benchmark. The procedure placed our index at a disadvantaged position, as some
259 of the added datasets had been used to develop and train the other methods and our approach was
260 thus truly blind to outcome in these new cases.
261 Despite the disability, and gratifyingly, the **Q2PD** achieved in terms of performance, the highest
262 average accuracy and AUC across all datasets {(AUC=0.61, accuracy=0.58) > hiPCA (AUC=0.58,
263 accuracy=0.57) > GMHI (AUC=0.55, accuracy=0.55) > Shannon entropy on species (AUC=0.52,
264 accuracy=0.53) > Shannon entropy on functions (AUC=0.44, accuracy=0.43)} (See Figure 5a). The
265 consistently poor performance of both entropy-based measures suggested their highly limited utility
266 as predictive indices.

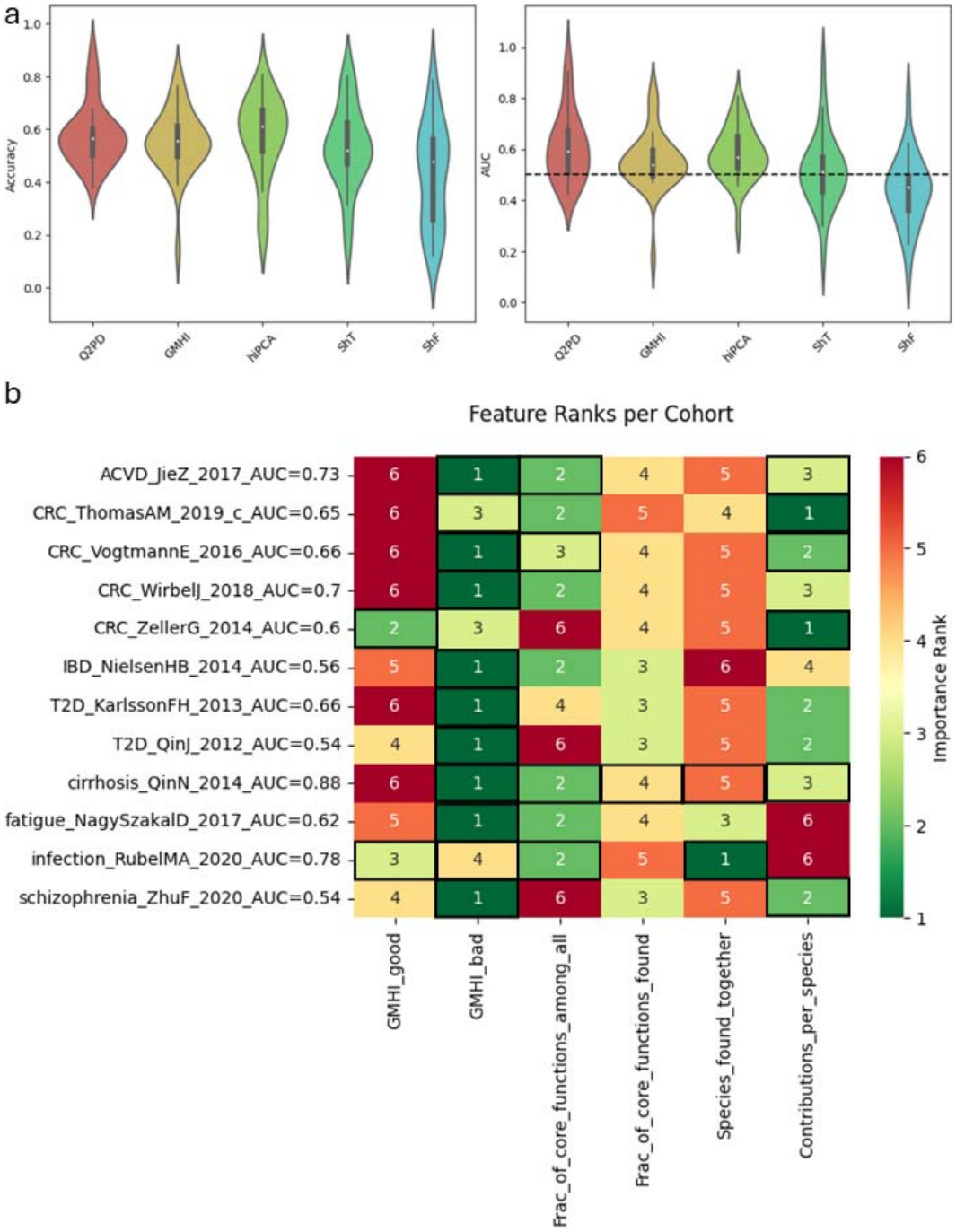


Figure 5 (**a**) Average accuracy and AUC values for each cohort. (**b**) Average importance ranks of features for each cohort. A lower rank indicates greater importance. Cells with black borders indicate variables identified as informative by the Boruta algorithm. The AUCs were produced as a

271 part of the feature importances analysis when training on individual cohorts, and therefore do not
272 align with the AUCs produced by the Q2PD. Only datasets with AUCs > 0.5 are shown.

273

274 We observed that where **Q2PD** classified a particular cohort better than the other indices, it did so
275 with a significantly greater margin than when it lost to the other methods. Its average winning AUC
276 margin over the hiPCA, the second best classifier, was 0.19 while the losing margin to the hiPCA was
277 0.09 when the hiPCA had the highest AUC. The average winning AUC margin of Q2PD against the
278 mean of the other indices was 0.20 and 0.11 if any other of them was better. In both cases, the T-
279 test statistics for the differences between the means of the Q2PD's winning AUC margins and that of
280 the hiPCA or the average of others produced p-values of 0.03 and 0.02, respectively, indicating a
281 significant classification improvement with our method in areas in which the remaining indices did
282 not classify well. The improvement was even more striking when we excluded datasets that had
283 been used for the training of either method. In this case, the winning margin of the Q2PD was 0.22
284 while the losing margin to the mean of the other indices if any of them was better was 0.09.
285 Our investigation into the accuracy and AUC of the indices for each cohort revealed substantial
286 variability in terms of the classes of cohort that each index was able to classify (Supplementary
287 Figure 6). We observed that while some cohorts such as Liss_2016 could be classified well with
288 function-based indices (Q2PD and Shannon entropy on functions), other cohorts such as
289 Gupta_2019 were slightly better separated with taxonomy-based indices (hiPCA, GMHI and Shannon
290 entropy on species). An exploration of the importance of feature(s) associated with the model
291 training on each dataset alone revealed a large amount of diversity, suggesting different kinds of
292 information used to classify different cohorts (Figure 5b). Interestingly, when trained on the level of
293 individual cohorts, a random forest would pick the "GMHI_bad" as its most informative parameter
294 (scoring the lowest rank in 9 cases) and the "GMHI_good" as the worst (appearing at the bottom of
295 the ranking 6 times). The importance of the function-based features would vary depending on the

dataset, with "Contributions_per_species" winning and losing twice and the other two being consistently in the middle of the ranking.

Having observed a discrepancy between the superior Q2PD performance and the greatest importance of the "GMHI_bad" parameter, we investigated this further by testing which types of diseases were best suited for each index to utilize. To this end, we designed a ranking method based on the number of times each index achieved the highest AUC for the most cohorts for each disease. Overall, Q2PD outperformed the other indices for 6 diseases (atherosclerotic cardiovascular disease, colon cancer, infection, metabolic disease, schizophrenia and fecal microbiota transplant: donor versus patient classification). This was followed by hiPCA for 5 (Behcet's disease, IBD, type 2 diabetes, chronic fatigue and cirrhosis), Shannon entropy on taxa for 2 (Parkinson's disease and acute diarrhoea) and GMHI for 1 (a cohort with mixed diseases). From this we also calculated an average rank for each method for all the diseases above and yet again, Q2PD ranked best (average ranking=2.39), followed by hiPCA (2.71), Shannon entropy on taxa (3.14), GMHI (3.32) and Shannon entropy on functions (3.57). The poor performance of the GMHI indicated a diminished role of the "GMHI_bad" parameter when combining all datasets, and implied a better generalization of the health status using function-based parameters.

Q2PD robustness to longitudinal alterations and sequencing depth

To construct a final model, we trained the random forest classifier on the complete set of 30 cohorts. We then took advantage of a longitudinal COVID-19 dataset that had been sequenced both shallowly and deeply, and was 'unseen' by any methods, in order to evaluate performance of Q2PD across sequencing depths. The set consisted of three groups – COVID-19 patients who, during the course of the treatment, were either i) transferred to the intensive care unit (ICU) or ii) recovered (noICU), and iii) controls (healthy hospital staff). For every individual, two timepoints were selected - "1" that was collected upon hospital admission, or in the case of staff, early in the pandemic, and "2" which was the final sample taken from each individual. As expected, sequencing data with fewer

than 300,000 reads per sample fell short in accurately classifying individuals, primarily because the limited read depth did not provide sufficient coverage for comprehensive functional annotation (Figure 6a).

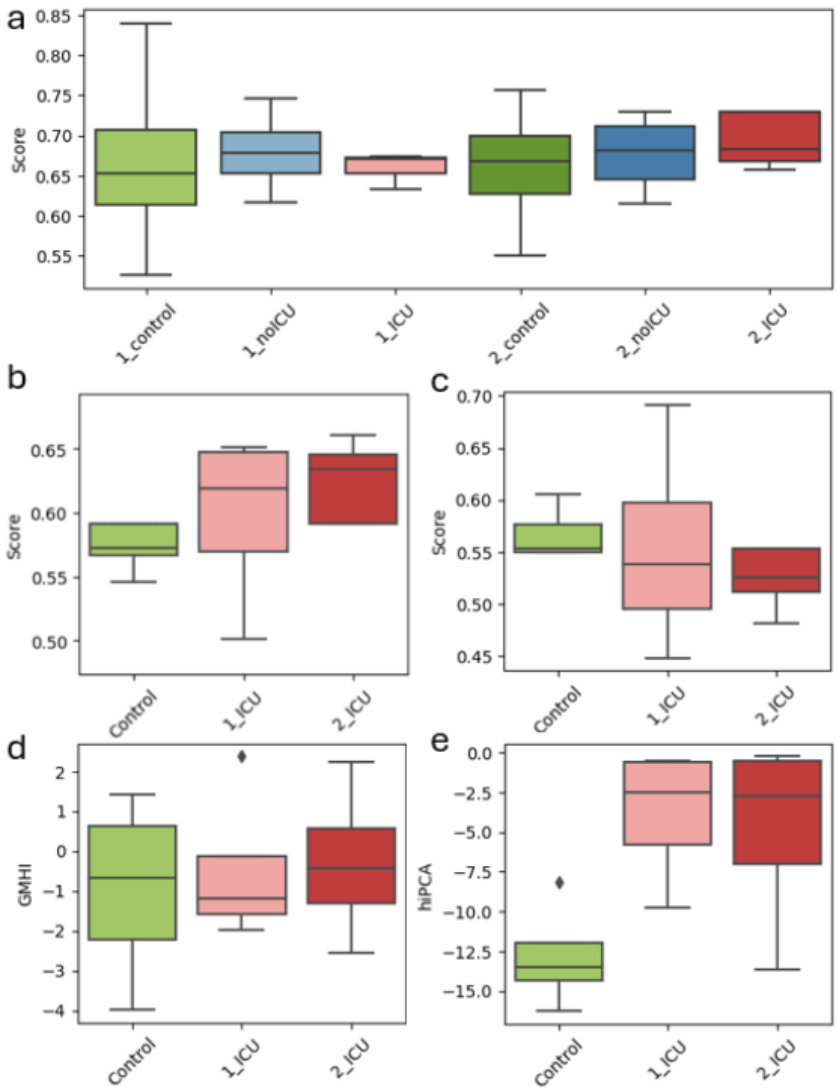


Figure 6. (**a**) **Q2PD** predictions on the shallow COVID-19 cohort. (**b**) . **Q2PD** predictions on the deep COVID-19 cohort. (**c**) **Q2PD** predictions on the deep COVID-19 cohort with the taxonomic parameters excluded. GMHI (**d**) and hiPCA (**e**) predictions on the deep COVID-19 cohort.

Surprisingly, despite its earlier described successes at classification, Q2PD was not able to distinguish between healthy and COVID-19 individuals sequenced deeply classifying ICU patients as healthy (Figure 6b). Curious about the reasons for this, we investigated feature importance and discovered

332 that the two taxonomic parameters “GMHI_good” and “GMHI_bad” both had negative values. This
333 was indicative of the detrimental influence of taxonomy-based features on the performance of the
334 model. As suspected, retraining the Q2PD without them led to the expected, correct predictions
335 (Figure 6c) suggesting that this dataset could be classified based on functional information alone. We
336 further validated this when we performed classification using the taxonomy-based GMHI (Figure 6d)
337 and the hiPCA (Figure 7e), which was inaccurate for GMHI and opposite for hiPCA.

338 Q2PD performed poorly on the shallow COVID-19 dataset which led us to ask where the threshold
339 lay for sequencing depth reliability. In order to investigate method robustness to depth, we applied
340 various degrees of rarefaction to deeply sequenced control samples expecting to see similar
341 prediction scores despite variable sequencing depth. Because the performance of the **hiPCA**
342 (second best classifier after the Q2PD overall) on the COVID-19 cohort was worse than that of the
343 **GMHI**, we decided to use the latter as our benchmark.
344 The scores produced by the Q2PD were robust and consistent regardless of sequencing depth,
345 whereas for **GMHI**, they increased with degree of rarefaction(Figure 7a,b). Furthermore, the **GMHI**
346 classification appeared to be only relevant for a foreboding depth of 2 million reads (score > 0), as
347 the healthy samples were defined as unhealthy at greater sequencing depth (score < 0). Our index
348 correctly identified healthy samples at any sequencing depth (scores were always above threshold).
349 We plotted the values for each feature separately and found that at low sequencing depths certain
350 species (Figure 7 c, d) or “core functions” (Figure 87e, f) were undetectable. The “Fraction of core
351 features among other features” and the “Function contributions per species” were those most
352 dependent on coverage, which was anticipated due to their dependence on low-abundance
353 functions (Figure 7 e - h). Interestingly, the numbers of “good” and “bad” species identified at any
354 sequencing depth covered only 32% of the complete **GMHI** list (2 and 14, versus 7 and 43 for good
355 and bad species, respectively). In line with the underlying theme of our work and in agreement with
356 all other presented data , this overlap argues strongly against indices that are based solely on
357 taxonomy.

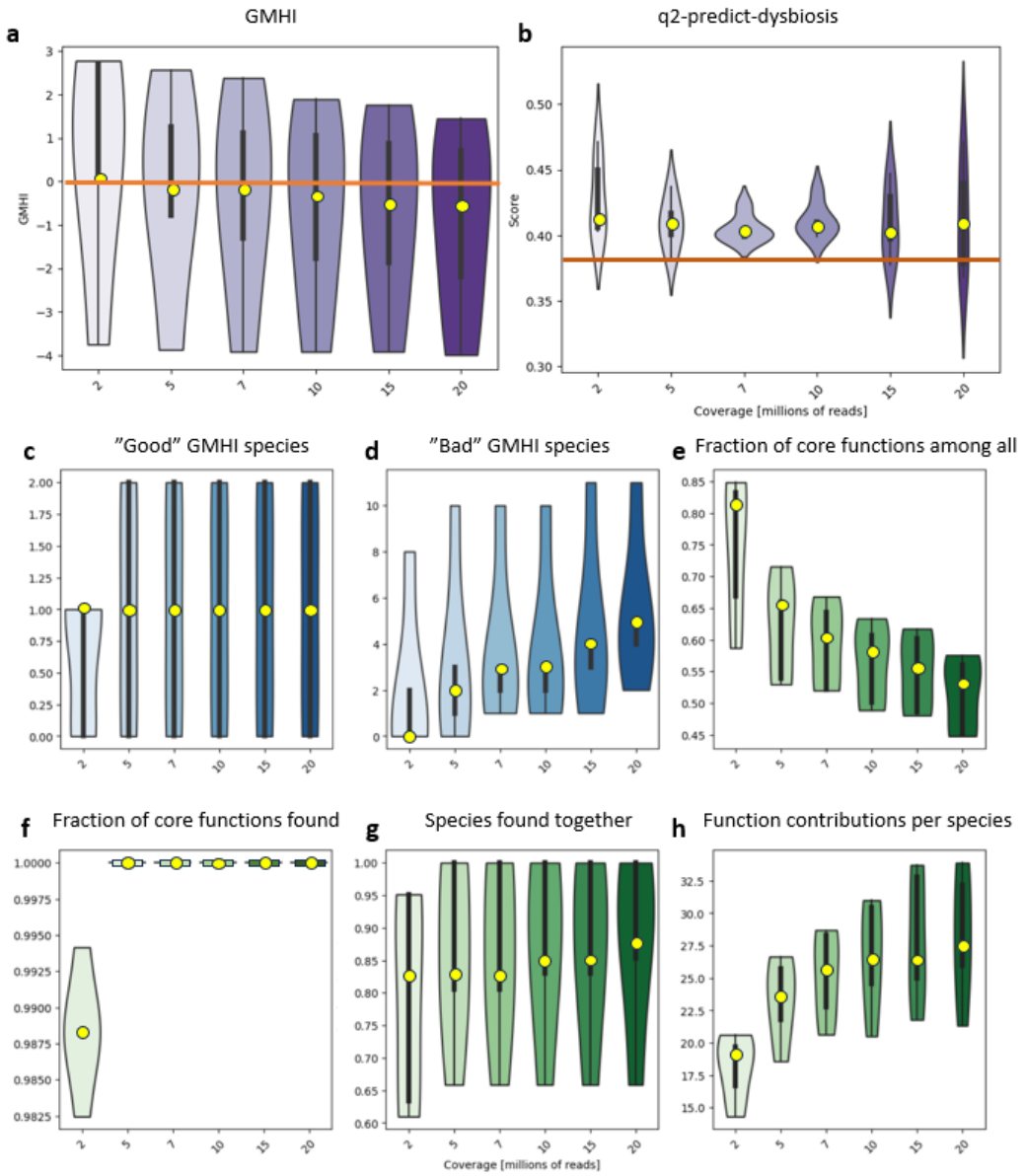


Figure 7. Robustness of health predictions to sequencing coverage. **`GMHI`** (**a**) and **`Q2PD`** (**b**) scores for deeply sequenced healthy samples from the COVID-19 cohort, rarefied to corresponding depths. The horizontal lines represent health thresholds relevant to the corresponding methods. Taxonomic **`GMHI`**-inspired **`Q2PD`** 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

366 A connection between the human gut microbiome and gut health is now well established, and a
367 number of approaches have been employed in an effort to identify gut dysbiosis from sequence-
368 based analysis of stool samples. Those methods are based on taxonomy and rely either on measures
369 of microbiome richness (alpha / beta diversity) or on the presence or absence of so-called "good"
370 and "bad" bacteria, with the health-indexes **GMHI** (Gupta et al., 2020) and **hiPCA** (Zhu et al.,
371 2023) proposed formally. Hampering such approaches, however, are ecological considerations of
372 metabolic or functional redundancies inherent within complex environments. Currently, re-defining
373 the microbiome to include **inferred functionality** is bolstered by recent studies that highlight the
374 importance of interactions between microbiome components and functional aspects thereof. To
375 address the inadequacy of dogmatic *Linnéan* (taxonomy-based) approach for evaluating change in
376 microbiomes, we developed a novel method that incorporates function in bioinformatics-based
377 assessment of microbiome dysbioses. We show that features based on microbiome functions and
378 interactions define a healthy microbiome more accurately. There exists a set of "core functions",
379 which are consistently identified as present in healthy gut microbiomes, and which disappear in the
380 advent of dysbiosis. We compare our results to the **hiPCA,** the **GMHI** and two **Shannon**
381 **entropy** measures (calculated on either species or on functions). Our index not only outperforms
382 the other methods in the originally targeted IBD classification, but also when applied to a range of
383 other diseases. Finally, it is robust to sequencing depth, unlike the **GMHI**, despite being based on
384 sequencing depth-sensitive parameters.
385 While we here present the index **Q2PD**, we acknowledge that challenges remain prior to eventual
386 clinical deployment as the method's robustness for diseases other than IBD and obesity is improved.
387 One of the most fascinating results from this work, apart from the model itself, was the finding that
388 different parameters were of varying importance across different diseases and cohorts. This finding
389 could have practical implications in the clinic, as it could indicate the directionality of the
390 microbiome-disease connections. By this we mean specifically that diseases originating in the gut,
391 e.g. IBD, are usually associated with taxonomic shifts and thus better classified with taxonomy-based

392 indices while diseases originating elsewhere may have (functional) effects on the gut and thus
393 better identified with function-oriented methods. It should be noted though that the etiology of
394 many diseases, i.e. whether they originate in the gut or not, remains unknown.

395 Clearly, a deeper understanding of how the identified function- and interaction-based microbiome
396 features respond to variations in sequencing depth and quality is required. Our initial test did find
397 that longitudinal data provides more insight into personalized "core features" of a microbiome, and
398 this may be indicative of individual deviations from the normal. Thus, tracking microbiome changes
399 in individuals over time may facilitate the early identification of microbiome trajectories or
400 alterations that could help determine risk for certain conditions or diseases, potentiating
401 personalized clinical intervention.

402 A caveat to this type of study is the infancy of certain fields, in particular, the availability of
403 functional annotation software. Currently, the only widely accepted functional annotation software
404 is HUMAnN (any version), which requires data to be processed in a particular way. In future work we
405 plan to explore other possibilities and ultimately migrate to other inputs (such as mifaser) as well as
406 towards a more universal solution, such as gene content (Zhu et al., 2017). Only then can recent
407 developments in augmented functional annotation be efficiently utilized, as we already have shown
408 in other applications (Maranga et al., 2023, Koehler Leman et al., 2023, Szydlowski et al., 2023).

409 We are also considering the introduction of a multi-omics approach to improve our predictions,
410 combining metagenomics, metabolomics and proteomics data. With decreasing costs for each of the
411 above, predicting patient health status based on all three -omics methods might be a reasonable
412 option that improves accuracy and prediction in the not-too-distant future. It would also allow us to
413 quantitatively estimate the functions directly, instead of basing our analyses on the functional
414 potential evaluated based on metagenomics data. As a primer for our current efforts, the
415 integration of multiple levels could be performed using the aforementioned MDFS, by investigating

relationships between the layers, or using more advanced network-based approaches such has ViLoN (Kańduła et al., 2023).

In conclusion, we highlight the performance and high accuracy obtained by **Q2PD,** based only on a limited number of highly relevant parameters. Despite its development to separate samples derived from healthy vs IBD individuals, its broad applicability was proven to accurately distinguish between healthy patients and individuals with other disease states and conditions. Motivated by this success we are convinced that our approach and methodology provide a better mechanistic understanding of the human gut microbiome and its health, one based on resource competition and interactions based on fundamental ecological principles that ultimately define microbial community structure.

Methods

Defining metagenomic parameters associated with health

The raw shotgun sequencing fastq files for the HMP2 and the two IBD validation cohorts used to derive metagenomic health-associated parameters 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., 2019, Mnich & Rudnicki, 2020) 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). Metagenomic table formatting and alpha diversity plots were done using **QIIME 2** (Bolyen et al., 2019) 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.

IBD index

Random Forest is a classification and regression method (Breiman, 2001). The algorithm utilises an ensemble of CART trees (Leo Breiman, 1984), in which each tree is built using different bootstrap samples of data and different random subsets of variables at each stage of the tree construction. It is a robust and versatile algorithm that works well on different types of data (Fernandez-Delgado et al., 2014).

Metagenomic features passed to the random forest were constructed based on healthy samples from the HMP2 and two validation cohorts (Val_1 and Val_2). The features are presented in Table 1.

Table 1. Parameters of the Q2PD model.

Name	Explanation	Derivation
Frac_of_core_functions_found	fraction of "core functions" found	number of “core functions” in a sample / number of all “core functions” in our list
Frac_of_core_functions_among_all	fraction of "core functions" among all functions	number of “core functions” in a sample / number of all functions in a sample
Species_found_together	fraction of species pairs commonly occurring together in healthy samples	SparCC correlation of species > 0.1, only nonnegative correlations in healthy cohorts
Func_contributions_per_species	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
GMHI_good	number of “good” GMHI species	list of health-associated GMHI species
GMHI_bad	number of “bad” GMHI species	list of disease-associated GMHI species

The importance of the features was assessed using a permutation approach and 5-fold cross validation.

From this step on, we used the manually curated samples from the curatedMetagenomicsRepository to ensure consistent sample processing (MetaPhlAn (Blanco-Míguez et al., 2023) and HUMAnN (Beghini et al., 2021), specifically we used the MetaCyc functional annotations) as well as to provide an additional layer of validation of our approach. Our manual curation excluded a samples for the following reasons: incomplete metadata, use of antibiotics or other similar drugs, repeats from the

458 same patients, young age (newborns), or healthy samples not being fully healthy (high BMI, chronic
459 conditions). During filtering, we set the following thresholds: species abundance >= 0.1%, pathway
460 coverage >= 20% and function abundance >= 0.01% - with any features below that changed to 0.

461 During random forest training, healthy individuals were labeled as "1" and the IBD as "0". A leave-
462 one-out cross-validation procedure was applied to predict health scores for each sample in order to
463 avoid model overfitting or the need to split the data into training and testing sets. The final health
464 score is the output of the random forest's *predict_proba* method, which expresses the probability of
465 each sample belonging to the healthy (or unhealthy) group.

466 The GMHI scores were calculated with the **`QIIME 2 q2-health-index`** plugin ("q2-health-
467 index," 2023) and the **`hiPCA`** (Zhu et al., 2023) predictions were obtained by substituting the
468 original test file with our test samples, due to a lack of instructions about how to do it better (this
469 could potentially result in an overlap of the train and test samples, possibly falsely increasing the
470 **`hiPCA`**'s accuracy). The Shannon entropies were calculated separately on taxonomic and functional
471 profiles for each sample using custom python scripts.

472 The GMHI was the only parameter with pre-defined threshold for health. In the case of the other
473 methods, the optimal threshold would be defined using the Youden's J statistic calculated on the
474 training data.

475 The relevance of each index in the context of IBD predictions was calculated using **`Boruta`** (Kursa
476 & Rudnicki, 2010). The Boruta algorithm is a wrapper around the Random Forest classifier. It works
477 in the following way: a data set is extended by adding a randomly permuted copy of each original
478 variable - a so-called shadow variable. A predictive model for the decision variable is then built using
479 the random forest algorithm. The importance of each variable is estimated using a permutation test.
480 Boruta collects the information about the importance and then compares the importance score of
481 each original variable with the maximal importance achieved by a shadow variable.

482 The procedure is repeated multiple times, and a statistical test is performed. The variables are
483 eventually assigned to three classes: Confirmed (better than random), Rejected (no better than
484 random) and Tentative (those that could not be assigned neither to Confirmed no to the Tentative
485 class. The utility of the variable is a good indicator of the information importance carried by the
486 variable, also in situations where the synergistic interactions are important.
487 In the Boruta section of our analysis, the predictions of each index were passed as parameters and
488 the health labels (0/1) were used as decision vectors. Index importances were redefined as ranks –
489 with the highest importance marked as rank "1".

490 Index testing: expanding to other diseases

491 For the purpose of testing, the index was re-trained on the complete set of 30 datasets. However, no
492 parameters were modified. To ensure a reliable benchmark, a leave-one-cohort-out procedure with
493 a 5-fold cross validation was performed. The accuracy and AUC statistics were defined separately for
494 each test cohort based on the validation set (which was a subset of the training dataset).

495 The performance of our index was evaluated by calculating its winning margin over hiPCA, the
496 second best index, and the mean of the other methods. Specifically, for each test dataset, the AUC
497 of the hiPCA or the mean AUC of the other methods was subtracted from the AUC produced by the
498 Q2PD. The winning margin was defined as a mean of the AUC differences, separately for cases when
499 Q2PD won and when not.

500 Importance of the features contributing to the Q2PD score was evaluated by training a random
501 forest model separately for each cohort and using a feature permutation approach. The importances
502 were changed to ranks, with the highest ranks (lowest values) representing the greatest importance.

503 Final index and application to the COVID-19 dataset
504

The final index was trained on the complete set of 30 cohorts using 5-fold cross validation. The health threshold was defined based on a mean of thresholds defined on training data for each iteration of the leave-one-cohort-out approach and was set at 0.38 (which was very similar for all iterations).

To evaluate the indices in the context of the robustness to sequencing depth, healthy deeply sequenced samples from the COVID-19 dataset were rarefied to different sequencing depths using **`seqtk 1.4`** (Li, 2023).

Data and code availability

The HMP2 data can be found in Sequence Read Archive (SRA) with the accession code PRJNA398089. The two IBD validation cohorts can be located under accessions PRJNA389280 and PRJEB1220. The test cohorts were downloaded from the curatedMetagenomicsData repository (Pasolli et al., 2017). The samples from this repository were manually selected - we removed samples from children, reported control yet unhealthy samples, as well as those of patients undergoing antibiotic treatment. The COVID-19 cohort can be found under the accession PRJEB64515 in the European Nucleotide Archive. The Q2PD is deposited on GitHub and can be accessed along with sample accessions and scripts needed to reproduce our results here: https://github.com/Kizielins/q2-predict-dysbiosis. The biotools identifier of the software is biotools:q2-predict-dysbiosis and the RRID is SCR_026038.

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, dr. Tomasz Kościółek from the Sano Centre for Computational Medicine and the members of Bioinformatics Research Group at MCB JU for their comments and support.

Ethics declarations

All authors declare no competing interests.

Bibliography

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

554 and new challenges. Microbiome 8, 103. https://doi.org/10.1186/s40168-020-00875-0

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

561 Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm,
562 E.J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J.E., Bittinger, K., Brejnrod, A., Brislawn, C.J.,
563 Brown, C.T., Callahan, B.J., Caraballo-Rodríguez, A.M., Chase, J., Cope, E.K., Da Silva, R.,
564 Diener, C., Dorrestein, P.C., Douglas, G.M., Durall, D.M., Duvallet, C., Edwardson, C.F., Ernst,
565 M., Estaki, M., Fouquier, J., Gauglitz, J.M., Gibbons, S.M., Gibson, D.L., Gonzalez, A., Gorlick,
566 K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G.A., Janssen, S.,
567 Jarmusch, A.K., Jiang, L., Kaehler, B.D., Kang, K.B., Keefe, C.R., Keim, P., Kelley, S.T., Knights,
568 D., Koester, I., Kosciolek, T., Kreps, J., Langille, M.G.I., Lee, J., Ley, R., Liu, Y.-X., Loftfield, E.,
569 Lozupone, C., Maher, M., Marotz, C., Martin, B.D., McDonald, D., McIver, L.J., Melnik, A.V.,
570 Metcalf, J.L., Morgan, S.C., Morton, J.T., Naimey, A.T., Navas-Molina, J.A., Nothias, L.F.,
571 Orchanian, S.B., Pearson, T., Peoples, S.L., Petras, D., Preuss, M.L., Pruesse, E., Rasmussen,
572 L.B., Rivers, A., Robeson, M.S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R.,
573 Song, S.J., Spear, J.R., Swafford, A.D., Thompson, L.R., Torres, P.J., Trinh, P., Tripathi, A.,
574 Turnbaugh, P.J., Ul-Hasan, S., vander Hooft, J.J.J., Vargas, F., Vázquez-Baeza, Y., Vogtmann,
575 E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K.C., Williamson,
576 C.H.D., Willis, A.D., Xu, Z.Z., Zaneveld, J.R., Zhang, Y., Zhu, Q., Knight, R., Caporaso, J.G., 2019.
577 Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2.
578 Nat. Biotechnol. 37, 852–857. https://doi.org/10.1038/s41587-019-0209-9

579 Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32.
580 https://doi.org/10.1023/A:1010933404324
581 Breiman, L., Friedman, J., Olshen, R.A., Stone, C.J., 2017. Classification and Regression Trees.
582 Chapman and Hall/CRC, New York. https://doi.org/10.1201/9781315139470
583 Bull, M.J., Plummer, N.T., 2015. Part 2: Treatments for Chronic Gastrointestinal Disease and Gut
584 Dysbiosis. Integr. Med. Clin. J. 14, 25–33.
585 DeGruttola, A.K., Low, D., Mizoguchi, A., Mizoguchi, E., 2016. Current understanding of dysbiosis in
586 disease in human and animal models. Inflamm. Bowel Dis. 22, 1137–1150.
587 https://doi.org/10.1097/MIB.0000000000000750
588 Friedman, J., Alm, E.J., 2012. Inferring Correlation Networks from Genomic Survey Data. PLOS
589 Comput. Biol. 8, e1002687. https://doi.org/10.1371/journal.pcbi.1002687
590 Gong, D., Gong, X., Wang, L., Yu, X., Dong, Q., 2016. Involvement of Reduced Microbial Diversity in
591 Inflammatory Bowel Disease. Gastroenterol. Res. Pract. 2016, 6951091.
592 https://doi.org/10.1155/2016/6951091
593 Gupta, V.K., Kim, M., Bakshi, U., Cunningham, K.Y., Davis, J.M., Lazaridis, K.N., Nelson, H., Chia, N.,
594 Sung, J., 2020. A predictive index for health status using species-level gut microbiome
595 profiling. Nat. Commun. 11, 4635. https://doi.org/10.1038/s41467-020-18476-8
596 Gut Microbiome Wellness Index 2 for Enhanced Health Status Prediction from Gut Microbiome
597 Taxonomic Profiles | bioRxiv [WWW Document], n.d. URL
598 https://www.biorxiv.org/content/10.1101/2023.09.30.560294v1 (accessed 12.3.23).
599 Hanawa, Y., Higashiyama, M., Kurihara, C., Tanemoto, R., Ito, S., Mizoguchi, A., Nishii, S., Wada, A.,
600 Inaba, K., Sugihara, N., Horiuchi, K., Okada, Y., Narimatsu, K., Komoto, S., Tomita, K., Hokari,
601 R., 2021. Acesulfame potassium induces dysbiosis and intestinal injury with enhanced
602 lymphocyte migration to intestinal mucosa. J. Gastroenterol. Hepatol. 36, 3140–3148.
603 https://doi.org/10.1111/jgh.15654
604 Hillmann, B., Al-Ghalith, G.A., Shields-Cutler, R.R., Zhu, Q., Gohl, D.M., Beckman, K.B., Knight, R.,

605 Knights, D., 2018. Evaluating the Information Content of Shallow Shotgun Metagenomics.
606 mSystems 3, 10.1128/msystems.00069-18. https://doi.org/10.1128/msystems.00069-18
607 Hong, Y.-R., Huo, J., Desai, R., Cardel, M., Deshmukh, A.A., 2019. Excess Costs and Economic Burden
608 of Obesity-Related Cancers in the United States. Value Health J. Int. Soc.
609 Pharmacoeconomics Outcomes Res. 22, 1378–1386.
610 https://doi.org/10.1016/j.jval.2019.07.004
611 Hrncir, T., 2022. Gut Microbiota Dysbiosis: Triggers, Consequences, Diagnostic and Therapeutic
612 Options. Microorganisms 10, 578. https://doi.org/10.3390/microorganisms10030578
613 Koehler Leman, J., Szczerbiak, P., Renfrew, P.D., Gligorijevic, V., Berenberg, D., Vatanen, T., Taylor,
614 B.C., Chandler, C., Janssen, S., Pataki, A., Carriero, N., Fisk, I., Xavier, R.J., Knight, R., Bonneau,
615 R., Kosciolek, T., 2023. Sequence-structure-function relationships in the microbial protein
616 universe. Nat. Commun. 14, 2351. https://doi.org/10.1038/s41467-023-37896-w
617 Kopera, K., Gromowski, T., Wydmański, W., Skonieczna-Żydecka, K., Muszyńska, A., Zielińska, K.,
618 Wierzbicka-Woś, A., Kaczmarczyk, M., Kadaj-Lipka, R., Cembrowska-Lech, D., Januszkiewicz,
619 K., Kotfis, K., Witkiewicz, W., Nalewajska, M., Feret, W., Marlicz, W., Łoniewski, I., Łabaj, P.P.,
620 Rydzewska, G., Kosciolek, T., 2023. Gut Microbiome Dynamics and Predictive Value in
621 Hospitalized COVID-19 Patients: A Comparative Analysis of Shallow and Deep Shotgun
622 Sequencing. https://doi.org/10.1101/2023.11.29.568526
623 Krueger, F., 2023. Trim Galore.Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.
624 https://doi.org/10.1023/A:1010933404324
625 Chang, D., Gupta, V. K., Hur, B., Cobo-López, S., Cunningham, K. Y., Han, N. S., Lee, I., Kronzer, V. L.,
626 Teigen, L. M., Karnatovskaia, L. V., Longbrake, E. E., Davis, J. M., Nelson, H., & Sung, J. (2024).
627 Gut Microbiome Wellness Index 2 enhances health status prediction from gut microbiome
628 taxonomic profiles. *Nature Communications*, *15*(1), 7447. https://doi.org/10.1038/s41467-
629 024-51651-9

Fernandez-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (n.d.). *Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?*

Kańduła, M. M., Aldoshin, A. D., Singh, S., Kolaczyk, E. D., & Kreil, D. P. (2023). ViLoN—a multi-layer network approach to data integration demonstrated for patient stratification. *Nucleic Acids Research*, *51*(1), e6. https://doi.org/10.1093/nar/gkac988

Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software*, *36*, 1–13. https://doi.org/10.18637/jss.v036.i11

Lee, J.-Y., Bays, D. J., Savage, H. P., & Bäumler, A. J. (2024). The human gut microbiome in health and disease: Time for a new chapter? *Infection and Immunity*. https://doi.org/10.1128/iai.00302-24

Leo Breiman, J. F. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC.

Mnich, K., & Rudnicki, W. R. (2020). All-relevant feature selection using multidimensional filters with exhaustive search. *Information Sciences*, *524*, 277–297. https://doi.org/10.1016/j.ins.2020.03.024

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. (2017). Accessible, curated metagenomic data through ExperimentHub. *Nature Methods*, *14*(11), Article 11. https://doi.org/10.1038/nmeth.4468

Piliszek, R., Mnich, K., Migacz, S., Tabaszewski, P., Sułecki, A., Polewko-Klim, A., & Rudnicki, W. (2019). MDFS: MultiDimensional Feature Selection in R. *The R Journal*, *11*(1), 198–210.

Zhu, C., Miller, M., Marpaka, S., Vaysberg, P., Rühlemann, M. C., Wu, G., Heinsen, F.-A., Tempel, M., Zhao, L., Lieb, W., Franke, A., & Bromberg, Y. (2017). Functional sequencing read annotation for high precision microbiome analysis. *Nucleic Acids Research*, *46*(4), e23. https://doi.org/10.1093/nar/gkx1209

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.

656 Differences in Alpha Diversity of Gut Microbiota in Neurological Diseases. Front. Neurosci.
657 16.
658 Lloyd-Price, J., Arze, C., Ananthakrishnan, A.N., Schirmer, M., Avila-Pacheco, J., Poon, T.W., Andrews,
659 E., Ajami, N.J., Bonham, K.S., Brislawn, C.J., Casero, D., Courtney, H., Gonzalez, A., Graeber,
660 T.G., Hall, A.B., Lake, K., Landers, C.J., Mallick, H., Plichta, D.R., Prasad, M., Rahnavard, G.,
661 Sauk, J., Shungin, D., Vázquez-Baeza, Y., White, R.A., Braun, J., Denson, L.A., Jansson, J.K.,
662 Knight, R., Kugathasan, S., McGovern, D.P.B., Petrosino, J.F., Stappenbeck, T.S., Winter, H.S.,
663 Clish, C.B., Franzosa, E.A., Vlamakis, H., Xavier, R.J., Huttenhower, C., 2019. Multi-omics of
664 the gut microbial ecosystem in inflammatory bowel diseases. Nature 569, 655–662.
665 https://doi.org/10.1038/s41586-019-1237-9
666 Maranga, M., Szczerbiak, P., Bezshapkin, V., Gligorijevic, V., Chandler, C., Bonneau, R., Xavier, R.J.,
667 Vatanen, T., Kosciolek, T., 2023. Comprehensive Functional Annotation of Metagenomes and
668 Microbial Genomes Using a Deep Learning-Based Method. mSystems 8, e01178-22.
669 https://doi.org/10.1128/msystems.01178-22
670 Martinez, J.E., Kahana, D.D., Ghuman, S., Wilson, H.P., Wilson, J., Kim, S.C.J., Lagishetty, V., Jacobs,
671 J.P., Sinha-Hikim, A.P., Friedman, T.C., 2021. Unhealthy Lifestyle and Gut Dysbiosis: A Better
672 Understanding of the Effects of Poor Diet and Nicotine on the Intestinal Microbiome. Front.
673 Endocrinol. 12.
674 Mirsepasi-Lauridsen, H.C., Vrankx, K., Engberg, J., Friis-Møller, A., Brynskov, J., Nordgaard-Lassen, I.,
675 Petersen, A.M., Krogfelt, K.A., 2018. Disease-Specific Enteric Microbiome Dysbiosis in
676 Inflammatory Bowel Disease. Front. Med. 5.
677 M'koma, A.E., 2013. Inflammatory Bowel Disease: An Expanding Global Health Problem. Clin. Med.
678 Insights Gastroenterol. 6, CGast.S12731. https://doi.org/10.4137/CGast.S12731
679 Mosca, A., Leclerc, M., Hugot, J.P., 2016. Gut Microbiota Diversity and Human Diseases: Should We
680 Reintroduce Key Predators in Our Ecosystem? Front. Microbiol. 7.

681 Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D. T., Beghini, F., Malik, F.,
682 Ramos, M., Dowd, J. B., Huttenhower, C., organ, M., Segata, N., & Waldron, L. (2017).
683 Accessible, curated metagenomic data through ExperimentHub. *Nature Methods*, *14*(11),
684 Article 11. https://doi.org/10.1038/nmeth.4468
685 Piliszek, R., Mnich, K., Migacz, S., Tabaszewski, P., Sułecki, A., Polewko-Klim, A., Rudnicki, W., 2018.
686 MDFS - MultiDimensional Feature Selection.
687 Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., Huttenhower, C., 2011.
688 Metagenomic biomarker discovery and explanation. Genome Biol. 12, R60.
689 https://doi.org/10.1186/gb-2011-12-6-r60
690 Szydlowski, L.M., Bulbul, A.A., Simpson, A.C., Kaya, D.E., Singh, N.K., Sezerman, U.O., Łabaj, P.P.,
691 Kosciolek, T., Venkateswaran, K.J., 2023. Adaptation to space conditions of novel bacterial
692 species isolated from the International Space Station revealed by functional gene
693 annotations and comparative genome analysis. https://doi.org/10.1101/2023.09.28.559980
694 Vijay, A., Valdes, A.M., 2022. Role of the gut microbiome in chronic diseases: a narrative review. Eur.
695 J. Clin. Nutr. 76, 489–501. https://doi.org/10.1038/s41430-021-00991-6
696 Zhu, J., Xie, H., Yang, Z., Chen, J., Yin, J., Tian, P., Wang, H., Zhao, J., Zhang, H., Lu, W., Chen, W.,
697 2023. Statistical modeling of gut microbiota for personalized health status monitoring.
698 Microbiome 11, 184. https://doi.org/10.1186/s40168-023-01614-x
699 Zou, Y., Xue, W., Luo, G., Deng, Z., Qin, P., Guo, R., Sun, H., Xia, Y., Liang, S., Dai, Y., Wan, D., Jiang, R.,
700 Su, L., Feng, Q., Jie, Z., Guo, T., Xia, Z., Liu, C., Yu, J., Lin, Y., Tang, S., Huo, G., Xu, X., Hou, Y.,
701 Liu, X., Wang, J., Yang, H., Kristiansen, K., Li, J., Jia, H., Xiao, L., 2019. 1,520 reference
702 genomes from cultivated human gut bacteria enable functional microbiome analyses. Nat.
703 Biotechnol. 37, 179–185. https://doi.org/10.1038/s41587-018-0008-8
704 Zouiouich, S., Loftfield, E., Huybrechts, I., Viallon, V., Louca, P., Vogtmann, E., Wells, P.M., Steves,
705 C.J., Herzig, K.-H., Menni, C., Jarvelin, M.-R., Sinha, R., Gunter, M.J., 2021. Markers of
706 metabolic health and gut microbiome diversity: findings from two population-based cohort

707 studies. Diabetologia 64, 1749–1759. https://doi.org/10.1007/s00125-021-05464-w

708