GMHI2: An Enhanced Predictive Index for Health Status Using Taxonomic Gut Microbiome Profiling

Daniel Chang[1], Vinod K. Gupta[2,3], Benjamin Hur[2,3], Kevin Y. Cunningham[4], and Jaeyun Sung[2,3,5]

[1]Department of Computer Science and Engineering, University of Minnesota. [2]Microbiome Program, Center for Individualized Medicine, Mayo Clinic.
[3]Division of Surgery Research, Department of Surgery, Mayo Clinic. [4]Bioinformatics and Computational Biology Program, University of Minnesota. [5]Division of Rheumatology, Department of Internal Medicine, Mayo Clinic

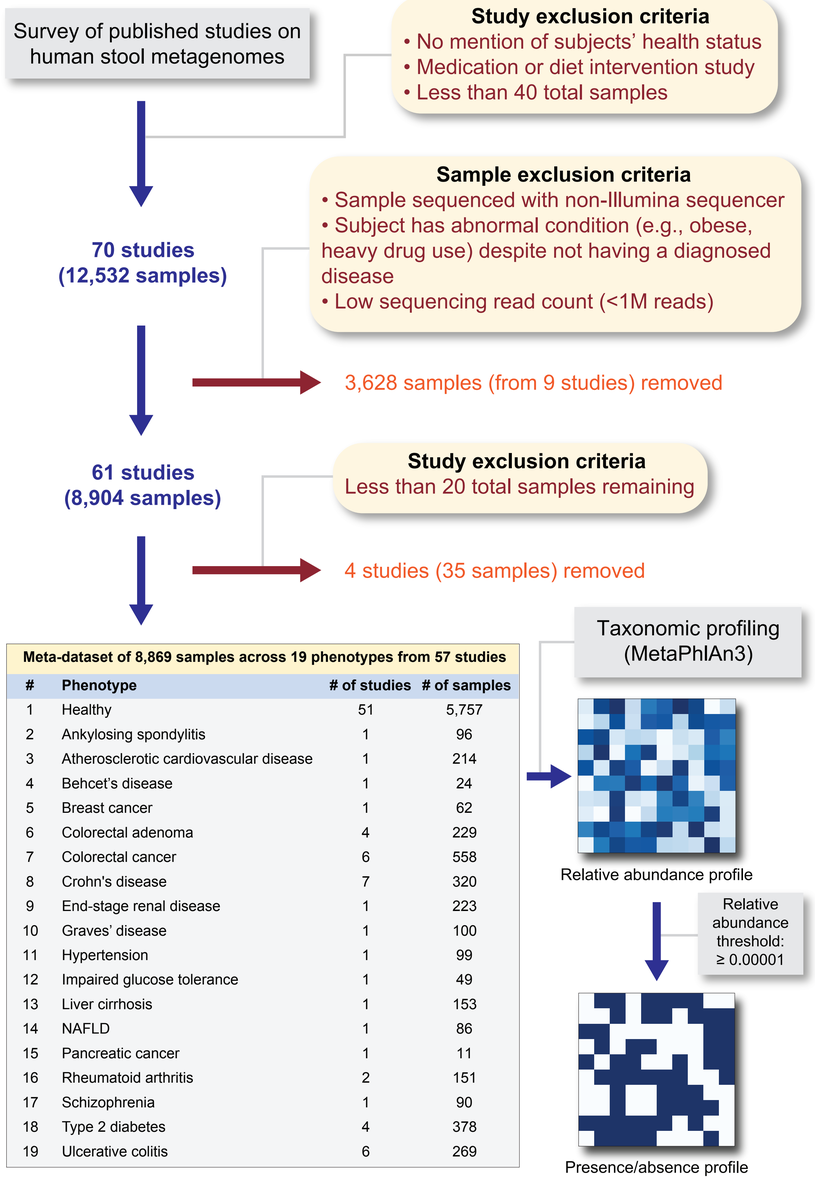
Background

To date, human gut microbiome research has given us various convincing associations and mechanistic insights into chronic diseases, as well as promising predictive tools for clinical applications. To demonstrate the utility of gut microbiome data for public health, we have recently introduced the **Gut Microbiome Health Index (GMHI)**, a stool-based indicator for monitoring the health state of one's gut microbiome. In its original version, GMHI depends on the abundances of health-prevalent and health-scarce species, which were determined using a pooled dataset of 4,347 stool shotgun metagenomic samples from 34 independent studies. Encouragingly, GMHI has already been used in many other studies focusing on identifying differences in the gut microbiome between healthy subjects and patients with a clinically diagnosed disease. In this study, using a vastly expanded pooled dataset of 8,869 stool shotgun metagenomic samples from 57 independent studies, we developed **GMHI2**, an improved version of GMHI in predicting disease presence independent of the clinical diagnosis.

Study aims

- Develop a Lasso-penalized logistic regression model (GMHI2) that predicts disease presence from gut taxonomic features
- Identify taxa associated with health and disease
- Verify that GMHI2 generalizes well to new clinical settings using inter-study validation
- Demonstrate that GMHI2 accurately quantifies intra-individual health fluctuations using gut microbiome samples from longitudinal intervention studies

Study design

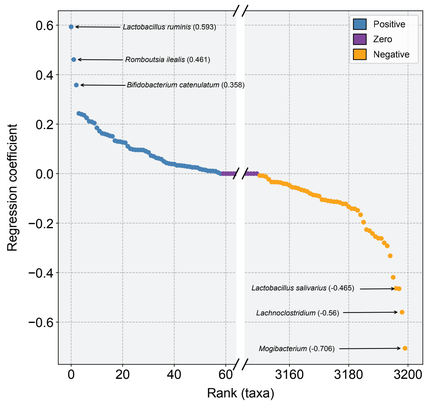


Meta-dataset of 8,869 samples across 19 phenotypes from 57 studies

#	Phenotype	# of studies	# of samples
1	Healthy	51	5,757
2	Ankylosing spondylitis	1	96
3	Atherosclerotic cardiovascular disease	1	214
4	Behcet's disease	1	24
5	Breast cancer	1	62
6	Colorectal adenoma	4	229
7	Colorectal cancer	6	558
8	Crohn's disease	7	320
9	End-stage renal disease	1	223
10	Graves' disease	1	100
11	Hypertension	1	99
12	Impaired glucose tolerance	1	49
13	Liver cirrhosis	1	153
14	NAFLD	1	86
15	Pancreatic cancer	1	11
16	Rheumatoid arthritis	2	151
17	Schizophrenia	1	90
18	Type 2 diabetes	4	378
19	Ulcerative colitis	6	269

A comprehensive survey on PubMed and Google Scholar of published studies on human stool metagenomes was conducted to create a pooled meta-dataset of 8,869 gut microbiome samples from healthy and nonhealthy individuals. The initial dataset consisted of 12,532 samples from 70 independent studies. Studies and metagenome samples were then removed based on several exclusion criterias. Finally, a total of 8,869 samples (5757 and 3112 metagenomes from healthy and nonhealthy individuals, respectively) from 57 studies ranging across healthy and 18 nonhealthy phenotypes were assembled into a meta-dataset for downstream analyses. All samples were downloaded and reprocessed uniformly using identical bioinformatics methods. Taxonomic profiling (using MetaPhlan3) and relative abundance thresholding were performed to generate a presence/absence profile for each sample. Binarizing relative abundance information simplifies many issues regarding the compositional structure of microbiome data, batch effects, and downstream model interpretation.

Lasso-penalized logistic regression extracts health and disease associated taxa

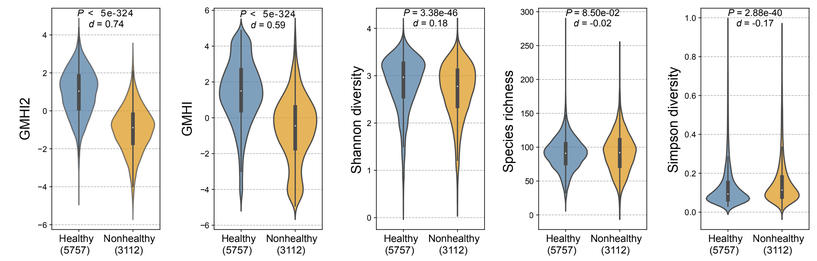


A Lasso-penalized logistic regression model was trained to predict disease presence using the entire pooled meta-dataset, yielding **59** taxa with **positive coefficients**, **3,091** taxa with **zero coefficients**, and **50** taxa with **negative coefficients**. Of the taxa with nonzero coefficients, there were 1 class, 3 orders, 7 families, 19 genera, and 79 species.

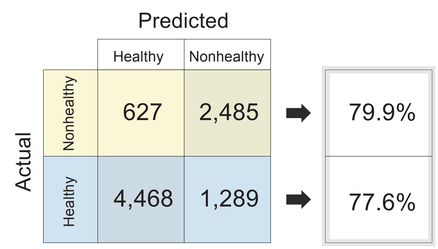
A coefficient is defined as the expected change in log-odds of a sample belonging to a healthy patient when its associated taxon changes from absent to present within a sample. Thus, taxa with positive and negative coefficients may be interpreted as health- and disease-associated, respectively.

GMHI2 stratifies healthy and nonhealthy groups stronger than α-diversity metrics

We define a sample's GMHI2 score as the trained lasso penalized logistic regression model's output (i.e., predicted log-odds of the sample belonging to a healthy patient).

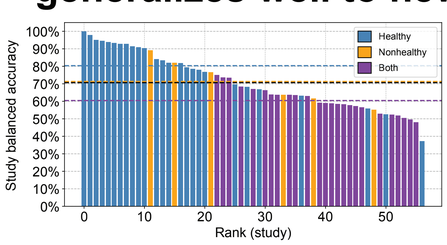


For the 8,869 samples in our meta-dataset, healthy subjects have significantly higher GMHI2 scores ($P < 5.0 \times 10^{-324}$), GMHI scores ($P < 5.0 \times 10^{-324}$), and Shannon diversity ($P = 3.38 \times 10^{-46}$), whereas nonhealthy subjects had significantly higher species Richness ($P = 8.50 \times 10^{-2}$) and Simpson diversity ($P = 2.88 \times 10^{-40}$). *P*-values are from the Mann–Whitney *U* test. The strongest effect size (Cliff's delta *d*) was seen with GMHI2.



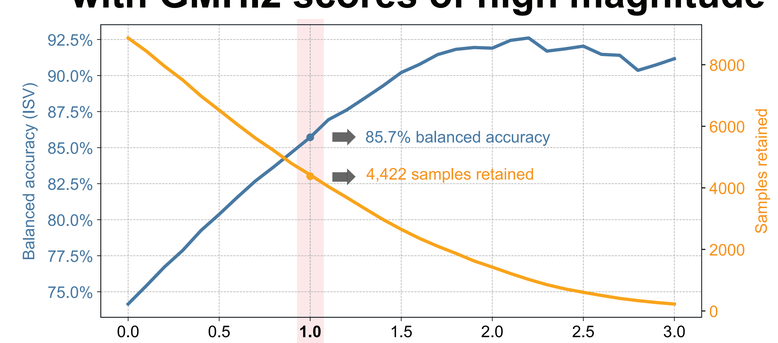
To gauge the theoretical upper limit of GMHI2's classification accuracy, the model was fit and evaluated on the entire dataset (i.e., training set = test set). GMHI2 achieved a balanced accuracy of **78.7%.** This performance is a **~9% improvement** over the original version of GMHI.

Inter-study validation reveals that GMHI2 generalizes well to new clinical settings



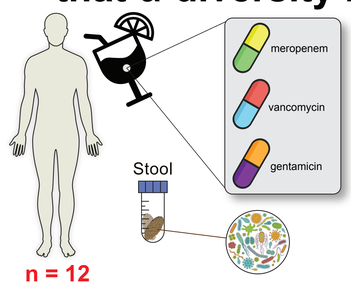
Classification performance for each held out study is displayed. Average study accuracies were: **80.4%** for studies with only **healthy samples**, **71.3%** for studies with only **nonhealthy samples**, **60.5%** for studies with **both phenotypes**, and **70.7%** for **all studies.**

Classification performance drastically improves when considering only samples with GMHI2 scores of high magnitude



GMHI2's classification performance increases dramatically with a magnitude cut-off: in inter-study validation (ISV), a balanced accuracy of **85.7%** was obtained when considering only the 4,422 samples (~50.0% of total samples) with GMHI2 scores lower than -1 or higher than 1.

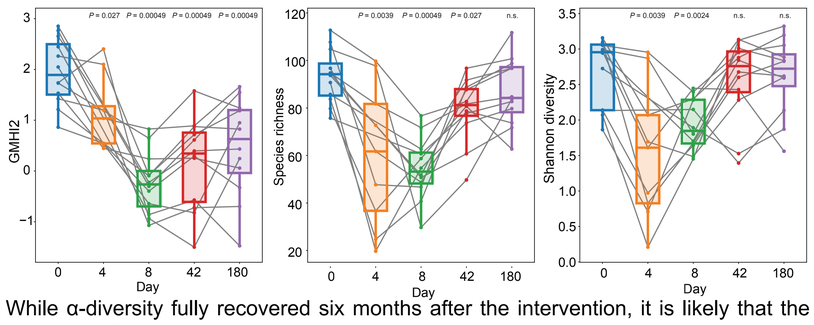
GMHI2 captures long term antibiotic effects that α-diversity metrics fail to detect



12 healthy men underwent a 4-day partial gut microbiota eradication in which they were treated with a cocktail of 3 last-resort antibiotics (**Palleja *et al*. Nature Microbiology (2018)**).

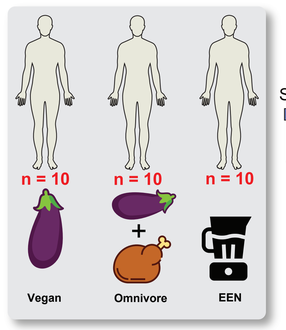
Stool samples and taxonomic profiles were collected at baseline (day 0), immediately after the intervention (day 4), and at three further timepoints (days 8, 42, and 180).

α-diversity was dramatically reduced from baseline on days 4 and 8 (Wilcoxon signed-rank test) but gradually recovered (days 42 and 180). On day 180, there was no significant difference in α-diversity between baseline. Likewise, GMHI2 scores were also dramatically reduced immediately after the treatment and gradually recovered over the next six months. However, **GMHI2 scores never fully recovered.**



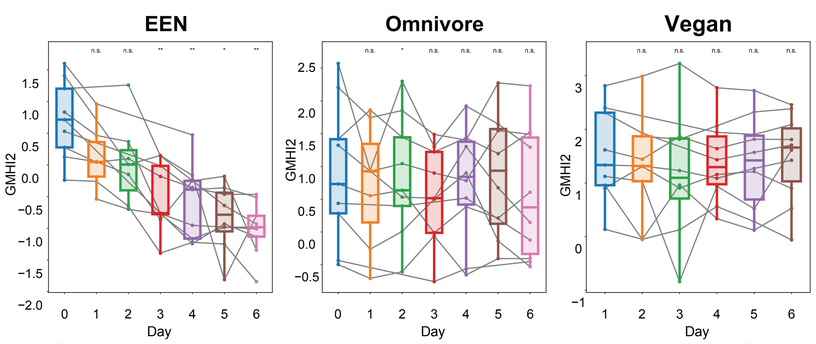
While α-diversity fully recovered six months after the intervention, it is likely that the subjects still have decreased overall gut health from baseline at this time. ***Clostridium spp.*** were undetectable pre-intervention but were significantly increased in relative abundance in all 12 subjects after day 42. Several known probiotics went extinct after treatment **(gone in all subjects and timepoints)** including: members of the genus ***Bifidobacterium*** that are considered pathogen-protective and immunostimulatory, butyrate producers such as ***Coprococcus eutactus*** and ***Eubacterium ventriosum*** (important for fiber digestion). As such, GMHI2 scores align with (hypothesized) health fluctuations and thus reveal gut health with higher granularity than α-diversity metrics.

Impacts of dietary fiber on human health are reflected in GMHI2 fluctuations



30 healthy volunteers were split into 3 groups of 10: Vegan, Omnivore, and EEN (**Tanes *et al.*, Cell Host and Microbe (2021)**). The **Vegan group** consisted of self-reported vegans, and resumed their regular diet during the study. The **EEN group** consumed a synthetic enteral nutrition diet lacking fiber. The **Omnivore group** consumed a menu of regular foods.

Stool samples and subsequent taxonomic profiles were collected at baseline and daily.



The Omnivore and Vegan groups did not experience noticable changes in GMHI2 scores. In contrast, by day 3 and after, subjects within the EEN group had significantly lower GMHI2 scores relative to baseline ($P < 0.05$, Wilcoxon signed-rank test). These results indicate that removing dietary fiber from an individual's diet may result in overall decreased gut health. In sum, our GMHI2 results further validate the well known health benefits of dietary fiber.

Conclusions

By using a vastly expanded pooled dataset and validating for both cross-study (batch) and longitudinal analyses, we introduce an enhanced predictor of health and disease based on gut microbiome taxonomic profiling (GMHI2). Our results show that GMHI2 generalizes reasonably well to new clinical settings, and potentially enables our vision of long-term stool-based health monitoring.

Acknowledgements

This work was supported by the Mayo Clinic Center for Individualized Medicine and the Translational Product Development Fund Award from the Minnesota Partnership for Biotechnology and Medical Genomics.