“Introducing the Health-Associated Core Keystone (HACK) index: A universal ranking order of gut microbial taxa for selection of next generation probiotic and biotherapeutics”

By

Abhishek Goel

Under the supervision of

Dr. Tarini Shankar Ghosh

Submitted in partial fulfillment of the requirements for the degree of Master of Technology, Computational Biology

Center for computational Biology, Indraprastha Institute of Information Technology – Delhi

May, 2024

Certificate

This is to certify that the thesis titled "*Introducing the Health-Associated Core Keystone (HACK) index: A universal ranking order of gut microbial taxa for selection of next generation probiotic and biotherapeutics*" being submitted by **Abhishek Goel** to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part of full to any other university or institute for the award of any degree/diploma.

May, 2024

Dr. Tarini Shankar Ghosh

Department of Computational Biology
Indraprastha Institute of Information Technology Delhi
New Delhi 110020

Acknowledgements

I sincerely thank Dr/ Tarini Shankar Ghosh for his unwavering guidance and mentoring throughout this journey. Without his support, this project would not have been possible. From the beginning of my thesis, he continuously challenged me to reach higher goals and excel in my work. He generously offered his time to address any issues or clarify my doubts, even during busy days. Whenever I encountered challenges in my work, he provided valuable insights and refined our approach, significantly contributing to the timely completion of my tasks. I am truly grateful for his constant encouragement and support, which have been instrumental in my academic and personal growth.

I would like to express my sincere gratitude to my parents and brother for their unwavering support, which has been invaluable throughout my work. I am thankful for their outstanding commitment, which has played a vital role in shaping my academic journey.

I also thank my senior Amit Samal, and other lab mates Lavanya, Omprakash, Shivangi, Sourav for supporting me throughout the work. Their presence and encouragement were invaluable.

I extend my heartfelt gratitude to all the faculty members and staff of the Department of Computational Biology at IIIT Delhi for their constant support and assistance throughout my college journey.

Abstract

Microbiome restoration has been considered as the potential therapeutic strategies for many diseases. As of now the focus has been done on identifying and administering specific microbes associated with health and administering either single or consortia of these microbes as probiotics or biotherapeutics, initially driven by observations of certain microbes being associated with health across multiple diseases. However, variations in baseline composition of 'healthy' microbiome across demographies and recent conflicting associations of some promising probiotics like Akkermansia muciniphila with diseases have highlighted the limitations of strictly compartmentalizing microbes as good or bad. This complicates the process of selecting microbiome therapeutic targets. In addition, most methods of selecting next generation probiotics and biotherapeutics have ignored two key properties that are required for a species to survive in an ecosystem, influence and stability.
To address, in this study we adopted a different approach where instead of compartmentalizing the taxa into good or bad, we for the first time build gut microbial species-scoring index that ranks 196 taxa, called the Health-Associated Core Keystone (HACK) index, by quantifying their association with three key hallmark properties, namely microbiome influence, stability association and health association. To develop this ranked score, we took performed a meta-analysis of publicly available 39,936 adult microbiomes (> 18 years) from 127 different study cohorts spanning 42 countries and 28 disease conditions. We also included 9,434 microbiomes to investigate the stability-association of the taxa in the gut microbiome. We identified set of 18 taxa, which we referred as the HACKs of the adult human gut microbiome, that have scores in the top 30 percentile with respect to all the three properties.

The HACK Score can potentially provide the new set of next generation probiotics. We can take the group of high HACK scoring taxa and customize them to rectify the microbiome dysbiosis. Not only it can help with finding the important gut microbes responsible to upkeep the gut health but it can also be used to grade donor microbiomes in cases of Fecal microbiome transplantation.

Contents

INTRODUCTION 6

MATERIALS AND METHODS 9

2.1 Collation of gut microbiome datasets 9

2.2 Data Processing and Generation of Taxonomic Abundance Profiles 10

2.3 Rank-scaling of values 11

2.4 Computation of the Core-Influencer scores 11

2.5 Association of Core-Influencer Scores with study-specific microbiome and host properties 12

2.6 Computation of Health-Association scores using global cohort of matched disease-control subjects covering 28 different diseased conditions 14

2.7 Computation of Taxa-specific Health-Associated Core Keystone (HACK) index 16

2.8 Functional Analysis 16

RESULTS 21

3.1 Global ranking of gut microbial taxa based on the first hallmark characteristic associated with microbiome/host health: Prevalence and Community association in non-diseased adult gut microbiomes 21

3.2 Large scale investigation of matched control-case gut microbiome study cohorts encompassing 28 disease categories enabled ranking of taxa based on a third hallmark property: Association with Health 28

3.3 Computation of the taxa-specific HACK 30

3.4 Identification of specific nutrient requirements and metabolic/functional signatures associated with high HACK-index species 31

REFERENCE 36

Chapter 1

INTRODUCTION

A key rapidly evolving area of clinical research is the development of microbiome-derived therapeutics. This class of therapeutic strategies aim to restore or replenish one's microbiome (a common target being the gut microbiome) to ameliorate diseases and improve various aspects of human health. These strategies range from external supplementation of first-generation (specific combinations of different *Lactobacilli* and *Bifidobacteria*) or next-generation (resident 'beneficial' microbiome members like *Faecalibacterium* and *Akkermansia*) probiotics, administration of prebiotics (administration of specific compounds that promote the 'beneficial' members of the gut community) or synbiotics (combinations of both) to finally, the transfer of an entire gut microbial community from a donor to a recipient using faecal microbiome transplantation[1,2]. The success of these therapeutic interventions have shown considerable variations across studies and targeted disease conditions[3,4]. Intermediate in this spectrum are supplementation strategies that involve specific microbial consortia (i.e. combinations of few gut microbiome members) that tend to constitute the core of the human gut microbiome[5,6]. While initial results of these on preclinical model are promising, their applicability in humans requires evaluation, especially in context of their inclusion of certain species like *Flavonifractor plautii*, also observed to be enriched in multiple diseases[7].

The formulation of such microbiome-restoration based therapeutics is critically dependent on identifying which members (and functional signatures) are associated with a healthy and resilient microbiome. The precision definition of these hallmarks is still one of the most intriguing aspects of human microbiome research[8]. Microbiome alterations, even for the same disease, can show large scale variations across studies and population, confounded by multiple host-associated factors like region, socio-economic status, age, and even personal/life-style-associated factors[9-11]. There are also examples of microbiome members showing context dependent association with diseases (e.g. next generation probiotic *Akkermansia muciniphila*)[12]. These observations advise against imposing microbial 'stereotypes', labelling species strictly as either 'beneficial or 'detrimental'. However, despite these conflicting associations and global variations, multiple meta-analyses have identified groups of gut microbial community members to be commonly either depleted (or enriched) across multiple diseases[7,10,13-15], indicating the presence of a shared microbiome-response common to multiple

diseases. Utilizing this information, generic microbiome health associated indices, like the Gut Microbiome Health Index (GMHI) have been proposed that aim to provide a quantitative measure of gut microbiome health by profiling the relative abundance/diversity of these specific groups of taxa[15].

An alternative approach to the problem of identifying hallmarks of microbiome health is not to assume stereotypes and compartmentalize gut microbiome members as 'good' or 'bad', but assume a ranked order with certain species showing more consistent association with gut microbiome health and stability as compared to others. These species higher up in this order, are those that can be potentially utilized/probed further as reliable next-generation probiotics or consortia for promoting general host/microbiome health. An ideal candidate for this purpose should have the following properties: it should be frequently associated with health; it should be positively linked with resilience and long-term gut microbial stability and; it should be a reasonably universal core and 'active' (being an influencer as well as influenced by others) member of a non-diseased gut microbiome. Such a universal ranking order of gut microbial taxa does not exist. But it is computable given the recent deluge of (both cross-sectional and longitudinal) gut microbiome sequence data from diverse population cohorts and spanning a wide-variety diseases.

In this study, we aimed to achieve these objectives and obtain the first ever ranking of gut microbial taxa, referred to as the 'HACK Index' (Health-Associated Core-Keystone Index). We perform a global meta-analysis encompassing 45,003 publicly available adult gut microbiomes (with host age > 18 years), including 9,434 longitudinal samples, from 123 studies (spanning 36 countries and 28 different disease conditions) and investigate 196 gut microbial taxa for their association with three quantifiable properties associated with microbiome/host health. These include prevalence/community-influence/association in non-diseased subjects, longitudinal-stability (both properties indicative of core-keystone-ness) and negative association with disease (indicative of host health association). We observe that taxa with higher HACK Index tend to show lower geography-specific variations. Using predictive machine learning, we next identify a total of 20 literature-annotated experimentally validated metabolic functionalities and a total of 871 genome-level functional categories, encompassing 53 Biosynthetic Gene Clusters (BiGGs), 22 Carbohydrate Active Gene Families (CAZymes), 257 Clusters of Orthologous Groups (COGs), 141 Enzyme Classes (E. C. Numbers), 251 KEGG Ortholog Groups, 31 KEGG Pathway Modules, 45 KEGG Reactions and 71 PFAMS, that show significant positive association with species-level HACK indices and (in

simulations) predict HACK indices of newer test species with median RMSD of 0.04 and median predicted-vs-actual Spearman Correlation of 0.64.

Chapter 2

MATERIALS AND METHODS

2.1 Collation of gut microbiome datasets

The study involved three systematic global investigations of grading the various gut microbial taxa based on their abundance associations with three key hallmark properties of microbiome health and resilience. Overall, the investigation encompassed 39,926 gut microbiomes from adult individuals (age >= 18 years) collated from 127 study cohorts covering 42 countries and six continents.

The distinct subsets of gut microbiomes (and study cohorts) included for the four different investigations are described below:

a. *Stage 1: Computation of Core-Influencer Scores for the gut microbial of 'normal' (apparently non-diseased) adult subjects across a consortium of globally spread study cohorts*: The investigation focused on a total of 18,642 gut microbiomes from apparently 'non-diseased' control subjects collected from 72 study cohorts across 42 countries (**Figure 2**). Of these, 9,736 (52%) were whole genome sequenced (WGS) microbiome datasets collected from 64 studies. The species-level taxonomic abundance profiles of 58 of these study cohorts (8,922 gut microbiomes) were directly available in the curatedMetagenomicData repository[16]. To this, we added WGS gut microbiome data from six more studies (822 gut microbiomes), corresponding to study populations from Ireland, Australia and the New Zealand[17,18,30-32]. To this collection, we further added 8,906 16S-amplicon derived gut microbiome data from eight additional cohorts from US, Japan, China, India and Columbia (including data from the AmericanGut project, the Arivale cohort)[20-23,33-36].

b. Stage 2: Computation of Stability Association Scores using a meta-analytic on diverse longitudinal study cohorts: We collected a total of 23 longitudinal gut microbiome sequence data as well as corresponding sample metadata from a total of 23 longitudinal study cohorts, conducted in different regions around the world. This combined dataset comprised 9,435 samples obtained from 2,243 individuals. Of these, 7,134 gut microbiome profiles contained at least one follow-up sample from the same subject at a subsequent time-point. The details of these sample metadata are in. The subjects constituting belonged to different clinical (or phenotypic) categories that were identified in multiple study cohorts, namely 'non-diseased' controls, Crohns' Disease, Ulcerative Colitis, Fiber intervention trials. Additionally, the

individuals suffering from difference Colitis sub-variants, (Lymphocytic Colitis, Collagenous colitis), Irritable Bowel Disease, Helminth Infections, individuals harbouring Antibiotic-resistant Enterobacteriaceae, along with antibiotic-treated individuals, fecal microbiome transplantation trials were present in single study cohorts.

c. Stage 3: Computation of health-association scores using a collection of gut microbiome datasets with matched disease-control samples spanning diverse diseases: The computation of the health-association scores for the different taxa involved an investigation of 18,020 microbiomes from 62 study cohorts covering 22 nationalities globally. The selection of this set of gut microbiomes was performed in a step-wise manner as described below.

Initially, we considered 35,447 microbiomes from 114 study cohorts. This contained a total of 14,136 from diseased subjects. These subjects (or individuals) from whom the samples were collected either affected by single or multiple clinical conditions. The metadata for the same were either obtained from the curatedMetagenomicData repository (for those datasets that were derived from this repository) or from the corresponding research papers. We observed that were many related disease conditions that could be further grouped that could be clubbed into disease categories. There were 69 such diseases that were clubbed into 37 disease categories. For each of these disease categories, we computed the study wise control and disease sample counts. Disease-Study combinations which did not contain a minimum of either 15 disease and control-matched samples were discarded. To further ensure that our results are not biased by high variations in the representation of diseased and matched-control subjects, we took only those study cohort for a given disease category for which neither the disease or control-matched samples comprised less than 20% of the total samples. This exercise yielded a final set of 18,020 gut microbiome profiles from 62 study cohorts, encompassing 28 disease categories.

2.2 Data Processing and Generation of Taxonomic Abundance Profiles

For the datasets downloaded from the curatedMetagenomicData repository, the metaphlan3 generated taxonomic profiles along with the sample metadata were already available[5]. For the rest of the whole genome sequencing datasets, the same protocol as adopted by the curatedMetagenomicData was utilized. All 16S derived gut microbiome datasets were downloaded from the corresponding data sources and processed using the SPINGO taxonomic classification tool[6]. The corresponding metadata for all the datasets that were not part of the curatedMetagenomicData, were obtained by looking into the respective research papers. We

thank the authors of the Arivale cohort for providing us with the metadata of the corresponding participants.

2.3 Rank-scaling of values

A key component of the computation of the different association scores was the rank-scaling function utilized here. For a given set x of length k, denoted by $x = \{x_1, x_2, x_3, x_4, \ldots, x_k\}$, the values are first ranked. Next, the corresponding rank-scaled values, denoted by $x^R = \{x_1^R, x_2^R, x_3^R, x_4^R, \ldots, x_k^R\}$, of this set were obtained as:

$x_i^R = [\text{Rank}(x_i) - \text{Minimum}(\text{Rank}(x))] / [\text{Maximum}(\text{Rank}(x_i)) - \text{Minimum}(\text{Rank}(x))]$

The transformation was essentially an approach of ranking the values and then converting the ranks to values between 0 and 1.

2.4 Computation of the Core-Influencer scores

The Core-Influencer scores for the different taxa were computed simply by counting the number of the cohorts a given taxa was identified as a Core-Keystone and subsequently rank-scaling these numbers obtained for all the taxa (as described above) to values from 0 and 1.

Identification of Core-Keystones for a given cohort: For a given study cohort 'Y', a Core-Keystone was identified as a taxon that satisfied the following two criteria:

- Prevalence of >= 70% (i.e. detected or present in at least 70% of the samples or gut microbiomes of Y)
- Community-Influence Rank >= 70%. The procedure to compute the Community-Influence Rank as well as the approach used to select the threshold of 70% is described below.

For each taxa, the absolute Community-Influence was first calculated using the Remove-Renormalize-Relate approach (3R) as described below.

Computation of community influence using the Remove-Renormalize-Relate (3R) approach:

The Remove-Renormalize-Relate (3R) approach in Figure 3 Briefly, for a given study cohort A, to measure the community influence of a given taxon (X), we first removed the abundances of X from all microbiomes belonging to cohort and recomputed the abundance profiles of the microbiome by renormalizing the abundances of the remaining taxa. The abundances of all remaining taxa were then renormalized and the community variations of the resulting microbiome profiles were then computed using Principal Coordinate Analysis using the Bray-Curtis distance matrix. The PCoA of the 3R approach was performed using 'dudi.pco' function of the ade4 package v1.7.22 of the R-programming framework. The input was the

Bray-Curtis Distance matrix containing the distances between each microbiome-pair provided as the input. The Bray-Curtis distances were computed using the 'vegdist' function of the vegan R package v2.6.4.

The removal of the test taxon X during the computation of community-wide variations ensured that the estimated variations in community composition were not simply due to the large variations in the relative abundances of X, but were instead solely based on the differences in the mutual representation of remaining microbiome members. The community-influence of X was then computed by inter-relating the abundance of X with the community-wide variations calculated (after the 'removal' of X) as described above. The envfit approach, encoded in the function with the same name in the vegan package as described above, was utilized for this purpose. The method provides two outputs: an R-squared value, indicating the extent of community- influence the abundance of X has on the ordination scores (depicting the absolute effect the abundance of X has on the community-wide variations of the remaining taxa); a p-value is the probability of observing the R-squared value by random chance (being computed using a permutation test) and indicative of the significance of the effect. Taxa were then rank-scaled (as described above) based on their community-influence (R-squared value obtained using envfit).

The 3R approach is derived from on the recently published Empirical Presence-Absence Interrelation (EPI) approach[24], with exception that it is based on abundance variation rather than the EPI's presence-absence information. We did not adopt the presence-absence based approach as the core-influencers are expected to be prevalent, classifying microbiomes simply based on the presence-absence of these members may result in a bias in the number of microbiomes belonging to different groups. Furthermore, a taxon is also likely to have variations in its community-influence depending on how abundant is it in a microbiome.

2.5 Association of Core-Influencer Scores with study-specific microbiome and host properties

Abundance: In this investigation, we checked and ensured that the detection as Core-Influencers was not a simple reflection of abundances of the identified species (i.e. only the highly abundant species being identified as Core-Influencers). For this purpose, we computed the study-specific mean abundance of each species and correlated the same with their study-specific community-influence (rank-scaled R-squared of the envfit) and prevalence using Pearson correlations.

Cohort Life-Style and Cohort Sequencing-Type: All gut microbiomes comprising the study cohorts were derived from individuals in the age-range >= 18 years. Individual microbiomes had a mix of different host demographics like age, gender, BMI etc. Some of the studies had a mix of nationalities. However, there were certain properties that were especially cohort-specific and could be used to broadly categorize almost all cohorts. These were the microbiome profiling strategy (referred to here as 'SequencingType') (amplicon-based or 16S-based; whole genome sequencing or WGS-based) and the life-style of the study populations (referred to here as 'CohortLifeStyle'). For the later, we checked the individual research papers pertaining to each study and subsequently categorized the life-style pattern of the study populations into three groups, namely Industrialized-Urban, Urban-Rural-Mixed and Rural-Tribal.As the names suggest, the first category was "IndustrializedUrban" (consisting of studies from Europe, North America and East Asia, where a majority of individuals reside in Industrialized Societies). The exceptions to these were the cohorts of Irish Traveller from Ireland (KeohaneD_2020) and a large Chinese population cohort (He_et_al), which have investigated individuals from mixed urban/settled and rural/nomadic life-styles[9,38]. These two studies, along with those focusing on populations from India, Fiji, Native New Zealanders, Pacific Islanders, and other South American non-nomadic cohorts, having a mixture of both Urban and Rural sub-cohorts, were placed into a second category called 'UrbanRuralMixed'. The third category consisted of nomadic/semi-nomadic or settled-rural cohorts from Africa, Peru and Mongolia and was tagged as 'RuralTribal'. Investigating with the SequencingType and CohortLifeStyle factors on the properties (prevalence, community-influence) of the gut microbiome members as well as the detection pattern of Core-Keystone taxa across cohorts was performed as described below.

We first performed study-level PCoAs (Euclidean distance for Prevalence and Community-Influence Rank; Jaccard Distance for Core-Keystone Detection). Subsequently, for each property, pairwise distances were computed between the studies by computing the Euclidean distances between the corresponding Principal Coordinates. Subsequently, by using the Permutational Multi-Variate Analysis of Variance (PERMANOVA) approach, we investigated if these observed variations between studies could be explained by their 'SequencingType' or the 'CohortLifeStyle' of the study populations. PERMANOVA was performed using the adonis2 function of the vegan R package (described above).

2.6 Computation of Health-Association scores using global cohort of matched disease-control subjects covering 28 different diseased conditions

As for the previous steps, we specifically investigated the same 196 species as used for the computation of the Core-Influencer and Stability-Association scores. The selection of the study cohorts and the 28 disease conditions have been described previously in this methods section. As in the previous section, the Health-Association scores were computed using a similar bootstrapped iterative approach as adopted for the Stability-Association scores. Here also, we performed 10 iterations, each time selecting 18 (of the 28) disease/clinical conditions (**Figure 1**).

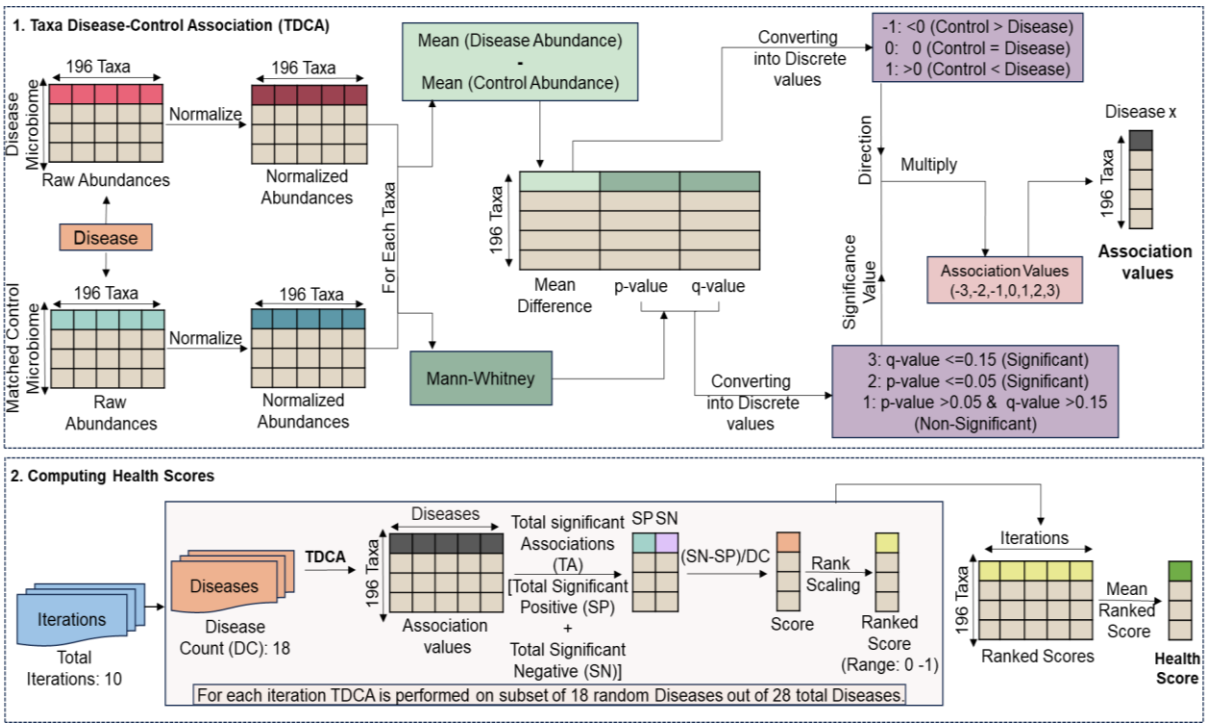

Figure 1

Within each iteration for each disease and a given taxon, we determined whether the taxon was associated with health or disease by comparing its abundance in control samples with respect to the diseased. For a given disease, we combined all study cohorts corresponding to the disease (based on the criteria described above). The abundance comparison was performed using a two-step approach. First, we checked the directionality of the change (i.e. whether the bacteria is more abundant in disease or in control-matched samples). Directionality was computed as either 1 (Mean Abundance in Disease > Mean Abundance in Controls) or -1 (Mean Abundance in Controls > Mean Abundance in Disease) or 0 (No Difference). Next, we checked what is the significance of the difference in the abundance of the species between

control and disease samples. For this purpose, we performed Mann-Whitney test using the disease and matched-control samples. The p-values were obtained and were also subsequently corrected using the Benjamini-Hochberg method ("fdr" method parameter provided to the p.adjust function in R) to obtain the Q-values. Taxa were assigned to different significance tags based on their significance of their associations. The taxa with Q-values <= 0.15 was assigned tag of 3; those with Q-value > 0.15 but p-value <= 0.05 were categorized as 2; the rest were assigned tag 1. Finally, the Association values for each taxon for each of the 18 diseases were computed as a product of directionality (-1, 0, 1) and significance tags (3 or 2 or 1) (**Figure 1**). Consequently, all taxa-disease associations were categorized as one of -3, -2, -1, 0, 1, 2, 3. Negative numbers indicated taxon was more abundant in control samples than in disease samples and vice versa. After the computation for each of the 196 taxa across the 18 diseases considered for the iteration, the iteration-specific unranked health-association scores were computed for each taxa as:

[Total significant Negative association values (i.e. total number of -3s) – Total significant Positive association values (i.e. total number of +3s)] / [Total number of -3s and +3s]

These unranked health-association scores obtained for the 196 species were the rank-scaled from 0 to 1 (using the same formula as described above) to obtain the iteration-specific health-association scores. The health-association scores for each taxon obtained for the 10 iterations were then averaged to obtain the final taxon-specific Health-Association score.

Identification of consistent disease-positive and disease-negative taxa: For this purpose, we also analysed all the 28 clinical conditions together in a single analysis using the same approach. We deduced the set of species that were negatively (-3) and positively (+3) associated with disease. Taxa falling into either the disease-positive and disease-negative categories were those that were significantly (association values of -3 and +3) associated with at least 30% of the total diseases and for whom the total disease-positive (association values of +3) or disease-negative (association values of -3) constitute at least 70% of the total significant associations (association values of -3 and +3). Using the above method we identified 45 disease-negative species and 35 disease-positive taxa.

2.7 Computation of Taxa-specific Health-Associated Core Keystone (HACK) index

Taxa-specific HACK index were envisioned as a combinatorial index that combines the association scores obtained for a given taxa with respect to the three properties (core-influence, stability-association and health-association) into a single measure. HACK index combines three scores (computed as previously described) as a product of two values:

- *Average Score*: This is the mean of the three scores (Core Influence: CS, Stability-Association: SS, Health-Association: HS). Higher mean values indicate higher overall association with the three properties. However higher mean values may also originate because of one of the scores showing an especially high value where as others don't. To address, we calculated the second score called 'Reward Score'
- *Reward Score*: Reward Score simply indicates how close are the values of the three scores with respect to each other. This was obtained by first computing GINI index across the three values. The GINI index is a measure of the extent to which a given set of values deviate from a perfectly equal distribution. The more similar are the values of a set, the closer will be the GINI index to 0. Thus, GINI index is an inverse measure of similarity within a set and varies from 0 to 1.Consequently, in this case, the Reward Score which measured the extent of similarity was thus calculated simply as: 1 – GINI Index of the three scores. Higher Reward Scores for a taxa indicates higher similarity of the CS, SS and HS. To compute this, we used the Gini function of the DescTools package (v0.99.50) in R

The HACK index for a given taxa 'I' was thus computed as

$$\text{HACK-Index}^{I} = \text{Average-Score}^{I} \text{ X } \text{Reward-Score}^{I}$$

2.8 Functional Analysis

- Sequence Data Collection: Towards the goal of capturing the functional insights of the core human gut microbiome, around 65,000 genomes (including Metagenome assembled genomes) encompassing 121 different core gut microbial species were curated from globally renowned open-source and comprehensive databases like EMBL-EBI MGnify and UNINA MAGs Genome Repository. Among these huge number of genomes, 31913 genomes which are of high quality (with greater than 90 percent completeness and less than 5 percent contamination) were taken for building the framework.

- Genome Annotation: All these high-quality genomes were thoroughly annotated against eggNOG Database 5.0.2 using the latest eggNOG-mapper pipeline (version 2.1.11), proteomes corresponding to the genomes were predicted using Prodigal (version 2) which is incorporated inside the eggNOG-mapper pipeline. The annotation files have been leveraged to build the functional profile matrices of the genomes by counting the frequency of occurrences of each functional features in a genome using an R script. The quantitative presence of functional signatures has been captured through these matrices. Profiles of functional features have been generated for eight different categories, those are CAZy (carbohydrate-active enzymes), COG (cluster of orthologous groups), BiGG (Biosynthetic gene clusters), KEGG Orthologs, KEGG Reactions, KEGG Modules, EC number and PFAMs.
- Functional Profiles of Unique Species: From the prior sections it is obvious that multiple genomes have been taken into account for every unique core gut microbial species. In order to synthesize a profile matrix for 121 core human gut microbial species from the profile matrix of 31,913 genomes and to assign a single row for each representative species, the arithmetic mean values were computed for each functional feature across all the genomes present for a particular species, thus condensing the multivariate information into a singular summary row representing the profile of a unique species. Thus, profile matrices for eight different functional categories as mentioned above have been generated covering 121 core species.
-
- Previous study has experimentally found the consumption and production of the 361metabolites for the 120 species. Out of these 120, we had an overlap of 101 species with our highly prevalent 196 species. We performed random forest on the metabolite profile to predict the association of the features with the HACK Score. To get the important features from random forest model we took the ratio of the cumulative importance of the feature and the maximum cumulative importance >=0.05, resulting in a set of 96 metabolite features. Finally, we took metabolite profile of 96 metabolites to compute the metabolite activities associated with HACKs and Non-HACKs. To achieve our goal, we performed logistic regression keeping the HACK Score as the target variable. We found 14 (Galactomannan Consumption, Glycogen Consumption, Xylan Consumption, Laminarin Consumption, D Glucuronic Acid Consumption, D Galacturonate Consumption, Acetate Consumption, Isobutyrate Production, Inulin Consumption, Succinate Production, NH3 Consumption, FOS Consumption,

Glucosamine Consumption, Butyrate Production) metabolite functions and 8 metabolite functions (H2O2 Production, Urea Consumption, Pantothenic Consumption, Trehalose Consumption, NH3 Production, Ethanol Production, L Lactate Production, Sucrose Consumption) that were positively and negatively associated with the HACK Score. Later, we applied random forest model on these 22 features to predict the HACK Score. Interestingly, even as small as 22 features resulted in the Pearson correlation of 0.54 between predicted and actual value of the HACK Score.

- Given the high dimensionality of the data consisting of 47,621 (Pfam: 9,807; BiGG: 10,095; cazy: 118, EC: 2,969; COG: 9,370; KEGG Orthologs: 9,617; KEGG Module: 717; KEGG Reaction: 4,928), the first step was to find the reduced set of important functional features. Out of 196 species, we proceeded with only 121 species due to absence of functional profiles of remaining one. To get the subset of important features from 47,621 features we followed multiple steps. We applied random forest on each of the 8 functional profiles keeping the HACK Score as target variable. To get the important features from random forest model, we took cumulative importance of features. Those features were deemed as important whose ratio with the maximum cumulative value was greater than or equal to 0.05. Combination of the important features from different functional profiles generated a set of 6,782 important features (Pfam: 1,674; BiGG: 373; cazy: 54, EC: 750; COG: 1,386; KEGG Orthologs: 1,330; KEGG Module: 284; KEGG Reaction: 931). Our next target was to identify the set of features that were associated with the HACK Species. We deduced such features using two approaches where one was based on frequency count of the features in HACK and Non-HACK species and another was based on the gaussian model of the logistic regression. In first approach, we determined the directionality of the feature association with HACK and Non-HACK by taking the difference of Mean feature count of HACKs from Non-HACKs.
- Mean Feature frequency count in Non-HACKs- Mean Feature Abundance in HACKs
- Then we investigated whether there was significant difference in the feature frequency count of the HACKs and Non-HACKs using the Mann-Whitney test. We opted for only those features (582) that were significantly negatively associated (Q-value<=0.15 and P-value <=0.05) with Non-HACKs. As a different approach to divide the features into HACKs and Non-HACKs by keeping the continuous disposition of the functional profiles we used logistic regression. We shortlisted a set of 584 features that were significantly (Q-value<=0.05) positively associated with the HACK Score. By taking

the union of features from the two approaches we got 912 features that were potentially associated with the HACK Score. These features were used to develop the protein database using the proteome information of 18 HACK species.

- TMHMM 2.0 and SignalP 6.0: In the investigation of 196 core species regarded for their resilience, 121 species demonstrated matches within the microbiome repository. Specifically, 101 of these species were found to align with UNINA MAGs Genome Repository, which hosts a total of 39,592 high quality genomes (completeness > 90% & contamination < 5%). Using this rigorous filtering method, we narrowed down an initial set of 31831 genomes to identify the top 101 genomes. This selection was made by carefully ranking them and choosing the best one based on specific criteria. Our emphasis was on genome completeness, ensuring it exceeded 90%, and keeping contamination levels below 5%. The remaining 20 species underwent mapping using the EMBL-MGnify Database, followed by additional predictions conducted for the genomes of 121 species. TMHMM is a highly accurate membrane protein prediction method rooted in a hidden Markov model. It reliably identifies transmembrane helices, distinguishing between soluble and membrane proteins with over 99% accuracy. TMHMM version 2.0 was utilized to forecast transmembrane proteins, gaining insights into their cellular localization. The results analyze the distribution of proteins within and outside cell membranes, as well as across them. We've extracted transmembrane fasta files from the source proteins of the respective genome species. Our focus extended to proteins prevalent in extracellular spaces, as they have a prevalence of signal peptides, integral for extracellular protein identification. SignalP 6.0 predicts signal peptides and their cleavage sites in various life domains, distinguishing different types in Bacteria and Archaea. We opted for SignalP 6.0's slow mode due to its enhanced accuracy in computing precise region borders, crucial for our need to accurately delineate signal peptides, particularly in Bacteria and Archaea. A specialized pipeline was developed, employing TMHMM 2.0 to initially predict transmembrane and extracellular regions within proteins. Subsequently, extending the output regarding extracellular proteins, this information was utilized as input for SignalP 6.0, facilitating the investigation and identification of signal peptides. From our comprehensive results, we've specifically isolated the transmembrane and signal protein fasta files for the 18 representative genomes associated with the 18 HACK species. This comprehensive approach aimed to elucidate pivotal proteins crucial for species resilience, specifically focusing on their roles within cellular membranes.

- Finally, to find HACK tags (215 Features) from the HACK associated features we considered only those features that were detected in at least 25% of the HACK species and whose detection rate in HACK Species was at least twice as much as the Non-HACK detection rate.

Chapter 3

RESULTS

3.1 Global ranking of gut microbial taxa based on the first hallmark characteristic associated with microbiome/host health: Prevalence and Community association in non-diseased adult gut microbiomes

A key characteristic for a microbial determinant of microbiome/host health/stability is that it should be a prevalent member and have high community-influence/association in the gut microbiome in non-diseased individuals. We referred to this property, as Core-Influence ('Core' as these members are more prevalent; 'Influence' as they are likely to have strong association with the composition of the other members of the gut microbial community). Supplementation of such members will have a higher likelihood towards transitioning into a normal-like microbiome. Thus, we started this investigation by first investigating the taxa for their association with this specific property.

For this and the subsequent investigations, we specifically selected a list of 196 species-level taxa that were detected in at least 10% of samples in at least 35% of the study cohorts ('Gut-Associated'). Cumulatively, these taxa accounted for a median abundance of 97.7% across all 72 studies (with only two studies having a median cumulative abundance < 75%) (**Figure 2**). The collated repository of 39,996 gut microbiomes contained data from primarily three kinds of datasets: population level studies on apparently non-diseased (or 'control') subjects; longitudinal study cohorts collecting samples from subjects at multiple time-points and; matched disease-control cohorts focusing on different diseases.

Figure 2: Cumulated abundance of 196 gut associated taxa.

For this specific analysis, we utilized a repository of 18,642 gut microbiomes from apparently non-diseased (or control) adult subjects (age >= 18 years) collected from 72 studies. These studies covered 34 nationalities across six continents the world (Figure 2; See and Methods for a detailed protocol). These included taxa abundance profiles derived from both shotgun sequencing (N=9.736)16-18, and 16S rRNA amplicon sequencing projects (N=8,906) 9,19-23. This large collation of gut microbiome profiles from control study populations encompassed diverse life-style patterns ranging from industrialized/urban populations of Europe, North America and East Asia to rural and hunter-gatherers from Africa, India and South America; from population cohorts with mix of rural and urban life-styles from India, China, Columbia, Fiji, El-Salvador, Mongolian and Oceania to populations with transitional life-styles like the Irish Travellers

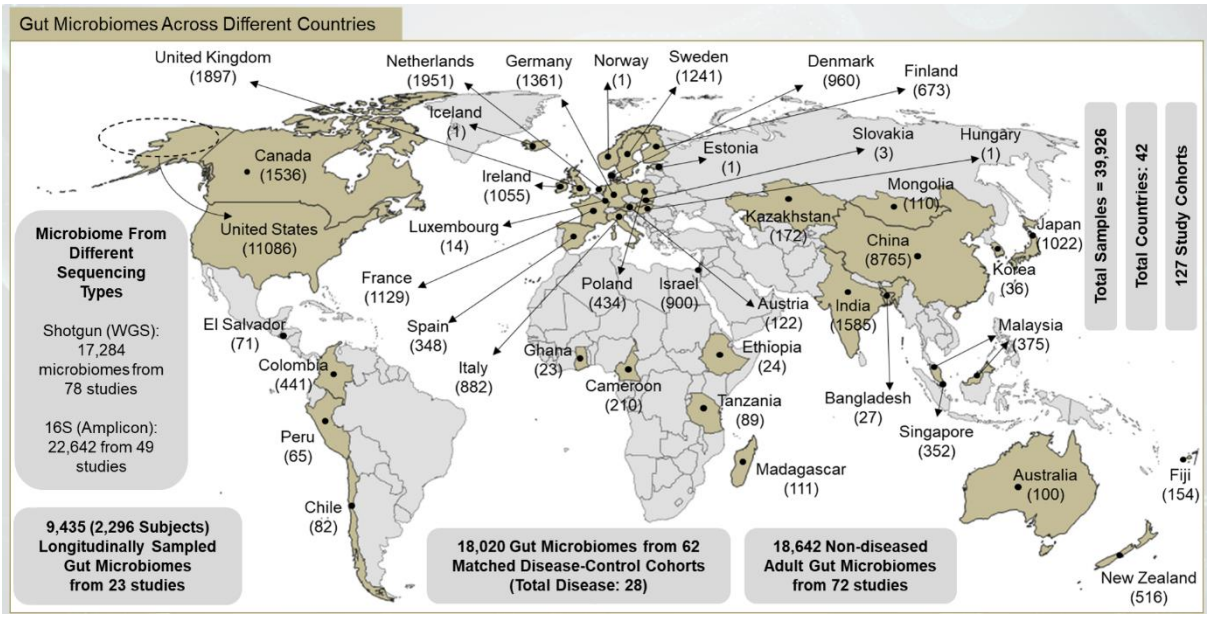

Figure 2: Data collation

The next step was the identification of the core and influential members of the gut microbiome individually in each of the study cohorts. Both of these properties (prevalence and community-association) are quantifiable. While obtaining the prevalence of each taxon is trivial, recently developed Empirical Presence-absence Interrelation (EPI) approach enables quantifiable prediction of the degree of community influence of each microbiome member by measuring the community-wide differences in microbiomes in the presence/absence of the taxon using various distance measures[24]. However, given that high prevalence was an essential characteristic of core members (which was one of the key attributes to be investigated here), classifying microbiomes simply based on the presence-absence of these members may result in a bias in the number of microbiomes belonging to different groups. Consequently, we utilized an alternate modified version of the EPI approach that utilizes the abundance of a taxon rather than its presence-absence information. We referred to this as the Remove-Renormalize-Relate (3R) approach (**Figure 3; See Methods**). The basic premise of this method was to compute the community influence (or association) of a given taxon X (for a given study) by correlating its abundance variation with the community variations of all other taxa in gut microbiomes of the study. The key step was to remove the taxon from the community abundance profiles and renormalize the profiles prior to computation of community variations. This ensured that the estimated variations in community composition were not simply due to the large variations in the relative abundances of X, but were instead solely based on the differences in the mutual representation of remaining microbiome members. For each study, the above steps were

repeated for all 196 gut-associated taxa (**see above**). Taxa showing significant associations (using the 3R approach; **see Methods and Figure 3**) were then ranked based on their extent of community influence. Taxa ranked in the top 70% and also detected in greater than >= 70% samples of the given study cohort were identified as the Core-Influencer -Taxa of that study cohort. This process was repeated for all study cohorts and taxa were then again ranked based on their consistency of detection as Core-Influencer-Taxa across the 72 study cohorts. This produced a first ranked order of taxa with respect to the first probed hallmark of microbiome health, that of high Prevalence and high Community Association in non-diseased subjects .

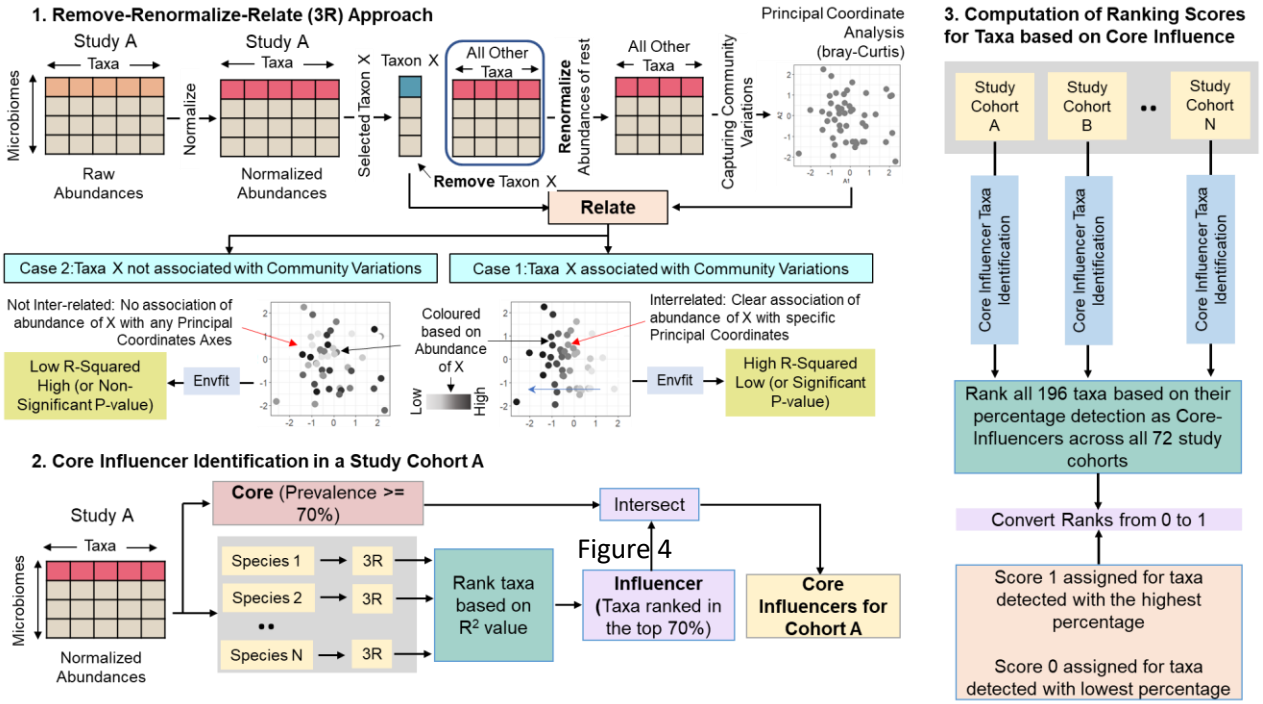

Figure 3: RRR approach

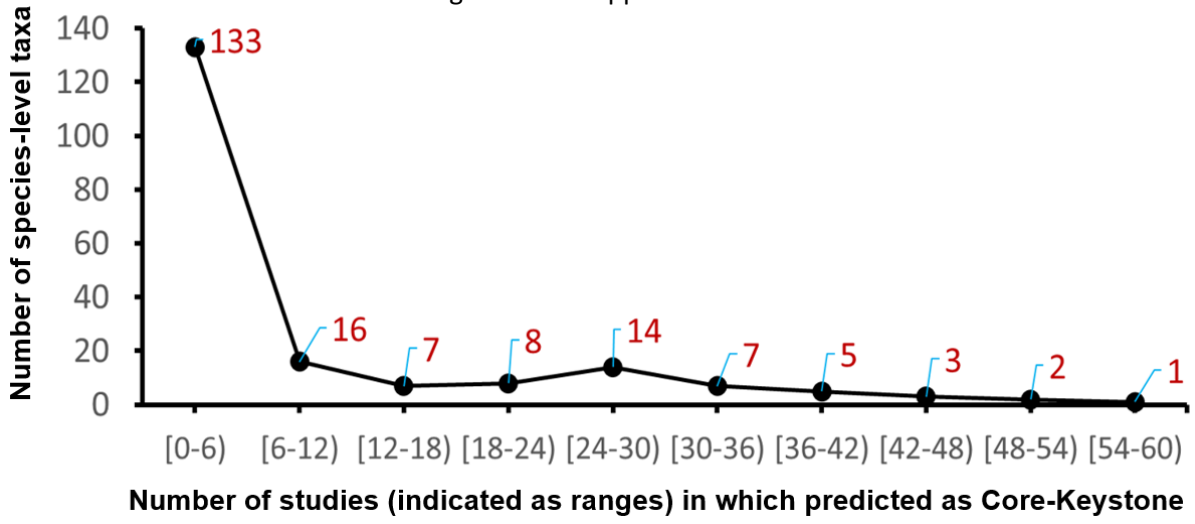

Figure 4

A large majority (76%) of the species (149) did not show any strong Core-Influencer properties (i.e. not detected as Core-Influencer-Taxa in > 80% of studies). The rest (47) formed a gradient with increasing consistency of prediction across cohorts (**Figure 4**). These 47 taxa comprised greater than 70% of the list of Core-Influencer-Taxa identified in greater than 80% of the study cohorts. This indicates that Core-Influencer property is restricted to a small subset of the gut microbiome. The top 5 ranked taxa (in terms of their detection consistency as Core-Interactive-Taxa) consisted of two Bacteroides species (*B. uniformis*, *B. vulgatus*), followed by *Faecalibacterium prausnitzii*, *Fusicatenibacter saccharivorans* and *Dorea longicatena*. While many of these taxa have been associated with health in prior studies[7,13], others (especially some members of the *Bacteroides*/*Parabacteroides*, *Faecalibacterium*, *Roseburia, Barnesiella, Prevotella* and *Bifidobacterium* genera) have been associated with faster gut microbiome recovery post-antibiotic treatment and resilient donor microbiome colonization post faecal microbiome transplantation[25,26]. Notably, the list of 47 top predicted Core-Influencer-Taxa also included members like *Flavonifractor plautii* and *Collinsella aerofaciens*, previously shown to have consistent enrichment across diseases in general[7]. Thus, the Core-Influencer-Taxa while being associated with high community influence and prevalence in apparently "healthy" microbiome, may not be directly linked to health.

Furthermore, in only 21 (29%) and 34 (48%) studies, there was a significant positive correlation between the study-specific mean abundance of taxa and their community-influence and prevalence, respectively (**Figure 5**). This indicates that the Core-Influencer property is not directly related to species abundance. In other words, simply being highly abundant does not necessarily qualify a taxon to be a Core-Influencer-Taxa of a normal gut microbiome. As expected, investigating the variation in the detection patterns of Core-Influencer-Taxa across different study cohorts revealed significant association with the life-style pattern of the study cohorts ('Industrialized-Urban', 'Urban-Rural-Mixed' or 'Rural-Tribal') (**Figure 7**). However, taxa detected with higher consistencies across studies showed significant lower Cohort-Lifestyle-specific variability (**Figure 6**), i.e the consistently identified Core-Influencer-Taxa are more likely to be universal.

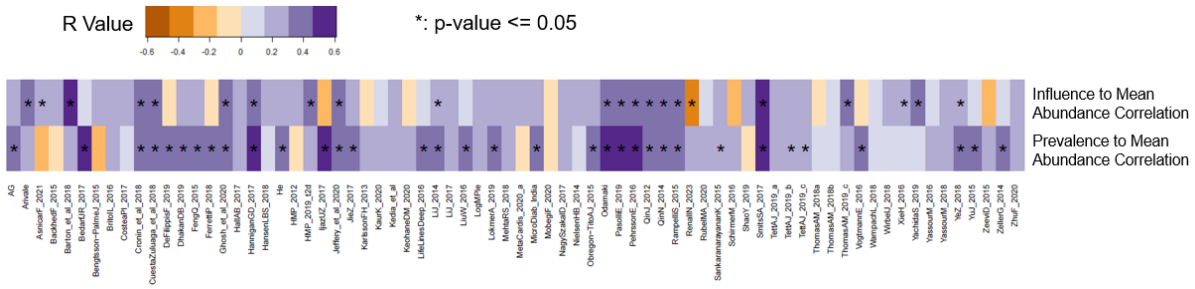

Figure 5

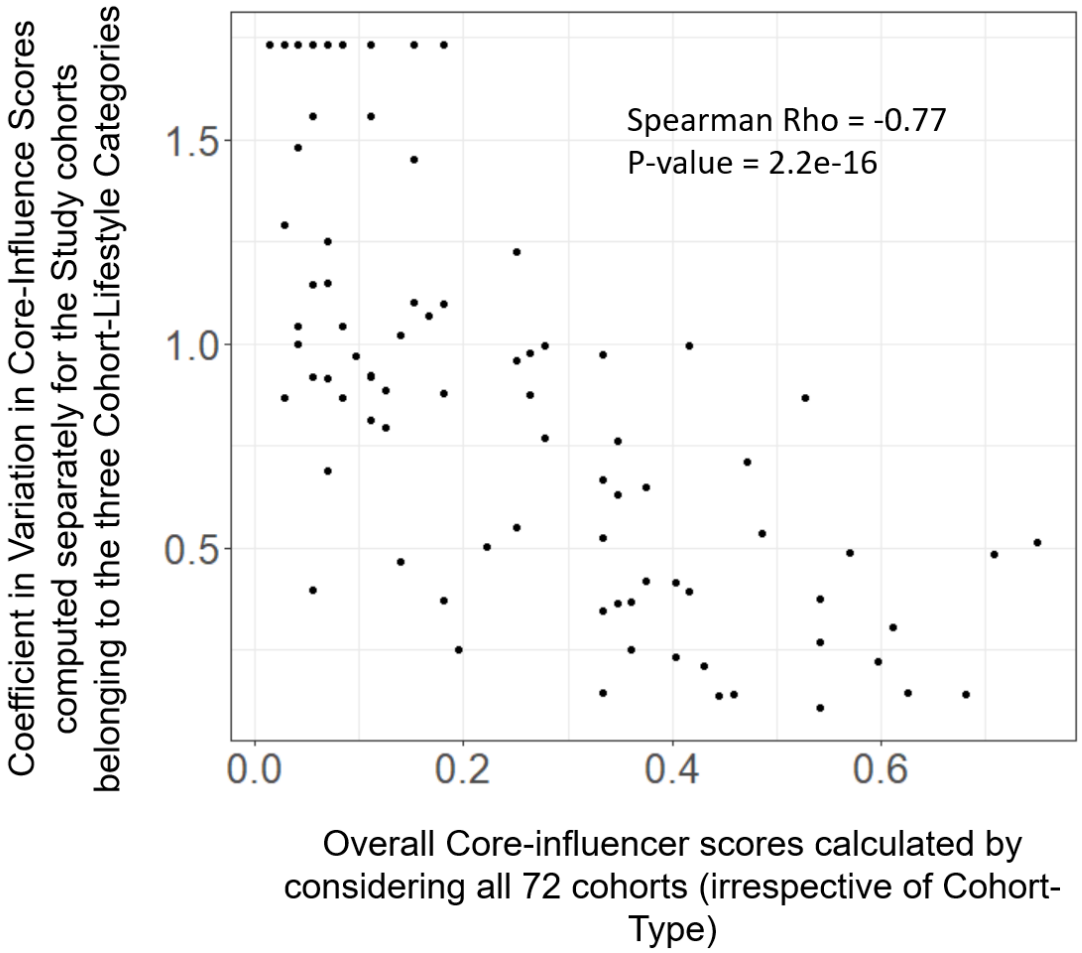

Figure 6

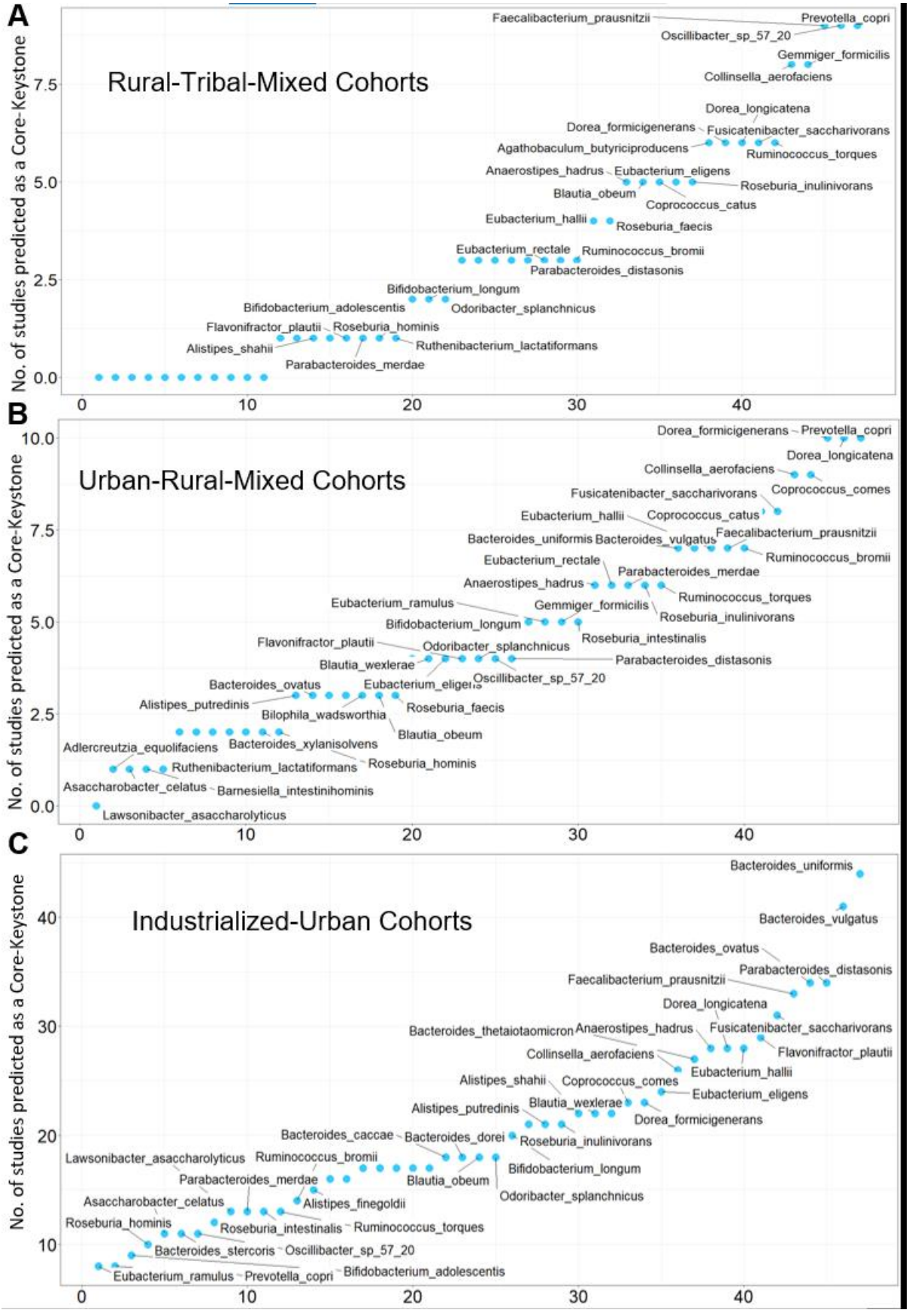
A
Rural-Tribal-Mixed Cohorts
No. of studies predicted as a Core-Keystone
Faecalibacterium_prausnitzii
Prevotella_copri
Oscillibacter_sp_57_20
Gemmiger_formicilis
Collinsella_aerofaciens
Dorea_longicatena
Dorea_formicigenerans
Fusicatenibacter_saccharivorans
Agathobaculum_butyriciproducens
Ruminococcus_torques
Anaerostipes_hadrus
Eubacterium_eligens
Roseburia_inulinivorans
Blautia_obeum
Coprococcus_catus
Eubacterium_hallii
Roseburia_faecis
Eubacterium_rectale
Ruminococcus_bromii
Parabacteroides_distasonis
Bifidobacterium_longum
Bifidobacterium_adolescentis
Odoribacter_splanchnicus
Flavonifractor_plautii
Roseburia_hominis
Alistipes_shahii
Ruthenibacterium_lactatiformans
Parabacteroides_merdae
B
Urban-Rural-Mixed Cohorts
No. of studies predicted as a Core-Keystone
Dorea_formicigenerans
Prevotella_copri
Dorea_longicatena
Collinsella_aerofaciens
Coprococcus_comes
Fusicatenibacter_saccharivorans
Eubacterium_hallii
Coprococcus_catus
Bacteroides_uniformis
Bacteroides_vulgatus
Faecalibacterium_prausnitzii
Ruminococcus_bromii
Eubacterium_rectale
Parabacteroides_merdae
Anaerostipes_hadrus
Ruminococcus_torques
Eubacterium_ramulus
Gemmiger_formicilis
Roseburia_inulinivorans
Bifidobacterium_longum
Roseburia_intestinalis
Flavonifractor_plautii
Odoribacter_splanchnicus
Parabacteroides_distasonis
Blautia_wexlerae
Oscillibacter_sp_57_20
Bacteroides_ovatus
Eubacterium_eligens
Alistipes_putredinis
Roseburia_faecis
Bilophila_wadsworthia
Blautia_obeum
Adlercreutzia_equolifaciens
Bacteroides_xylanisolvens
Roseburia_hominis
Ruthenibacterium_lactatiformans
Asaccharobacter_celatus
Barnesiella_intestinihominis
Lawsonibacter_asaccharolyticus
C
Industrialized-Urban Cohorts
No. of studies predicted as a Core-Keystone
Bacteroides_uniformis
Bacteroides_vulgatus
Bacteroides_ovatus
Parabacteroides_distasonis
Faecalibacterium_prausnitzii
Dorea_longicatena
Bacteroides_thetaiotaomicron
Anaerostipes_hadrus
Fusicatenibacter_saccharivorans
Collinsella_aerofaciens
Flavonifractor_plautii
Eubacterium_hallii
Alistipes_shahii
Coprococcus_comes
Eubacterium_eligens
Alistipes_putredinis
Blautia_wexlerae
Dorea_formicigenerans
Bacteroides_caccae
Bacteroides_dorei
Roseburia_inulinivorans
Lawsonibacter_asaccharolyticus
Ruminococcus_bromii
Bifidobacterium_longum
Parabacteroides_merdae
Blautia_obeum
Odoribacter_splanchnicus
Asaccharobacter_celatus
Alistipes_finegoldii
Roseburia_hominis
Roseburia_intestinalis
Ruminococcus_torques
Bacteroides_stercoris
Oscillibacter_sp_57_20
Eubacterium_ramulus
Prevotella_copri
Bifidobacterium_adolescentis

Figure 7

3.2 Large scale investigation of matched control-case gut microbiome study cohorts encompassing 28 disease categories enabled ranking of taxa based on a third hallmark property: Association with Health

The previous investigations enabled the gradation of the 196 taxa based on two key hallmark properties: Core-Influence and Stability-Association. Taxa with higher scores with respect to both these properties have a higher likelihood to act as Core-Keystones of the adult gut microbiome. It indicates that they are prevalent and active members of a normal gut microbiome and are associated with its long-term stability. However, the final key property that needed to be investigated was their association with health. We subsequently investigated the same.

Starting from an overall initial study set of 117 study cohorts encompassing 35,447 gut microbiomes, we selected a core set of 18,020 microbiomes from 62 study cohorts. These cohorts were selected in a step-wise manner such that the sizes of both the diseased and control-matched sub-cohorts in each selected cohort exceeded 15 and neither the diseased or the matched-control sub-cohort sizes were below 20% of the total cohort size (thus removing cohorts with high disparity in sample numbers between the two groups) (**See Methods**). This selected set of 62 cohorts covered 66 different diseases/complications that could be further categorized into 28 different disease categories.

The next step was the identification of shared markers of health and disease across the 62 study cohorts. For this purpose, for each disease category, we identified the disease-positive (enriched in disease with Q-value <= 0.15) and disease-negative (depleted in disease with q-value <= 0.15) taxa using Mann-Whitney tests. Our goal was to provide equal weightage to all disease categories (without any bias towards disease-groups having higher number of study cohorts). Thus, we investigated these alterations at the disease-category level (by grouping together cohorts belonging to multiple disease categories). There were considerable variations in the overall number of taxa alterations across the 28 different disease categories, with some diseases like Colorectal Cancer (CRC), Cardiovascular Diseases (CVD), Prediabetes, Type-II-Diabetes, Covid, IBD and other gut inflammations, especially pronounced with respect to the number of alterations. Nevertheless, as has been previously seen in multiple meta-analysis studies (including some by a subset of the current authors)[1-4], there was a reasonable consistency in the directionality of the alterations observed for the individual taxa across the different diseases. There were 70 (of the 196) taxa, which were consistently and significantly (Q-value <= 0.15) either enriched or depleted across at least 30% of the considered disease categories. Of these, 37 (health-associated taxa) were consistently depleted and 33 (disease-

associated taxa) were consistently enriched. Amongst the health-associated taxa, the taxa with strongest associations were *Faecalibacterium prausnitzii*, *Ruminococcus lactaris*, *Eubacterium ventriosum*, *Barnesiella intestinihominis*, *Eubacterium eligens*, *Coprococcus eutactus* (all significantly depleted in disease in greater than 50% of the disease categories). At the other end of the spectrum, were the strongest disease-associated taxa, starting from *Ruminococcus gnavus*, *Clostridium symbiosum*, *Clostridium bolteae*, *Bifidobacterium dentium*, *Eggerthella lenta* a)nd *Flavonifractor plautii* (which interestingly was also identified as a major core-influencer;(all significantly enriched in greater than 50% of the disease categories). All the 196 taxa could be placed into a graded spectrum based on the consistency of their health association patterns. This provided the basis for our third ranking index, which we referred to as the Health-Association Score.

Furthermore, to avoid biases in the computation of Health-Association scores because of certain specific diseases showing aberrant microbiome alteration patterns, the health-association scores were computed as described in, using a similar boot-strapped iteration-based approach (with strong correlations in the pattern of associations across iterations, ranging from 0.84 to 0.95) as for the Stability-Association Scores. Integrated evaluation of the three scores (Core-Influence, Stability-Association and Health-Association) divided the 196 taxa into specific groups based on their association with three hallmark properties. A core set of 18 (of the 196 species) showed strong association (having all the corresponding rank scores in the top 30%) with all the three properties. We referred to these as the set of primary Health-Associated Core-Keystones (HACKs). Besides the 18 HACKs, an additional 74 taxa showed association with only or two specific properties. Some notable examples include *Prevotella copri*, *Escherichia coli*, *Flavonifractor plautii*, *Collinsella aerofaciens*, *Gordonibacter pamaleae*, *Anaerostipes hadrus* and the two *Dorea* species, which showed strong associations only with Core-Influence, but not with stability and health. There were others like *Paraprevotella xylaniphila*, *Ruminococcus callidus*, *Bifidobacterium catelunatum*, *Lactobacillus*, *Butyrivibrio crossotus*, which showed association only with health, but not with stability and influence.

In order to account for these variabilities of association of the different species with the three individual properties, we subsequently devised a combinatorial taxa-specific index called the ‘HACK-Score’ that captures the consensus association of a taxon with these three individual hallmark properties as a single measure. In simple terms, the HACK-Score was a calculated as product of two scores, namely the mean score of association of a taxon with all three properties and a reward score that measures how similar (or unevenly distributed) are the three association scores with respect to each other.

3.3 Computation of the taxa-specific HACK

We subsequently devised a single combinatorial taxa-specific index called the 'HACK-Index that captures the consensus association of a taxon with these three individual hallmark properties as a single measure. In simple terms, the HACK index was a calculated as product of two scores, namely the mean of the association scores of a taxon for all the three properties and a reward score that measures how similar (or evenly distributed) are the three association scores with respect to each other. The top of this list was the well-known microbiome health marker *Faecalibacterium prausnitzii*, followed by *Bacteroides uniformis*. Other members in the top 20 on this list included multiple species of *Roseburia*, *Alistipes*, *Eubacterium* and *Coprococcus catus*. Interestingly, just succeeding *B. uniformis* on this list were *Fusicatenibacter saccharivorans*, *Agathobaculum saccharivorans*, and *Odoribacter splanchnicus*. The higher HACK index of these three taxa (also *Oscillibacter sp* at 10) as compared to other known markers like *Roseburia*, *Alistipes*, *Coprococcus* and *Eubacterium* was surprising. All three have previously shown sporadic associations with health (especially in cognitive function and healthy aging)[5-7], but the number of such studies are limited. However, we could not find studies showing their enrichment in any disease. Other surprising entrants to the top 20 list (also identified as HACK-species were two equol-producing taxa of the *Eggerthellaceae* family, namely *Adlercruetzia equilofaciens* and *Asaccharobacter celatus*. In contrast, the taxa with low HACK indices were dominated by oral taxa (*Actinomyces*, *Veillonella*, *Rothia* and *Streptococcus*) as well as by the known generally disease-associated lineages like *Ruminococcus gnavus*, pathobiont *Clostridium* (*C. ramosum*, *C. bolteae*, *C. clostridium*, *C. innocuum*, etc), *Eggerthella lenta*, *Klebsiella* and *Escherichia*. Notably, many of the earlier 'beneficial' taxa recently shown to have conflicting associations with health including *Akkermansia muciniphila* and *Prevotella copri* (also shown to have high longitudinal variability)[8], *Eubacterium hallii*, *Anaerostipes hadrus* and *Dorea longicatena*[9,10], were identified with intermediate HACK indices. Thus, the obtained HACK-indices clearly reflected the known positive, negative as well as conflicting associations of specific taxa with health and microbiome stability, besides identifying other promising members of the gut microbiome that could be explored for their beneficial traits.

3.4 Identification of specific nutrient requirements and metabolic/functional signatures associated with high HACK-index species

Identification of the functional/metabolic signatures associated with HACK-index is expected to reveal two crucial features of the generic markers of health and disease (reflected by HACK-index). First is to identify what metabolic/functional properties were contributed by putatively beneficial and detrimental members, thereby enabling us to short-list the putative functional determinants of a generic health-associated microbiome. Second was to identify the nutrient preferences associated with generally 'beneficial' and 'detrimental' taxa (i.e. high and low HACK-index, respectively), thereby providing the guidelines on how to externally enrich the putatively beneficial members of our gut microbiome (with higher HACK-index) and impede the growth of the low HACK-index-associated detrimental ones.

We first examined the experimentally-validated species-level metabolic functionalities independently collated from published studies[11,12], and utilized in previous studies[4,13]. These experimentally-validated species-to-metabolite-production/consumption mapping encompassed a total of 361 production/consumption capabilities (271 compounds/ions) from 990 species (120 overlapping with the list of 196 taxa considered in the current study). We then utilized a two-step procedure to identify the metabolic functionalities associated with HACK-indices (**Figure 8**). First, using Random Forest (RF) models, we first checked whether these species-level experimentally-validated metabolic profiles were predictive of the corresponding HACK-indices (as calculated in this study). Next, we selected 96 top metabolic functionalities that had the most diagnostic ability in the RF models (**See Methods**). Next, we further validated these 96 metabolic features by relating their detection across different species with the taxa-level HACK-indices using logistic regression models. We identified 21 metabolic functionalities showing strong association (p-value < 0.05) and 15 metabolic functionalities having marginal association (0.05 <= p-value < 0.1) with species-level HACK-scores (**Figure 9**). The RF models, built using the species-level detection profiles of these 36 metabolic functionalities yielded a significant positive correlation between the actual and the predicted species-level HACK-indices (obtained using the RF-prediction models built upon the) (Spearman Rho = 0.52, P-value = 1.3e-9) (**Figure 10**), indicating that HACK-indices have strongly associated metabolic signatures.

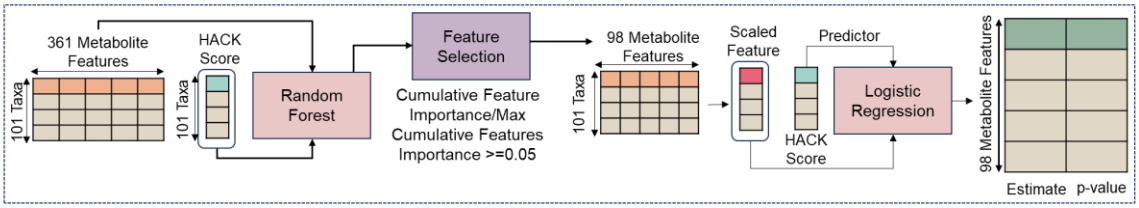

Figure 8

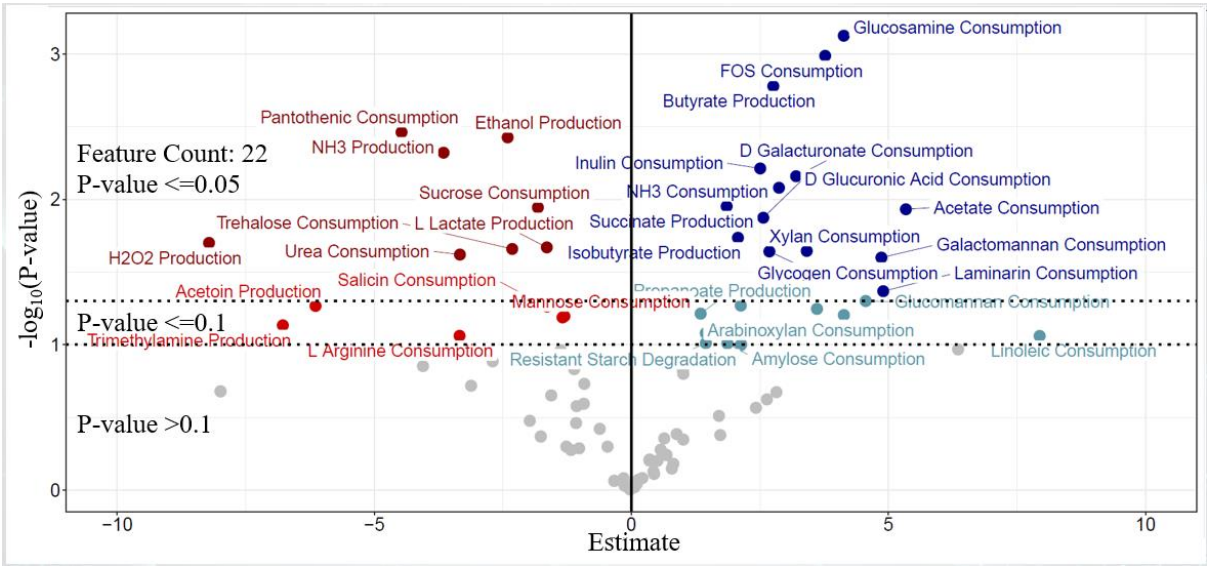

Figure 9

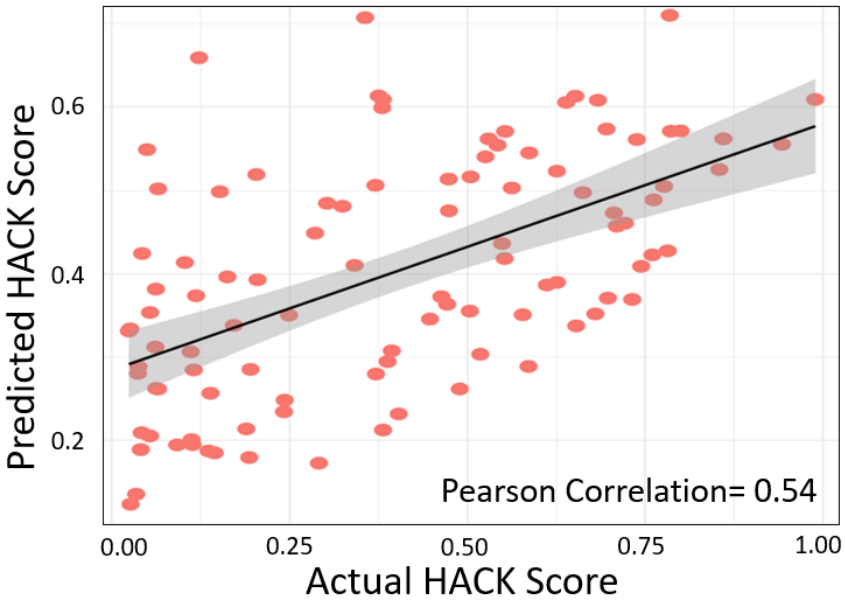

Figure 10

The most noticeable aspect was the pattern of these metabolic functionalities. Taxa-level HACK-indices showed significantly positive association with butyrate, isobutyrate and succinate production and marginally positive association with proponoate indicating that with increasing HACK-indices, species tended to be enriched for production of major short chain and branched chain fatty acids. Species with higher HACK-indices also had unique nutrient preferences characterized by specific complex carbohydrate consumption, with strong enrichments in the consumption of like Fructo-oligosaccharides (FOS), xylan, galactomannan, glucosamine, glucuronic acid and marginal enrichments in the consumption of glucomannan. Xyloglucan and Chondroitin sulfate (many being known prebiotics). At the other end of the spectrum, taxa with lower (or decreasing) indices were characterized by the production of many metabolites with detrimental effects like Hydrogen-peroxide (H2O2), ethanol, Trimethyalmine and Ammonia(NH3) and importantly with the nutrient preferences linked to consumption of several simple sugars like Sucrose and Fructose. A similar functional signature associated with a transition to a non-communicable disease-associated microbiome has been previously observed in our previous studies on Elderly Irish and the Irish Travellers[14,15]. By validating many of the previously known experimentally validated metabolic functionalities with HACK-indices, this association analysis clearly provided the proof-of-principle of utilizing HACK-scores as a measure of identifying functionalities associated with generic microbiome health and disease. However, the number of associated metabolic functionalities were limited and the level of significance of association for a majority of these were also low. This was primarily because not all metabolic functionalities were experimentally across all species-level taxa and not all species-level taxa had information pertaining to their experimentally validated metabolic profiles.

Thus, next we extended our investigation to the genome-level functional annotations for each of these taxa and attempted to identify a comprehensive list of functional profiles significantly associated with HACK-index and the HACK-species (**Figure 11**). Utilizing the repositories of EMBL-MGNify, NCBI-RefSeq and UNINA-Metagenome-Assembled-Genomes[16,17], we collected 31913 high-quality genomes (completeness > 90% and contamination < 5%) from 121 species-level-taxa (overlapping with our 196 taxa). For each genome, we obtained functional profiles corresponding to 8 different schema (encompassing 47,621 functional features) (**Figure 11**). These genome-level function detection profiles were then reduced at the species-level. Subsequently, for each schema, a similar RF-based approach (as in **Figure 12**) was used to identify most informative functional features associated with

species-level HACK-indices, resulting in a total of 6782 functional features. As for metabolic functionalities, iterative RF models with 2-fold cross validation (trained on 50% of taxa and tested on the remaining 50%) built using only the informative features indicated strong predictability of HACK-indices based on species-level functional profiles, with high correlations (ranging from 0.5 to 0.6) between predicted and actual HACK-indices observed for all functional schema (except Biosynthetic Gene Clusters: BiGG). These observations (especially the validation using the two-fold-cross-validation) suggest that species-level functional profiles are predictive of HACK-indices, can potentially identify newer taxa with high HACK indices.

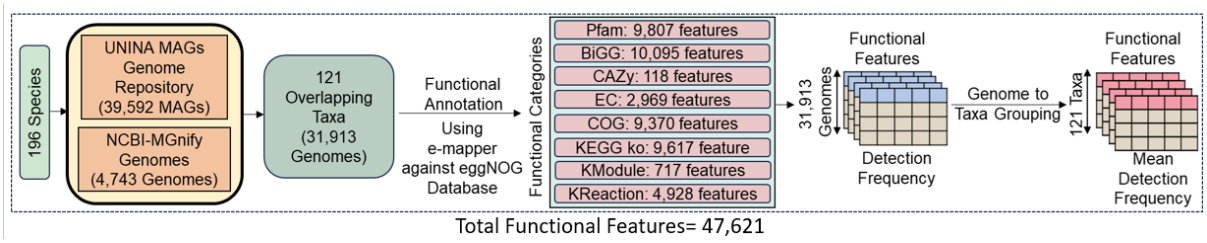

Figure 11

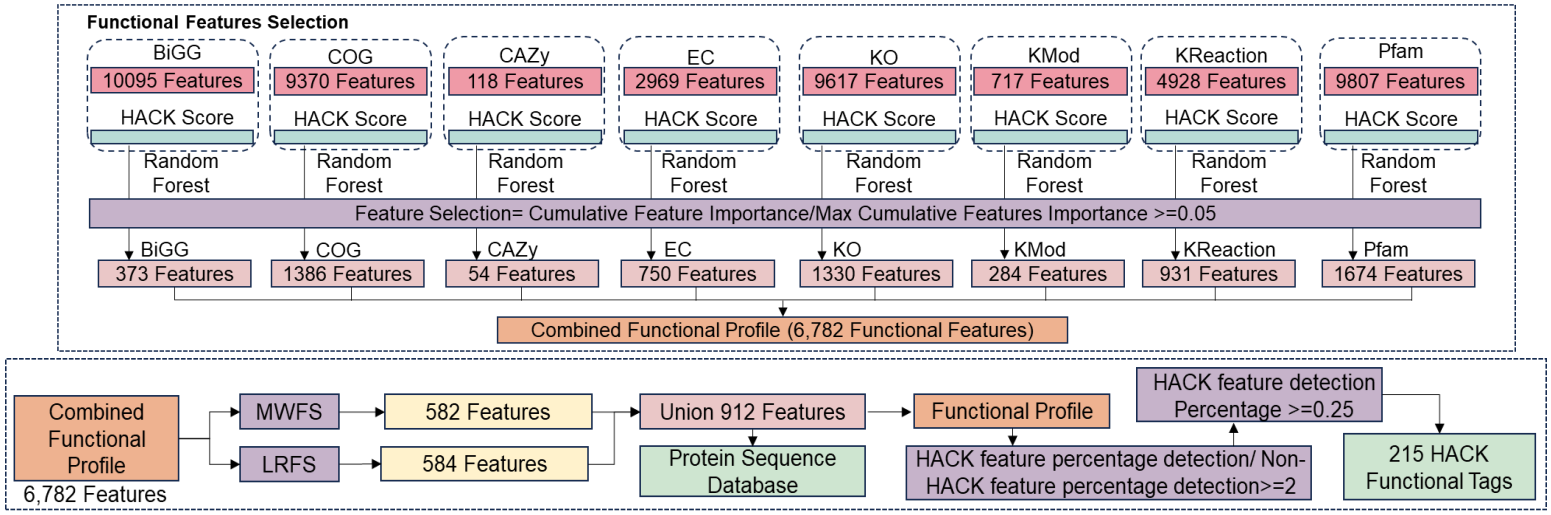

Figure 12

Encouraged by this observation, we thus proceeded to further probe this select list of 6782 functional features using a combination of logistic regression (as in **Figure 12**) (abbreviated as LRFS) and Mann-Whitney tests (abbreviated as 'MWFS'), followed by further filtering based on the detection percentage of the different functional features within the 18 HACK taxa and the detection percentage difference of the same between the HACK and the non-HACK taxa. Using these elaborate steps, we identified a select set of 215 functional features (or tags) from all the 8 schema that were significantly with only the 18 HACK taxa. We mapped back this list back to representative genomes/proteomes of to identify a list of 4617

protein sequences containing at least one of the 215 functional annotations/tags and originating from the representative genomes of the 18 HACK taxa. These protein sequences represent candidates that can be explored further from an evolutionary, mechanistic and pre-clinical perspective to unravel the key functions that characterize the health-associated keystone members of the human gut microbiome and aid in the identification and formulation of novel-biotherapeutics.

REFERENCE

1 Swanson, K. S. *et al.* The International Scientific Association for Probiotics and Prebiotics (ISAPP) consensus statement on the definition and scope of synbiotics. *Nature reviews. Gastroenterology & hepatology* **17**, 687-701, doi:10.1038/s41575-020-0344-2 (2020).

2 Khoruts, A. & Sadowsky, M. J. Understanding the mechanisms of faecal microbiota transplantation. *Nature reviews. Gastroenterology & hepatology* **13**, 508-516, doi:10.1038/nrgastro.2016.98 (2016).

3 Ghosh, T. S. & Valdes, A. M. Evidence for clinical interventions targeting the gut microbiome in cardiometabolic disease. *BMJ (Clinical research ed.)* **383**, e075180, doi:10.1136/bmj-2023-075180 (2023).

4 Zhang, T., Zhang, C., Zhang, J., Sun, F. & Duan, L. Efficacy of Probiotics for Irritable Bowel Syndrome: A Systematic Review and Network Meta-Analysis. *Frontiers in cellular and infection microbiology* **12**, 859967, doi:10.3389/fcimb.2022.859967 (2022).

5 Shetty, S. A. *et al.* Inter-species Metabolic Interactions in an In-vitro Minimal Human Gut Microbiome of Core Bacteria. *NPJ Biofilms Microbiomes* **8**, 21, doi:10.1038/s41522-022-00275-2 (2022).

6 van der Lelie, D. *et al.* Rationally designed bacterial consortia to treat chronic immune-mediated colitis and restore intestinal homeostasis. *Nat Commun* **12**, 3105, doi:10.1038/s41467-021-23460-x (2021).

7 Ghosh, T. S., Shanahan, F. & O'Toole, P. W. Toward an improved definition of a healthy microbiome for healthy aging. *Nature Aging* **2**, 1054-1069, doi:10.1038/s43587-022-00306-9 (2022).

8 Shanahan, F., Ghosh, T. S. & O'Toole, P. W. The Healthy Microbiome-What Is the Definition of a Healthy Gut Microbiome? *Gastroenterology* **160**, 483-494, doi:10.1053/j.gastro.2020.09.057 (2021).

9 He, Y. *et al.* Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat Med* **24**, 1532-1535, doi:10.1038/s41591-018-0164-x (2018).

10 Ghosh, T. S., Das, M., Jeffery, I. B. & O'Toole, P. W. Adjusting for age improves identification of gut microbiome alterations in multiple diseases. *eLife* **9**, e50240, doi:10.7554/eLife.50240 (2020).

11 Vujkovic-Cvijin, I. *et al.* Host variables confound gut microbiota studies of human disease. *Nature* **587**, 448-454, doi:10.1038/s41586-020-2881-9 (2020).

12 Cirstea, M., Radisavljevic, N. & Finlay, B. B. Good Bug, Bad Bug: Breaking through Microbial Stereotypes. *Cell host & microbe* **23**, 10-13, doi:10.1016/j.chom.2017.12.008 (2018).

13 Pasolli, E., Truong, D. T., Malik, F., Waldron, L. & Segata, N. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. *PLoS Comput Biol* **12**, e1004977, doi:10.1371/journal.pcbi.1004977 (2016).

14 Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A. & Alm, E. J. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat Commun* **8**, 1784, doi:10.1038/s41467-017-01973-8 (2017).

15 Gupta, V. K. *et al.* A predictive index for health status using species-level gut microbiome profiling. *Nature Communications* **11**, 4635, doi:10.1038/s41467-020-18476-8 (2020).

16 Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nat Methods* **14**, 1023-1024, doi:10.1038/nmeth.4468 (2017).

17 Renall, N. *et al.* The fecal microbiotas of women of Pacific and New Zealand European ethnicities are characterized by distinctive enterotypes that reflect dietary intakes and fecal water content. *Gut Microbes* **15**, 2178801, doi:10.1080/19490976.2023.2178801 (2023).

18 Mobegi, F. M. *et al.* Intestinal microbiology shapes population health impacts of diet and lifestyle risk exposures in Torres Strait Islander communities. *Elife* **9**, doi:10.7554/eLife.58407 (2020).

19 McDonald, D. *et al.* American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems* **3**, doi:10.1128/mSystems.00031-18 (2018).

20 Dubey, A. K. *et al.* LogMPIE, pan-India profiling of the human gut microbiome using 16S rRNA sequencing. *Sci Data* **5**, 180232, doi:10.1038/sdata.2018.232 (2018).
21 Odamaki, T. *et al.* Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study. *BMC microbiology* **16**, 90, doi:10.1186/s12866-016-0708-5 (2016).
22 Kedia, S. *et al.* Gut microbiome diversity in acute severe colitis is distinct from mild to moderate ulcerative colitis. *Journal of gastroenterology and hepatology* **36**, 731-739, doi:10.1111/jgh.15232 (2021).
23 de la Cuesta-Zuluaga, J. *et al.* Gut microbiota is associated with obesity and cardiometabolic disease in a population in the midst of Westernization. *Scientific reports* **8**, 11356, doi:10.1038/s41598-018-29687-x (2018).
24 Amit, G. & Bashan, A. Top-down identification of keystone taxa in the microbiome. *Nature Communications* **14**, 3951, doi:10.1038/s41467-023-39459-5 (2023).
25 Chng, K. R. *et al.* Metagenome-wide association analysis identifies microbial determinants of post-antibiotic ecological recovery in the gut. *Nature ecology & evolution* **4**, 1256-1267, doi:10.1038/s41559-020-1236-0 (2020).
26 Ianiro, G. *et al.* Variability of strain engraftment and predictability of microbiome composition after fecal microbiota transplantation across different diseases. *Nat Med* **28**, 1913-1923, doi:10.1038/s41591-022-01964-3 (2022).
27 Power, M. E. *et al.* Challenges in the Quest for Keystones: Identifying keystone species is difficult—but essential to understanding how loss of species will affect ecosystems. *BioScience* **46**, 609-620, doi:10.2307/1312990 (1996).
28 Zaneveld, J. R., McMinds, R. & Vega Thurber, R. Stress and stability: applying the Anna Karenina principle to animal microbiomes. *Nat Microbiol* **2**, 17121, doi:10.1038/nmicrobiol.2017.121 (2017).
29 Ma, Z. S. Testing the Anna Karenina Principle in Human Microbiome-Associated Diseases. *iScience* **23**, 101007, doi:10.1016/j.isci.2020.101007 (2020).
30 Ghosh, T. S., Das, M. Jeffery I. B., O' Toole P. W. Adjusting for age improves identification of gut microbiome alterations in multiple diseases. *eLife* (2020).
31 Barton, W. *et al.* The microbiome of professional athletes differs from that of more sedentary subjects in composition and particularly at the functional metabolic level. *Gut* **67**, 625-633, doi:10.1136/gutjnl-2016-313627 (2018).
32 Cronin, O. *et al.* A Prospective Metagenomic and Metabolomic Analysis of the Impact of Exercise and/or Whey Protein Supplementation on the Gut Microbiome of Sedentary Adults. *mSystems* **3**, doi:10.1128/mSystems.00044-18 (2018).
33 McDonald, D. American Gut Project fecal sOTU counts table. *figshare*, doi:https://doi.org/10.6084/m9.figshare.6137192.v1 (2018).
34 McDonald, D. Full American Gut Project mapping file. *figshare*, doi:https://doi.org/10.6084/m9.figshare.6137315.v1 (2018).
35 Wilmanski, T. *et al.* Heterogeneity in statin responses explained by variation in the human gut microbiome. *Med* **3**, 388-405.e386, doi:10.1016/j.medj.2022.04.007 (2022).
36 Pinna, N. K. *et al.* Trans-ethnic gut microbial signatures of prediabetic subjects from India and Denmark. *Genome medicine* **13**, 36, doi:10.1186/s13073-021-00851-9 (2021).
37 Oksanen, J., Guillaume, F., Friendly, M., Kindt, R., McGlinn, D., Minchin, P., O'Hara, R., Simpson, G., Solymos, P., Henry, P., Stevens, H., Szoecs, E., Wanger, H. . vegan: Community Ecology Package. *R package* **2.5-7** (2020).
38 Keohane, D. M. *et al.* Microbiome and health implications for ethnic minorities after enforced lifestyle changes. *Nat Med*, doi:10.1038/s41591-020-0963-8 (2020).