Refined Enterotyping Reveals Dysbiosis in Global Fecal Metagenomes

Authors

Marisa Isabell Keller[1], Suguru Nishijima[1], Daniel Podlesny[1], Chan Yeong Kim[1], Shahriyar Mahdi Robbani[1], Christian Schudoma[1], Anthony Fullam[1], Jonas Richter[1], Ivica Letunic[2], Wasiu Akanni[1], Askarbek Orakov[1], Thomas Sebastian Schmidt[1], Federico Marotta[1], Jonel Trebicka[3], Michael Kuhn[1], Thea Van Rossum[1]*, Peer Bork[1,4,5] *

Affiliations

1. European Molecular Biology Laboratory, Structural and Computational Biology Unit, 69117 Heidelberg, Germany
2. biobyte solutions GmbH, Bothestr. 142, 69126 Heidelberg
3. Department of Internal Medicine B, University Hospital Muenster, Muenster, Germany
4. Max Delbrück Centre for Molecular Medicine, Berlin, Germany
5. Department of Bioinformatics, Biocenter, University of Würzburg, Würzburg, Germany

* Co-corresponding

Correspondence: thea.van.rossum@gmail.com and peer.bork@embl.de

Abstract

Enterotypes describe human fecal microbiomes grouped by similarity into clusters of microbial community composition, often associated with disease, medications, diet, and lifestyle. Numbers and determinants of enterotypes have been derived by diverse frameworks and applied to cohorts that often lack diversity or inter-cohort comparability. To overcome these limitations, we selected 16,772 fecal metagenomes collected from 38 countries to revisit the enterotypes using state-of-the-art fuzzy clustering and found robust clustering regardless of underlying taxonomy, consistent with previous findings. Quantifying the strength of enterotype classifications enriched the enterotype landscape, also reflecting some continuity of microbial compositions. As the classification strength was associated with the patient's health status, we established an "Enterotype Dysbiosis Score" (EDS) as a latent covariate for various diseases. This global study confirms the enterotypes, reveals a dysbiosis signal within the enterotype landscape, and enables robust classification of metagenomes with an online "Enterotyper" tool, allowing reproducible analysis in future studies.

Introduction

The human fecal microbiota has emerged as a critical axis of our health[1]. Unique in its composition for each individual[2], the microbiota is inherently shaped by microbial interactions and external influences[3]. Environmental and host factors, including genetics, physical activity, diet, and medications, contribute to the microbiome's variability, as do both specific and unspecific effects that appear in many diseases[4,5]. While specific alterations enable microbial signatures as biomarkers, broader microbiome dysregulations commonly called 'dysbiosis' tend to be more unspecific and are typically characterized by an unfavorable compositional divergence from healthy reference cohorts[6]. However, the definitions and causes of dysbiosis vary, and its implications are still poorly understood[7].

To simplify the complex structure of the fecal microbiome, taxonomic profiles have been grouped into distinct, reproducible microbial community clusters (often at the genus level) called 'enterotypes,' which are dominated by and typically named after specific taxa[8]. While most studies consistently detected enterotypes with high relative abundance of the *Prevotella* or *Bacteroides* genera, others reported communities dominated by Clostridiales[9], Firmicutes[10,11], or *Ruminococcaceae*[8]. Although the exact count and constitution partially differed between studies, enterotypes have been valuable in clinical contexts for stratifying study populations based on enterotype-specific phenotypes[12,13], responses to treatments[9,14], and dietary interventions[15].

The methods previously used to define enterotypes, as well as the size and diversity of underlying populations, have varied greatly, making comparisons across studies difficult. Enterotypes were first established in 2011 based on partitioning around medoid (PAM) clustering of 33 samples from three continents, from which three clusters emerged, each dominated by either *Bacteroides, Prevotella,* or *Ruminococcus*[8]. These enterotypes were supported by Dirichlet multinomial mixtures (DMM)[16] modeling on 1106 metagenomes from Belgian individuals[17] and a meta-cohort of 785 individuals from the USA, Europe, and China[10].

These studies do not rule out other counts of enterotypes, as the size of the study cohort was still below the reported lower limit for robust metagenome-wide association studies[18]. Indeed, similar work on 888 samples collected in France, Denmark, and Germany pointed to an additional *Bacteroides* enterotype associated with disease, inflammation, low fecal cell counts, and a high BMI[14]. Other analyses found between two and four enterotypes but relied on 16S rRNA gene sequencing data or were limited to homogenous study populations from single countries[19].

More recent work by Tap et al. applied DMM clustering to 3,233 metagenomes, which was then used to predict enterotypes and study their trajectory using the PHATE algorithm[20]. Frioux et al. leveraged bacterial co-occurrence patterns to identify enterotypes in 5,230 metagenomes but did not include

samples from non-Westernized populations[21]. Both studies found clusters that largely overlapped with the initially proposed enterotypes.

A recurring criticism of the concept of enterotypes was the initial assumption of discrete clusters of microbiome states. It has been argued that the microbiome follows a continuous gradient of inversely related *Prevotella* and *Bacteroides* abundances[22,23]. Thus, a hybrid model has been proposed that may better reflect reality[10], where low-density gradients connect high-density clusters (enterotypes).

Here, we developed such a quantitative model by revisiting the enterotype concept using fuzzy k-means (FKM) clustering[24] on a curated, comprehensive, diverse, and, to date, the most extensive set of 16,772 metagenomes collected globally from 38 countries. Using a rarefaction approach, we observe the enterotype classification becoming more stable with an increasing number of samples. Leveraging FKM properties, we quantify the enterotype classification strength for each sample, effectively accounting for the inherent continuity of the microbiome composition landscape. Following the Anna Karenina principle for microbiomes[25], which postulates that “all healthy microbiomes are similar; any dysbiotic microbiome is dysbiotic in its own way,” we observe that microbiomes with lower enterotype classification strength are associated with a variety of diseases, leading us to propose the 'Enterotype dysbiosis score' (EDS). Both, EDS and the enterotype classifications, should inform clinical decision making and are publicly accessible through the "Enterotyper" tool at https://enterotype.embl.de.

Results

Development of robust and global enterotypes.

We assembled a global dataset of 16,772 metagenomes from 129 studies (Table S1) collected from 13,517 individuals in 38 countries (Figure 1A). This population had a mean BMI of 24 (SD 5.34, n=5,036), a mean age of 40 years (SD 17.15, n=7,432) ranging from 5 to 70 years, and included 32 reported diseases and conditions.

To assess the applicability of enterotyping on the global dataset and compare the results with previous work, we applied two commonly used clustering methods, PAM and DMM[10]. DMM clustering did not indicate an absence of clusters or an exact number of clusters. Instead, the optimal number of clusters was the maximum attempted, 15 (Figure S1A). We benchmarked the effect of dataset size and heterogeneity and found that increasing the number of samples resulted in more clusters (Figure S1B-C). This is in line with previous DMM clustering on 3,233 samples resulting in 24 clusters[20]. Thus, DMMs alone are unlikely to reduce this dataset's complexity into simple models as they introduce too many degrees of freedom.

Using PAM clustering, the silhouette width suggested that two, three, and four enterotypes described the dataset well, with slightly higher support for two enterotypes (Figure S2A). Further evaluation with the CH index predicted the presence of three clusters (Figure S2C). According to a PERMANOVA

analysis, batch effects between studies are the predominant factor driving microbial variance across samples, explaining 10.5% of the total variance (adjusted $q < 0.01$; Table S2). Therefore, we wanted to estimate the effect of integrating multiple studies into a single dataset on predicting the optimal cluster number. We performed a rarefaction analysis by randomly selecting increasing subsets of studies. As the number of samples and studies increased, the optimal cluster number stabilized, with three clusters identified by the CH-index and two by the silhouette width (Figure S2B+D). These models differ because the Firmicutes and Bacteroides enterotypes are separate in the 3-cluster model and combined in the 2-cluster model (Figure S3). Lumping samples into one larger cluster loses information, but splitting them into two clusters may overestimate their distinctiveness. It is, therefore, necessary to quantify enterotypes taking the inherent continuity of microbiome composition into account. Fuzzy clustering allows overlapping clusters and quantifies the enterotype classifications, which results in more nuanced and stable enterotyping across diverse and large datasets.

Fuzzy *k*-means (FKM) clustering confirms and enriches enterotype landscape

We applied the fuzzy *k*-means (FKM) algorithm[24] and identified three enterotypes as the best approximation to describe the microbial compositional landscape (Figure 1B, Figure S4B). The rarefaction analysis for FKM, similar to the one for PAM clustering, demonstrated that an increased number of samples and studies stabilized the prediction (Figure S4A), emphasizing the importance of a large, global, and balanced study population for accurately calculating enterotypes given the heterogeneous nature of the microbiome. Larger cohort sizes are more likely to contain a critical mass of samples belonging to each enterotype. In contrast, analysis based on a single limited cohort may result in clusters reflecting study-specific effects rather than corresponding to globally consistent enterotypes. The wide range of different enterotype proportions in the individual studies (Figure 2A, Figure S5) supports this hypothesis that clustering on single cohorts would not result in the same enterotype classifications, as cluster methods are usually biased towards detecting equally sized clusters[26].

One identified global enterotype was associated with high *Bacteroides* and *Phocaeicola* abundance ('Bacteroides/Phocaeicola enterotype', Figure 2B). The latter genus was recently split off the *Bacteroides*[27]. Their relative abundances correlate ($R=0.64$, $p < 2.2e-16$, Pearson correlation, log10-transformed data) and jointly drive this enterotype. Prevotella dominated a second enterotype, and the third enterotype was driven by several genera, mainly belonging to the Firmicutes phylum (e.g., *Ruminococcus_E, Gemminger, Anaerostipes, Dialister, Blautia*) and was therefore named the 'Firmicutes' enterotype. These taxa are all based on the GTDB[27] taxonomy, which is increasingly used in metagenomic studies instead of the traditional NCBI[28] classification. We compared both taxonomies to assess their influence on enterotypes. The silhouette width showed consistent results (Figure S4E), and enterotype classification resulted in 94.8 % (n=15,852) of the samples being classified into the same enterotype regardless of the underlying taxonomy.

134 FKM calculated the strength with which the microbiomes belong to each enterotype by reporting the
135 consistency of enterotype classifications over multiple cluster iterations, which makes it suitable for
136 data sets that do not fit neatly into clear categories. This "classification strength" offers new perspectives
137 on the microbiome landscape, presenting a hybrid design between distinct enterotype clustering and the
138 gradient concept that considers the continuous nature of fecal microbial composition. We found that
139 FKM classifies 87.5% (n=14,675) of samples into the same three enterotypes as assigned by PAM
140 clustering (Figure S4C) while reporting lower classification strength for the samples that have different
141 classifications across the two models (Figure S4D).

142 **Enterotypes built on taxonomic profiles reflect microbial metabolic differences.**

143 We aimed to quantify the extent to which enterotypes, as preferred taxonomic compositional states,
144 vary in their functional profiles and, consequently, their impact on the host. Using a dimensional
145 reduction method (MDS), the functional profiles for each sample showed a remarkable alignment with
146 enterotype classifications (Figure 2D). The Firmicutes enterotype separates from the others across the
147 first dimension, while the Bacteroides/Phocaeicola and Prevotella enterotype separates across the third
148 dimension. We compared the differential abundance of KEGG-Orthologs between enterotypes (Figure
149 S6) and found, for example, pronounced differences in carbohydrate uptake mechanisms. The
150 Bacteroides/Phocaeicola and Prevotella enterotypes are characterized by specialized outer membrane
151 proteins for starch uptake[29]. In contrast, the Firmicutes enterotype shows an ATP-binding cassette (ABC
152 system) enrichment associated with glycan uptake[30].

153 **Robustness of enterotypes across methodology in integrated cohorts.**

154 Recent studies of combined cohorts used different approaches to summarize the complexity of the
155 human microbiome into higher-level community groups[20,21], which we wanted to compare to the FKM-
156 derived enterotypes. Tap and colleagues used DMM on 3,233 metagenomes, subsequently imposing
157 the clustering on around 15,000 additional samples followed by hierarchical clustering[20], while Frioux
158 and colleagues focused on bacterial co-occurrence patterns[21]. The high overlap of samples between Tap
159 et al. and this study enabled us to link the sample classifications directly. To include data from Frioux
160 et al. on the intersection, we used their online prediction tool on our cohort. The results across the
161 different approaches are consistent, identifying three major clusters associated with *Bacteroides,*
162 *Prevotella,* or Firmicutes species (Figure 3A & S7A). Between Tap et al. and this study, 4,132 samples
163 (75,5 %) are consistent in their enterotype classification, while Frioux et al. and this study agree in
164 11,551 (72,3 %) samples. Samples showing inconsistent classification were associated with a lower
165 FKM-derived classification strength (Figure 3B & S7B), reinforcing the substantial overlap between
166 the sample classification for central enterotype samples.

167 Additional groups in Tap et al. and Frioux et al. show only a small overlap with our dataset as they are
168 mainly found among young children, which motivated us to contextualize the mature human
169 enterotypes from the FKM on our global dataset. We combined our dataset with other human gut

170 microbiome samples from infants (< 5 years) and older people (>70 years), as well as samples from
171 different human body sites and other animal host-associated habitats (total n = 49,762). Using a
172 dimensional reduction method (UMAP)[31], we observed a clear distinction between the human gut
173 microbiomes of adults and infants and those of other human and non-human habitats (Figure S8). While
174 different dimensionality reduction methods were used to observe the host and body site separation,[32,33]
175 we also found a reflection of the enterotype classifications. This highlights the distinct enterotype signal
176 in the human gut microbiome even when clustered with vast heterogeneous samples.

177 **Stability of enterotypes and associations with demographic measures and diseases**

178 Our dataset is a rich resource for testing the temporal stability of enterotype classification, as it contains
179 651 subjects with multiple samples within one year (3,853 samples, Figure S9A). We used a Markov
180 chain model to assess the stability of classifications (Figure 4A; examples in Figure S9B-C) and found
181 that 57% (n=362) of the subjects showed high stability in their enterotypes. Among all enterotype
182 changes, the most frequent is to the Firmicutes enterotype, starting from the Bacteroides/Phocaeicola
183 (186 changes, 28%) and Prevotella (116 changes, 17%) enterotypes. Transitions between the Prevotella
184 and Bacteroides/Phocaeicola enterotypes were rare (49 changes in both directions, 7%) and possibly
185 induced by microbiome perturbations. Many subjects with changes in enterotypes suffer from
186 inflammatory bowel disease (n=55), a condition that is known to increase microbial variability[6], or took
187 part in a probiotic intervention study (n=32)[34].

188 After having established a robust global enterotype framework that is independent of underlying
189 taxonomy and study effects as well as stable over time in an individual, we then adequately explored
190 the relationships between enterotypes and host phenotypes. We found the Firmicutes enterotype
191 microbiomes had higher alpha diversity, and the subjects were often older, female, and had higher BMIs
192 (Figure 4B-E). Male subjects were overrepresented in the Prevotella enterotype (Figure 4E). Our dataset
193 includes 395 samples from five studies with documented dietary patterns, which revealed a significant
194 association between diet (omnivore, vegetarian, vegan) and enterotypes (chi-squared test, p = 2.3e-15;
195 Figure 4F). Plant-based diets were most associated with the Prevotella enterotype. Only 10% of
196 omnivores (n=26) were classified as the Prevotella enterotype, compared to 15% (n=14) of vegetarians
197 and 41% (n=16) of vegans. Notably, no vegan had microbiomes classified to the
198 Bacteroides/Phocaeicola enterotype. This adds to previous hypotheses based on smaller datasets
199 implying that dietary habits at least contribute to enterotype stratification[35].

200 To explore the relationship between disease and symptoms with the enterotypes, we used linear mixed-
201 effect models (lmer), adjusting for multiple studies and samples, and generalized linear models (glm)
202 for single-study, non-repeated sampling data (Figure 4G). Instead of limiting the analysis to rigid
203 enterotype classifications, we used the classification strength of each sample for each enterotype to
204 provide a more nuanced analysis. We only analyzed the study's internal samples to use matched cases
205 and controls. End-stage renal disease was associated positively with the Firmicutes enterotype (glm,

206 coefficient = 0.096, q = 3.2e-05) and negatively with the Prevotella enterotype (glm, coefficient = -
207 0.14, p = 8.8e-03), which is consistent with previous findings[36]. Diarrhea was associated with the
208 Bacteroides/Phocaeicola enterotype (glm, coefficient = 0.088, q = 2.2e-02), and cardiovascular disease
209 showed associations with the Firmicutes enterotype (lmer, coefficient = 0.04, q = 1.9e-02).

210 **Enterotype dysbiosis scores (EDS) are linked to subjects' disease status.**

211 Besides the associations of specific diseases or symptoms with the enterotypes, we wondered to what
212 extent samples from diseased individuals generally fit into the enterotype landscape. FKM clustering
213 quantifies how well a microbiome fits in each enterotype using the consistency of classification across
214 cluster iterations in the process of defining the best cluster number. This is reflected visually in a PCoA,
215 where samples central to an enterotype have more robust classification than those between enterotypes
216 (Figure 5A). We hypothesized that samples with lower classification strength are more dysbiotic and
217 developed a classification strength-based Enterotype dysbiosis score (EDS). A higher EDS indicates a
218 higher deviation from enterotype centers, which translates to dysbiotic microbiomes being located
219 somewhere in the space "between" enterotypes.

220 EDS showed a significant correlation with a previously established dysbiosis score, which defines
221 dysbiotic samples as different to a healthy reference set[6] (Figure 5B). Alpha diversity as a common
222 measure of gut health[37] showed a negative relationship with EDS (Shannon, R = -0.43, p < 2.2e-16).
223 Healthy subjects had lower EDS (Figure 5C), and disease severity correlated with EDS in a cohort of
224 liver cirrhosis patients (Figure 5E). Twenty out of 32 conditions, including colorectal cancer, Crohn's
225 disease, and ulcerative colitis, are associated with higher EDS (Figure S10). Most of these diseases have
226 been previously linked to alterations in microbiome composition[5,38,39] The most dramatic shift in EDS
227 was seen in patients with a reported *Clostridium difficile* infection. This could be explained by repeated
228 antibiotic usage of these patients, which is associated with a higher EDS (Figure 5D). The EDS provides
229 a latent measure of dysbiosis, which could be used for further patient stratification beyond taxonomic
230 profiling. It is independent of a particular disease or the availability of a healthy reference sample set,
231 and accounts for the underlying microbial compositional states across various disease phenotypes.

232 **The online 'Enterotyper' tool predicts enterotype and EDS**

233 To facilitate enterotype classification allowing an abstract view of the state of an individual
234 microbiome, we have developed an online prediction tool. It classifies enterotypes taking into account
235 the strength of the clustering as well as an associated dysbiosis estimation for newly generated human
236 fecal microbiome samples. We trained XGBoost regression models[40] and validated them on an
237 independent dataset (n = 347, details in methods). The XGBoost models showed areas under the curve
238 (AUC) ranging from 0.985 – 1 (Figure 6A & Figure S11) and predicted strength for the enterotype
239 classifications correlated with the validation dataset (Figure 6B).

Given the overlap between the enterotype classifications between the different approaches described above and the global, comprehensive training dataset used, we have made our XGBoost models publicly available via an easy-to-use online "Enterotyper" (https://enterotype.embl.de/) to facility easy classification of a given taxonomic profile into enterotypes and estimation of the EDS (Figure 6C).

Discussion

Enterotypes describe states of the human fecal microbiome composition associated with various factors including diseases and host lifestyle[3], providing a useful abstraction of the complex microbiome into an easily communicated latent variable. More than a decade after the first enterotype publication[8], we revisited the concept, quantified nuance within the model, and studied its relationship to host health and microbiome dysbiosis. A new enterotyping methodology was used on a curated set of 16,772 metagenomes, 500 times the size of the original study population. We observe that the three-enterotype landscape of Bacteroides/Phocaeicola, Prevotella, and Firmicutes-dominated communities remains the best approximation of the complex landscape of community composition.

The observation of enterotypes suggests that, across the human population, there are a few distinct alternative states into which microbial communities tend to converge. Given the symbiotic relationship between human hosts and their microbial communities, these states should promote the health of their host, at least in the sense of high-density areas. This supports a normative model for eubiosis and dysbiosis: what is typically observed in healthy individuals is the balanced, eubotic state, while rarer states and those seen in unhealthy individuals are the unbalanced or 'dysbiotic' states. This concept is reflected in the Anna Karenina Principle for microbiomes[25], which argues that dysbiotic microbiomes are expected to vary widely, while eubiotic microbiomes should be more constrained. We see this reflected in our results, where healthy individuals are more likely to have a solid membership for one enterotype, and unhealthy individuals are more likely to be intermediary or completely distinct from the enterotypes. A novel aspect of the enterotype model presented in this work is that this membership strength can be quantified and translated into the Enterotype dysbiosis ccore (EDS). The EDS builds on a previous approach[6] that is conceptually similar but does not consider compositional states within the eubiotic reference set. As we found associations of EDS with disease states and severity, it can be used to stratify individuals within statistical analyses and to study links between the microbiome, enterotypes, and human health.

In the enterotype landscape presented here, a second Bacteroides enterotype, previously identified through DMM in cohorts up to approximately 2,500 samples[14,41,42], is not represented. This cluster was not seen as a distinct enterotype in our study because its samples were clustered together with the samples in the Bacteroides/Phocaeicola enterotype. This is likely due to the combined effect of DMM tending to call more clusters and the larger sample size 'filling in the space' between the two Bacteroides clusters. The second DMM-derived Bacteroides enterotype was previously linked to high BMI[14],

275 obesity[43], metabolic disorders[13], IBD[41], and cardiovascular disease[42]. We acknowledge the importance
276 of these associations and the usefulness of DMM in describing these study populations. However, the
277 effect of reporting higher cluster numbers with higher sample numbers found in this study and by Tap
278 et al.[20] demonstrated the limitations of using DMM with large heterogeneous cohorts to describe broad,
279 overarching patterns. This is an excellent example of how clustering within smaller, narrowly defined
280 datasets might be more influenced by study- or disease-specific signals, as also shown in our rarefaction
281 analysis, but may provide additional insights. The two goals should be clearly distinguished: one is to
282 describe the structure of microbiomes as broadly and globally as possible (enterotypes) and the other is
283 to define population or phenotype-specific microbiome clusters, which can be useful for investigating
284 disease-specific and local effects. For the latter, we support the exploration of study-specific clusters
285 by de novo clustering. For the former, we recommend that the model developed and presented here be
286 used. To enable this, we provide researchers with the 'Enterotyper' tool. This provides reliable and
287 reproducible enterotype classification of new metagenomic samples against the background of the
288 global dataset, ensuring reproducibility and comparability between studies.

289 A major co-variate with enterotype is diet. The association of the Prevotella enterotype with a high-
290 fiber diet has been previously reported[10,44], while a diet richer in animal fat and simple sugars has been
291 linked to the Bacteroides enterotype[10,45]. Diet is, in turn, associated with geography, and indeed the
292 global distribution of enterotypes seen in this study overlaps with the typical dietary patterns prevalent
293 in the different regions. North America has both the largest proportion of individuals in the
294 Bacteroides/Phocaeicola enterotype and the highest rate of animal protein consumption (83g per day
295 per capita[46]), while Africa and Asia have the largest proportion of individuals in the Prevotella
296 enterotype and the highest rate of more plant-based protein consumption (50g and 58g, respectively).
297 Congruently, we observed that the proportion of samples classified into the Prevotella enterotype
298 increased with the strictness of plant-based diet (i.e., omnivore to vegetarian to vegan). In addition, the
299 biggest change in the functional analysis of KEGG orthologs across enterotypes was linked to microbial
300 metabolism of dietary carbohydrate compounds (starch and glycan uptake). Although diet is a major
301 influence on the gut microbiome[17] and enterotype classification, dietary metadata was only reported for
302 2% of samples (n=395) in our study population. Documenting reliable dietary data is challenging,
303 requiring restrictive measures to record or control all food intake or relying on subjective
304 questionnaires. To better understand effects of diet on host health and the microbiome, longitudinal
305 studies focusing on microbial metabolism in food digestion[47] and combined metagenomic and
306 metabolomic studies with rich contextual data are needed[48]. Other sample biases include urban and rural
307 differences, access to healthcare system and local dietary preferences is. While we included as much
308 data from non-Western populations as was available, there is still a general under-representation of
309 samples from non-Western communities in the microbiome field[49] and we encourage more sampling
310 efforts in non-Westernized communities.

311 This study advances our understanding of the diversity and constraints of the human fecal microbiome
312 by refining the enterotype model. We increased robustness by including a much larger and more
313 representative population and a modern taxonomy. The methods of fuzzy clustering adopted to
314 enterotyping also allowed to utilize the correlation of classification strength and dysbiotic to develop
315 and provide a dysbiosis estimation. Together with the easy-to-use and publicly available "Enterotyper"
316 classification tool that allows a standardized comparison of newly generated with historic datasets.

Online Methods

Data Overview

For the global analysis, we gathered 27,674 publicly available human fecal metagenomic samples from 173 studies from the SPIRE database [50]. For our enterotype cohort, we only kept samples from subjects between 5 and 70 years old and removed samples described in original publications as deriving from "elderly" subjects. For 21 studies that provided fecal metagenomic samples from infants, we removed the samples if no age was reported for the subject. We excluded samples from fecal microbiome transplantation (FMT) studies and individuals taking antibiotics, as these interventions are likely to alter the microbiome's composition substantially[51,52]. Samples with no reported disease status and no antibiotic treatment record (N=4,122) were not removed. If we could not link sample-specific clinical information but the study contained a general description of a disease or lack of disease for all samples in the study, we manually added that disease or lack of disease to the metadata. After filter out samples with irregular taxonomic composition (see taxonomic and functional profiling), the remaining sample collection comprised 16,772 samples from 129 studies (see Supplementary Table 1).

Metagenomic data processing

Taxonomic and Functional profiling

Metagenomes were taxonomically profiled with mOTUs v2.5 with default parameters. Analyses were performed using two taxonomic classifications of individual mOTUs: the default taxonomy (based on the NCBI taxonomy) and a newly generated GTDB-based taxonomy (based on GTDB r202). The GTDB taxonomy implemented the reclassification of many former Bacteroides species to the genus Phocaeicola. The GTDB taxonomy was obtained by classifying genomes in proGenomes2 (corresponding to "ref_mOTUs") using GTDB-tk and directly mapping "meta-mOTU" marker genes against GTDB marker genes, followed by rule-based and manual resolution of conflicting classifications within individual mOTUs. Two thousand two hundred forty-one samples with more than 10 % unclassified species ("-1" fraction from mOTUs) and samples with a richness of less than ten different mOTUs were removed. For a genus-level taxonomic profile, we summed the relative abundance for all mOTUs with the same genus.

Open reading frames (ORFs) were predicted using Prodigal (v2.6.3) and annotated with eggNOG-mapper (v2.1.3) against eggNOG database 5.0. Each predicted gene was multiplied by the average sequencing coverage of the contig it was found on. If a gene was associated with multiple KEGG orthologs, the average sequencing coverage count was distributed evenly between associated KEGG orthologs. A sample-wise compositional transformation yielded relative abundance values per KEGG orthologs.

351

352 **Clustering and enterotype classification.**

353 All cluster methods used an abundance matrix on the genus level. The Enterotype clusters based on the
354 partitioning around medoid (PAM) clustering on the samples' Jensen-Shannon divergence (JSD) have
355 been identified as described before[8]. In the PAM clustering, the Calinski-Harabasz (CH) index (ratio of
356 the sum of inter-clusters dispersion and intra-cluster dispersion) together with the silhouette width
357 (measure how similar a sample is to its cluster compared to other clusters) has been used to assess the
358 number of clusters and cluster robustness[53].

359 DMM clustering was run as reported before[16] using the DMM R package[54] on square root normalized
360 relative abundances. The optimal number of clusters was evaluated by detecting a local minimum using
361 the Laplace approximation, Bayesian information criterion (BIC), and Akaike's information criterion
362 (AIC). To assess the impact of sample size on DMM clustering, we randomly subsetted the global
363 dataset into fractions of 1/2 (n=8386), 1/4 (n=4193), 1/8 (n=2096), 1/16 (n=1048), 1/32 (n=524) of the
364 original sample size. This sub-setting was repeated three times for each fraction. To benchmark the
365 effect of heterogeneity on the predicted optimal cluster number, we performed DMM clustering on the
366 eight most extensive datasets within our global dataset from an increasing number of studies (number
367 of studies mixed = 1 to 8), keeping sample size approximately constant in the mixed cohorts (number
368 of samples = 996 to 1000). The optimal number of clusters was determined in both benchmarks using
369 the Laplace approximation.

370 The fuzzy *k*-means (FKM) clustering was obtained using the *FKM* function from the fclust R package[55]
371 on the relative abundance matrix, and silhouette width was used as the cluster validity index to select
372 the optimal number of clusters. The silhouette width as the clustering statistic is most comparable to the
373 cluster statistics in the PAM clustering. The CH-index is unavailable as the within and the between the
374 sum of squares cannot be computed in fuzzy clustering as cluster medoids are not defined. The
375 enterotype classification for each sample was obtained from the cluster with the highest enterotype
376 classification strength for this sample.

377 The PCoA was calculated using the *dudi.pco* function from the ade4 R package[56] on the JSD matrix
378 and the PCA based on functional profiles was calculated using *prcomp* function from the stats R
379 package[57].

380 We performed a rarefaction analysis on our cohort to assess the robustness of our PAM and FKM
381 clustering results. This approach allowed us to investigate the impact of varying sample sizes on the
382 stability of the predicted number of clusters. For this analysis, we selected samples from a subset of
383 randomly chosen studies representing 30%, 40%, 50%, 60%, 70%, 80%, and 90% of the studies in our
384 cohort. We repeated this selection process 20 times for iterations of each percentage) and performed
385 FKM clustering for each subset in the same way as described above, utilizing silhouette width as a

386 criterion to determine the optimal number of clusters. Additionally, we conducted PAM clustering using
387 the CH-index and the silhouette width to evaluate the clustering quality and the best number of clusters
388 for each subset.

389 For the UMAP, we compiled taxonomic profiles of animal-host-associated metagenomic samples from
390 the SPIRE dataset [50]. Samples with over 10% unidentified profiles were excluded from the analysis. We
391 eliminated unidentified features for the remaining profiles and re-normalized each to ensure the total
392 equaled one. For the remaining 49,762 samples, we calculated Bray-Curtis dissimilarity based on their
393 GTDB-based profiles. Finally, we conducted UMAP[31] dimensionality reduction, setting the
394 'n_neighbors' parameter to 50 and the 'min_dist' parameter to 0.01.

395

396 ## Statistical Analysis

397 To demonstrate the differential abundance of genera and KEGG orthologs between enterotype clusters,
398 we applied the Wilcoxon rank-sum test and Benjamini-Hochberg correction[58] to adjust the p-value for
399 multiple testing. Recurrent samples from the same subject were removed. To assess the impact of
400 various factors on the distance matrix of the microbiome data, we performed a permutational
401 multivariate analysis of variance (PERMANOVA) using the *adonis2* function from the vegan package
402 in R[59]. The categorical variables of interest, such as age, biological sex, diet, or disease status, were
403 individually tested as predictors of the microbial community structure represented in a JSD matrix.

404 **Longitudinal Analysis**

405 A Markov chain model was used to analyze the likelihood of transitions between enterotypes, providing
406 insights into the stability and dynamics of enterotype classification and acquired using the *markovchain*
407 R package[60]. To ensure a consistent and relevant timeframe for our analysis, we only considered
408 samples collected within one year for each subject. In addition, we excluded the subjects with 89
409 samples to prevent undue influence from this individual with numerous repeated samples from the
410 model.

411 **Enterotype Associations**

412 The associations of enterotype classifications with age, BMI, and alpha diversity (Shannon index) were
413 calculated using the Wilcoxon rank sum test. To test differences in the gender and diet distribution, we
414 performed a chi-squared goodness-of-fit test to determine if the observed frequency in each group
415 deviated significantly from the ratio in our cohort.

416 In available case and control studies, we calculated generalized linear models (glm) using the *glm*
417 function from the R "stats" package[57], to model the association with reported diseases and conditions
418 with each enterotype. For each enterotype, separate models were calculated using the strength value for
419 this enterotype derived from the FKM clustering as a measure of enterotype affiliation. We used only
420 the internal controls of the study to avoid artificially inducing significance through a considerable

421 sample size of controls and to benefit from matched cases and controls that are usually selected in the
422 case and control studies. We adjusted for random effects using the *lmer* function from the lme4 R
423 package[61] to account for longitudinal sampling and study effects where appropriate. Multiple testing
424 corrections were conducted on all p-values using Benjamini-Hochberg corrections.

425 **Enterotype Dysbiosis Score (EDS)**

426 To obtain the enterotype dysbiosis score (EDS), the enterotype strength derived from the FKM
427 clustering for each sample was inverted, z-score normalized, and scaled to have a zero to one scale. As
428 a result, samples with lower enterotype strength in FKM clustering have a higher dysbiosis score. The
429 dysbiosis score by Lloyd-Price et al. 2019 was calculated as described in the original study[6] using the
430 samples in our cohort with a reported absence of disease as the healthy reference cohort.

431 **Enterotype prediction for novel samples**

432 XGboost regression models[40] were constructed to predict the strength of each enterotype, which was
433 determined by FKM clustering. The prediction models were trained based on the genus-level taxonomic
434 profiles of the dataset using the *train* function of the caret package in R[62] with a 10-times repeated 10-
435 fold cross-validation procedure. Before the model construction, minor genera with an average relative
436 abundance of <1e-4 were excluded, and only abundant genera above the threshold were used for the
437 training. To avoid model overfitting due to multiple samples derived from the same individuals, we
438 ensured these samples were incorporated exclusively in the training or the evaluation data during each
439 cross-validation fold. The models' accuracies were evaluated by applying the prediction models to the
440 unused validation dataset (347 samples from three studies). Separate models were constructed for each
441 enterotype (i.e., two models for the two-enterotype clustering and three models for three-enterotype
442 clustering). The highest strength obtained from these models was used for each sample's enterotype
443 classification and EDS.

444 Additionally, binary LASSO classification models were constructed to predict enterotypes determined
445 by PAM clustering. The models were trained and built using the same procedure described above. The
446 highest score obtained from the LASSO classification models was used for enterotype classification.

447 The Enterotype classifications and EDS prediction for new independent datasets are offered via an
448 online and publicly accessible "Enterotyper" at https://enterotype.embl.de (Figure 5, C). Users will
449 receive the enterotype classifications as tabular data after submitting taxonomic profiles to the
450 "Enterotyper," which accepts data from various taxonomic profiling tools. Default settings use the
451 three-enterotype FKM model, which will also report the enterotype classification strength and the EDS.
452 The EDS is calculated by inverting and z-score normalizing using the center and standard deviation
453 from the global enterotype training data as a reference and scaling the maximum strength for an
454 enterotype for each sample. However, XGBoost models reflecting different clustering algorithms
455 (FKM, PAM) and numbers of enterotypes (2, 3, 4) are also provided to allow additional explorative

analysis by informed users. Furthermore, a first graphical overview of the classified enterotypes and EDS on the users' input sample will be displayed compared to the global dataset from this study.

R codes used for enterotyping, statistical analysis, figure plotting and building Enterotyper models are available at https://github.com/grp-bork/enterotypes.

Acknowledgments

We thank all members of the Bork group, Jeroen Raes and Jorge Francisco Vázquez Castellano from the Raes group, and all members of the MICROB-PREDICT study group, especially Mani Arumugam, for fruitful discussions. We are grateful to Julien Tap and Falk Hildebrand for their help and support of open science. We also thank Anna Głazek, Anna Schwarz, Roman Thielemann, Leonie Thomas, Ela Cetin, and Moritz von Stetten for their help with the metadata curation. We acknowledge funding and support from EMBL. The MICROB-PREDICT project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 825694. This paper reflects only the author's view, and the European Commission is not responsible for any use that may be made of the information it contains. This project has been partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 460129525.

Author Contributions

Conceptualization, M.I.K. T.V.R, M.K., P.B.; methodology, M.I.K, T.V.R, M.K., S.N., D.P., F.M.; data analysis, M.I.K, T.V.R., C.Y.K., S.N., J. R., T.S.S, A.O.; collection of public metagenomes and metadata curation, T.S.S., A.N.F., W.A., M.K.; Enterotyper prediction tool, M.I.K, S.N, S.M.R, D.P, C.Y.K, C.S, A.F, I.L, T.V.R, M.K; writing, M.I.K, T.V.R, M.K., D.P., T.S.S., S.N., D.P., P.B.; supervision, T.V.R, M.K., D.P., J. T., P.B. All authors discussed the results, reviewed the manuscript, and approved the final manuscript.

List of abbreviations

AUC	Area under the curve
CH-index	Calinski-Harabasz index
DMM	Dirichlet multinomial mixtures
EDS	Enterotype dysbiosis score
FKM	Fuzzy *K*-means clustering
glm	generalized linear model
GTDB	Genome taxonomy database
LASSO	Least absolute shrinkage and selection operator
JSD	Jenson-Shannon divergence
KEGG	Kyoto encyclopedia of genes and genomes
lmer	linear mixed effect model
mOTU	marker gene-based operational taxonomic unit
NCBI	National center for biotechnology information
PAM	partition around medoid clustering
PCA	Principal components analysis
PCoA	Principal coordinate analysis
PERMANOVA	permutational multivariate analysis of variance
PHATE	Potential of heat-diffusion for affinity-based trajectory embedding
UMAP	Uniform manifold approximation and projection
XGBoost	Extreme gradient boosting

References

1. Gomaa, E. Z. Human gut microbiota/microbiome in health and diseases: a review. *Antonie van Leeuwenhoek* **113**, 2019–2040 (2020).
2. Hildebrand, F. *et al.* Dispersal strategies shape persistence and evolution of human gut bacteria. *Cell Host & Microbe* **29**, 1167-1176.e9 (2021).
3. Schmidt, T. S. B., Raes, J. & Bork, P. The Human Gut Microbiome: From Association to Modulation. *Cell* **172**, 1198–1215 (2018).
4. 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 (2017).
5. Wirbel, J. *et al.* Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nature Medicine* **25**, 679–689 (2019).
6. Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).
7. Hooks, K. B. & O'Malley, M. A. Dysbiosis and Its Discontents. *mBio* **8**, 10.1128/mbio.01492-17 (2017).
8. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473**, 174–80 (2011).
9. Tap, J. *et al.* Identification of an Intestinal Microbiota Signature Associated With Severity of Irritable Bowel Syndrome. *Gastroenterology* **152**, 111-123.e8 (2017).
10. Costea, P. I. *et al.* Enterotypes in the landscape of gut microbial community composition. *Nature Microbiology* **3**, 8–16 (2018).
11. Kushugulova, A. *et al.* Metagenomic analysis of gut microbial communities from a Central Asian population. *BMJ Open* **8**, e021682 (2018).
12. Wang, J. *et al.* Enterotype Bacteroides Is Associated with a High Risk in Patients with Diabetes: A Pilot Study. *J Diabetes Res* **2020**, 6047145 (2020).

523 13. Forslund, S. K. *et al.* Combinatorial, additive and dose-dependent drug–microbiome associations.
524 *Nature* **600**, 500–505 (2021).

525 14. Vieira-Silva, S. *et al.* Statin therapy is associated with lower prevalence of gut microbiota dysbiosis.
526 *Nature* **581**, 310–315 (2020).

527 15. De-Oliveira, G. J. M., Schieferdecker, M. E. M. & Campos, A. C. L. ARE ENTEROTYPES IN
528 OBESE MODIFIED BY BARIATRIC SURGERY, THE USE OF PROBIOTIC SUPPLEMENTS
529 AND FOOD HABITS? *ABCD, arq. bras. cir. dig.* **34**, (2021).

530 16. Holmes, I., Harris, K. & Quince, C. Dirichlet Multinomial Mixtures: Generative Models for
531 Microbial Metagenomics. *PLOS ONE* **7**, e30126 (2012).

532 17. Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science* **352**, 560–564
533 (2016).

534 18. Rothschild, D. *et al.* An atlas of robust microbiome associations with phenotypic traits based on
535 large-scale cohorts from two continents. *PLOS ONE* **17**, e0265756 (2022).

536 19. Oh, H.-S. *et al.* Proposal of a health gut microbiome index based on a meta-analysis of Korean and
537 global population datasets. *J Microbiol* **60**, 533–549 (2022).

538 20. Tap, J. *et al.* Global branches and local states of the human gut microbiome define associations
539 with environmental and intrinsic factors. *Nat Commun* **14**, 3310 (2023).

540 21. Frioux, C. *et al.* Enterosignatures define common bacterial guilds in the human gut microbiome.
541 *Cell Host & Microbe* **0**, (2023).

542 22. Jeffery, I. B., Claesson, M. J., O'Toole, P. W. & Shanahan, F. Categorization of the gut microbiota:
543 enterotypes or gradients? *Nature Reviews Microbiology* **10**, 591–592 (2012).

544 23. Levy, R. *et al.* Longitudinal analysis reveals transition barriers between dominant ecological states
545 in the gut microbiome. *Proceedings of the National Academy of Sciences* 201922498 (2020)
546 doi:10.1073/pnas.1922498117.

547 24. Magdalena, L. Fuzzy Rule-Based Systems. in *Springer Handbook of Computational Intelligence*
548 (eds. Kacprzyk, J. & Pedrycz, W.) 203–218 (Springer, Berlin, Heidelberg, 2015). doi:10.1007/978-
549 3-662-43505-2_13.

550 25. Zaneveld, J. R., McMinds, R. & Vega Thurber, R. Stress and stability: applying the Anna Karenina
551 principle to animal microbiomes. *Nat Microbiol* **2**, 1–8 (2017).

552 26. Gates, A. J., Wood, I. B., Hetrick, W. P. & Ahn, Y.-Y. Element-centric clustering comparison
553 unifies overlaps and hierarchy. *Sci Rep* **9**, 8574 (2019).

554 27. Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a
555 phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic*
556 *Acids Research* **50**, D785–D794 (2022).

557 28. Schoch, C. L. *et al.* NCBI Taxonomy: a comprehensive update on curation, resources and tools.
558 *Database (Oxford)* **2020**, baaa062 (2020).

559 29. Foley, M. H., Cockburn, D. W. & Koropatkin, N. M. The Sus operon: a model system for starch
560 uptake by the human gut Bacteroidetes. *Cell. Mol. Life Sci.* **73**, 2603–2617 (2016).

561 30. Koropatkin, N. M., Cameron, E. A. & Martens, E. C. How glycan metabolism shapes the human
562 gut microbiota. *Nat Rev Microbiol* **10**, 323–335 (2012).

563 31. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and
564 Projection. *Journal of Open Source Software* **3**, 861 (2018).

565 32. Armstrong, G. *et al.* Uniform Manifold Approximation and Projection (UMAP) Reveals Composite
566 Patterns and Resolves Visualization Artifacts in Microbiome Data. *mSystems* (2021)
567 doi:10.1128/mSystems.00691-21.

568 33. Kim, P. S. *et al.* Host habitat is the major determinant of the gut microbiome of fish. *Microbiome*
569 **9**, 166 (2021).

570 34. Zmora, N. *et al.* Personalized Gut Mucosal Colonization Resistance to Empiric Probiotics Is
571 Associated with Unique Host and Microbiome Features. *Cell* **174**, 1388-1405.e21 (2018).

572 35. Bartsch, M., Hahn, A. & Berkemeyer, S. Bridging the Gap from Enterotypes to Personalized
573 Dietary Recommendations: A Metabolomics Perspective on Microbiome Research. *Metabolites*
574 **13**, 1182 (2023).

575 36. Jiang, S. *et al.* Alteration of the gut microbiota in Chinese population with chronic kidney disease.
576 *Sci Rep* **7**, 2870 (2017).

577 37. Kriss, M., Hazleton, K. Z., Nusbacher, N. M., Martin, C. G. & Lozupone, C. A. Low diversity gut
578 microbiota dysbiosis: drivers, functional implications and recovery. *Current Opinion in*
579 *Microbiology* **44**, 34–40 (2018).

580 38. Zhang, L. *et al.* Insight into alteration of gut microbiota in Clostridium difficile infection and
581 asymptomatic C. difficile colonization. *Anaerobe* **34**, 1–7 (2015).

582 39. Dahal, R. H., Kim, S., Kim, Y. K., Kim, E. S. & Kim, J. Insight into gut dysbiosis of patients with
583 inflammatory bowel disease and ischemic colitis. *Frontiers in Microbiology* **14**, (2023).

584 40. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd*
585 *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794
586 (Association for Computing Machinery, New York, NY, USA, 2016).
587 doi:10.1145/2939672.2939785.

588 41. Vieira-Silva, S. *et al.* Quantitative microbiome profiling disentangles inflammation- and bile duct
589 obstruction-associated microbiota alterations across PSC/IBD diagnoses. *Nat Microbiol* **4**, 1826–
590 1831 (2019).

591 42. Si, J. *et al.* Long-term life history predicts current gut microbiome in a population-based cohort
592 study. *Nat Aging* **2**, 885–895 (2022).

593 43. Alili, R. *et al.* Characterization of the Gut Microbiota in Individuals with Overweight or Obesity
594 during a Real-World Weight Loss Dietary Program: A Focus on the Bacteroides 2 Enterotype.
595 *Biomedicines* **10**, 16 (2022).

596 44. Adamberg, S. & Adamberg, K. Prevotella enterotype associates with diets supporting acidic faecal
597 pH and production of propionic acid by microbiota. *Heliyon* **10**, (2024).

598 45. Bresser, L. R. F., de Goffau, M. C., Levin, E. & Nieuwdorp, M. Gut Microbiota in Nutrition and
599 Health with a Special Focus on Specific Bacterial Clusters. *Cells* **11**, 3091 (2022).

600 46. FAO.FAOSTAT License: CC BY-NC-SA 3.0 IGO. Extracted from:
601 https://www.fao.org/faostat/en/#data/FBS/report. Data of Access: 17-05-2024.

602 47. Leeming, E. R., Johnson, A. J., Spector, T. D. & Le Roy, C. I. Effect of Diet on the Gut Microbiota:
603 Rethinking Intervention Duration. *Nutrients* **11**, 2862 (2019).

604 48. Leeming, E. R. *et al.* The complexities of the diet-microbiome relationship: advances and
605 perspectives. *Genome Medicine* **13**, 10 (2021).

606 49. Abdill, R. J., Adamowicz, E. M. & Blekhman, R. Public human microbiome data are dominated by
607 highly developed countries. *PLOS Biology* **20**, e3001536 (2022).

608 50. Schmidt, T. S. B. *et al.* SPIRE: a Searchable, Planetary-scale mIcrobiome REsource. *Nucleic Acids*
609 *Research* gkad943 (2023) doi:10.1093/nar/gkad943.

610 51. Ramirez, J. *et al.* Antibiotics as Major Disruptors of Gut Microbiota. *Frontiers in Cellular and*
611 *Infection Microbiology* **10**, (2020).

612 52. Fernandes, M. R. *et al.* Alterations of Intestinal Microbiome by Antibiotic Therapy in Hospitalized
613 Children. *Microbial Drug Resistance* **23**, 56–62 (2017).

614 53. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.
615 *Journal of Computational and Applied Mathematics* **20**, 53–65 (1987).

616 54. Martin Morgan. DirichletMultinomial: Dirichlet-Multinomial Mixture Model Machine Learning
617 for Microbiome Data. (2020).

618 55. Serafini, A., Giordani, P. & Ferraro, M. fclust: An R Package for Fuzzy Clustering. *The R Journal*
619 **9**, (2019).

56. Dray, S. & Dufour, A.-B. The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software* **22**, 1–20 (2007).

57. R Core Team. R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing (2023).

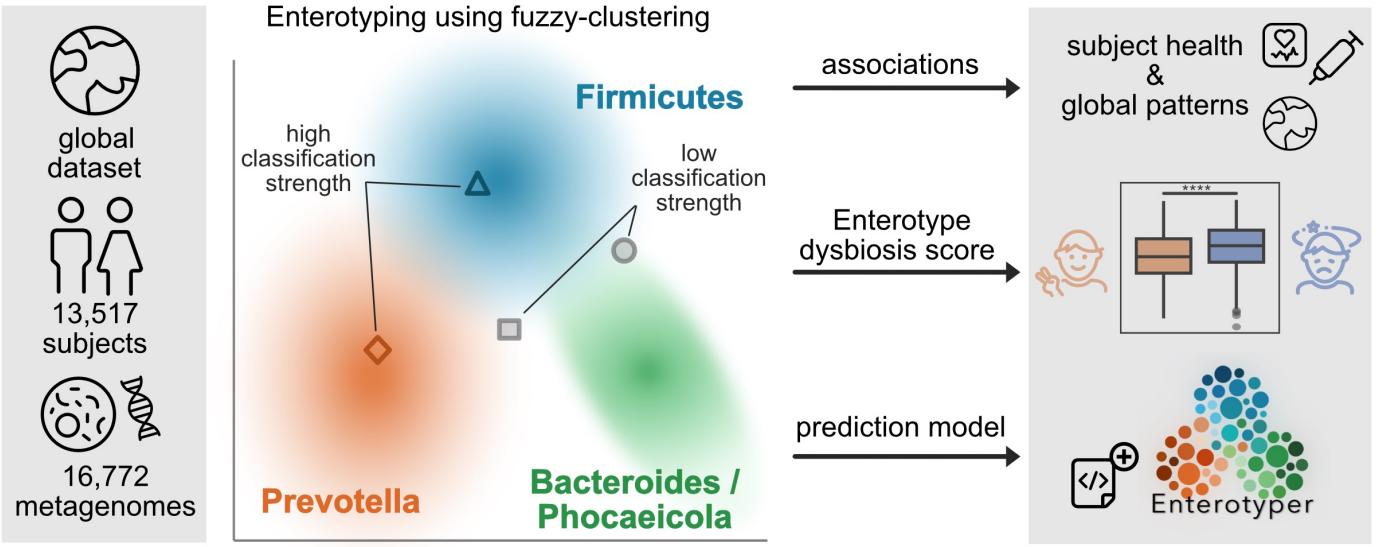
58. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

59. Oksanen, J. *et al.* vegan: Community Ecology Package. (2024).

60. Spedicato, G. A. Discrete Time Markov Chains with R. *The R Journal* **9**, 84–104 (2017).

61. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* **67**, 1–48 (2015).

62. Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* **28**, 1–26 (2008).



Graphical Abstract

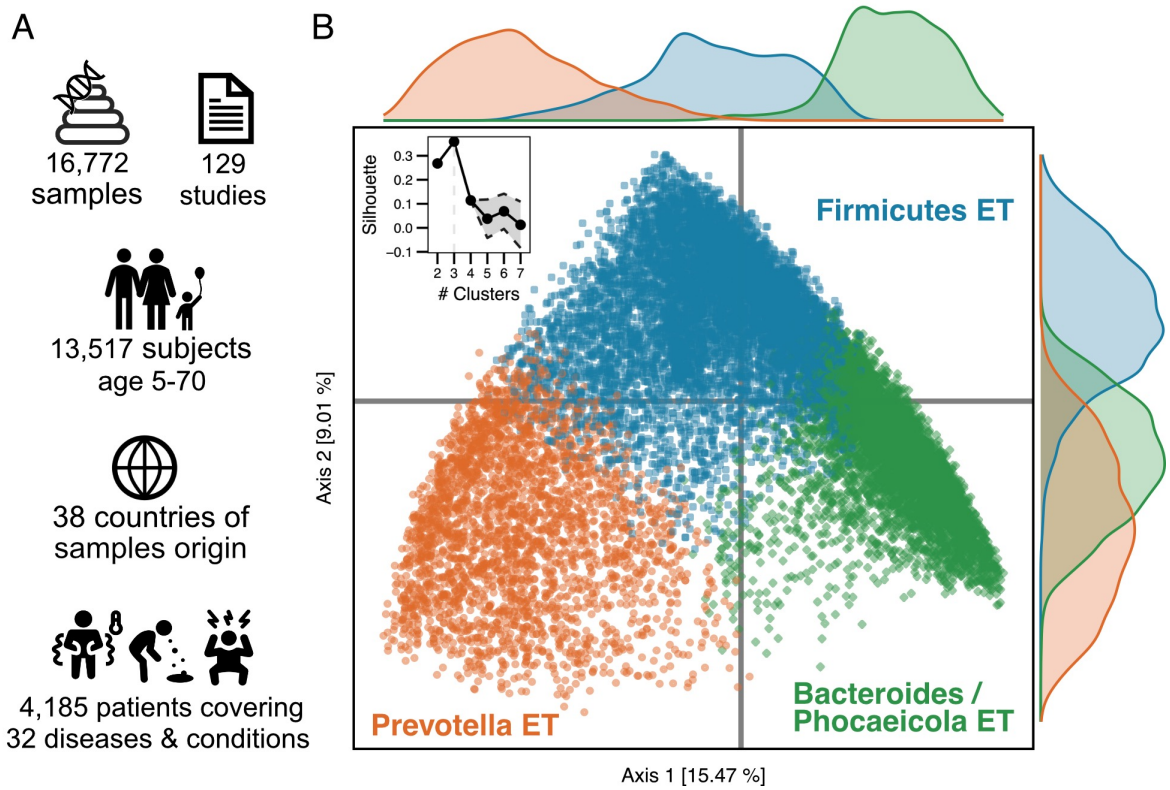


Figure 1: Data and enterotype classification overview. A. Data overview on our global dataset. **B.** PCoA generated from the JSD matrix. Each sample is colored based on its enterotype classification. Density plots indicate the distribution of points along the first two principal components for each enterotype. The inset plot in the upper left corner displays the silhouette width and grey confidence intervals for 20 repetitions of the algorithm.

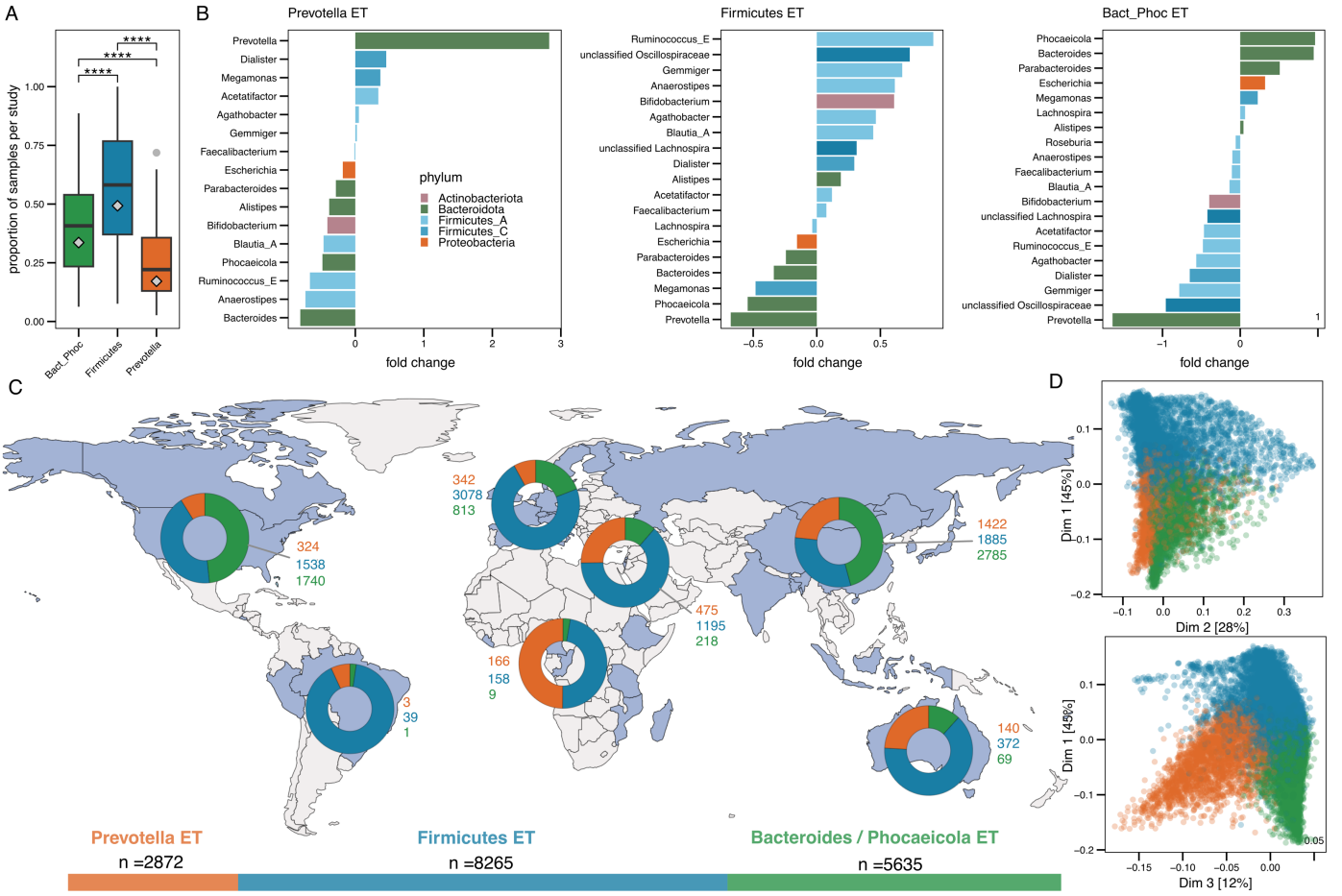


Figure 2: Enterotypes identified through FKM clustering. A. The proportion of samples classified into enterotypes for each individual study, with the proportion in the total cohort indicated by grey diamonds. **** p <=0.0001 **B.** Fold change from differential abundant genera between samples classified and not classified into each enterotype. Only the top significant genera after Benjamini-Hochberg p-value correction for multiple hypotheses are shown. **C.** World map showing the global enterotype distribution with the numbers of samples classified into each enterotype reported next to the donut plots. The interactive version of that map can be explored on our webpage https://enterotype.embl.de/, which includes more information about included studies, countries, samples, and country-wise enterotype distributions. The bar plot at the bottom shows the Enterotype classification distribution for the whole global dataset **D**. Multidimensional scaling plots with dimensions one and two (upper plot) and dimensions one and three (lower plot), calculated on the functional profiles with relative abundances of KEGG orthologs for each sample. The colors of the points correspond to the enterotypes, as reported in panel C.

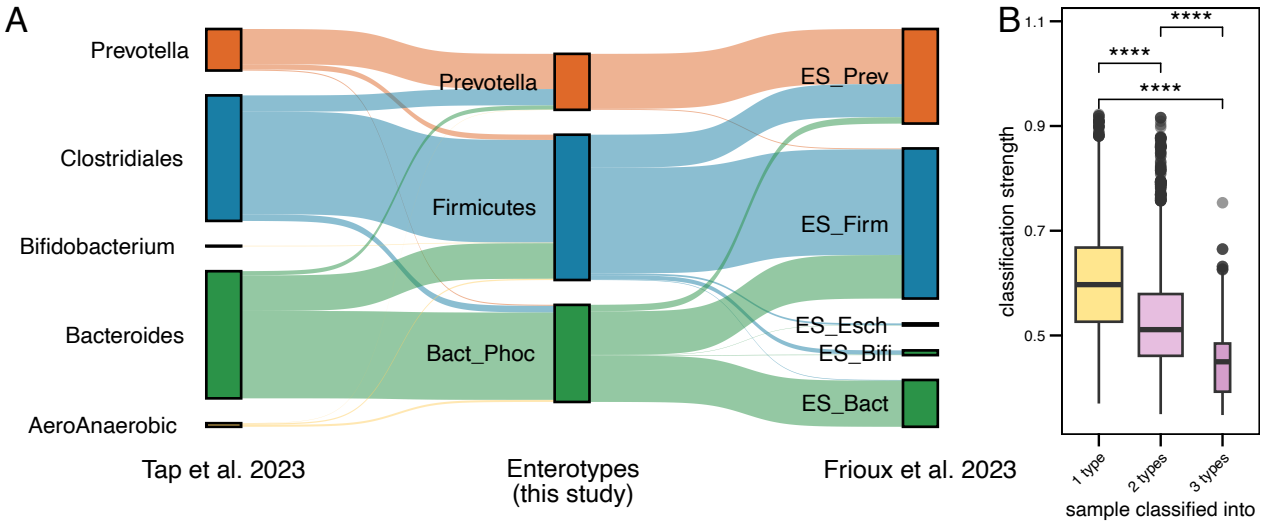


Figure 3: Enterotypes are robustly detected across different methodologies A. Comparison of the enterotype classifications for the overlapping samples between this study and Tap et al. 2023 and Frioux et al. 2023 (n = 5,668), which used alternative approaches to define compositional groups in the human fecal microbiome. **B.** The stability in the enterotype classification among the different studies in panel A is associated with FKM-derived classification strength, showing lower strength for samples with changing classifications. 2,749 samples are stable, 2,430 samples change once, and 295 samples change twice in their enterotype classification.

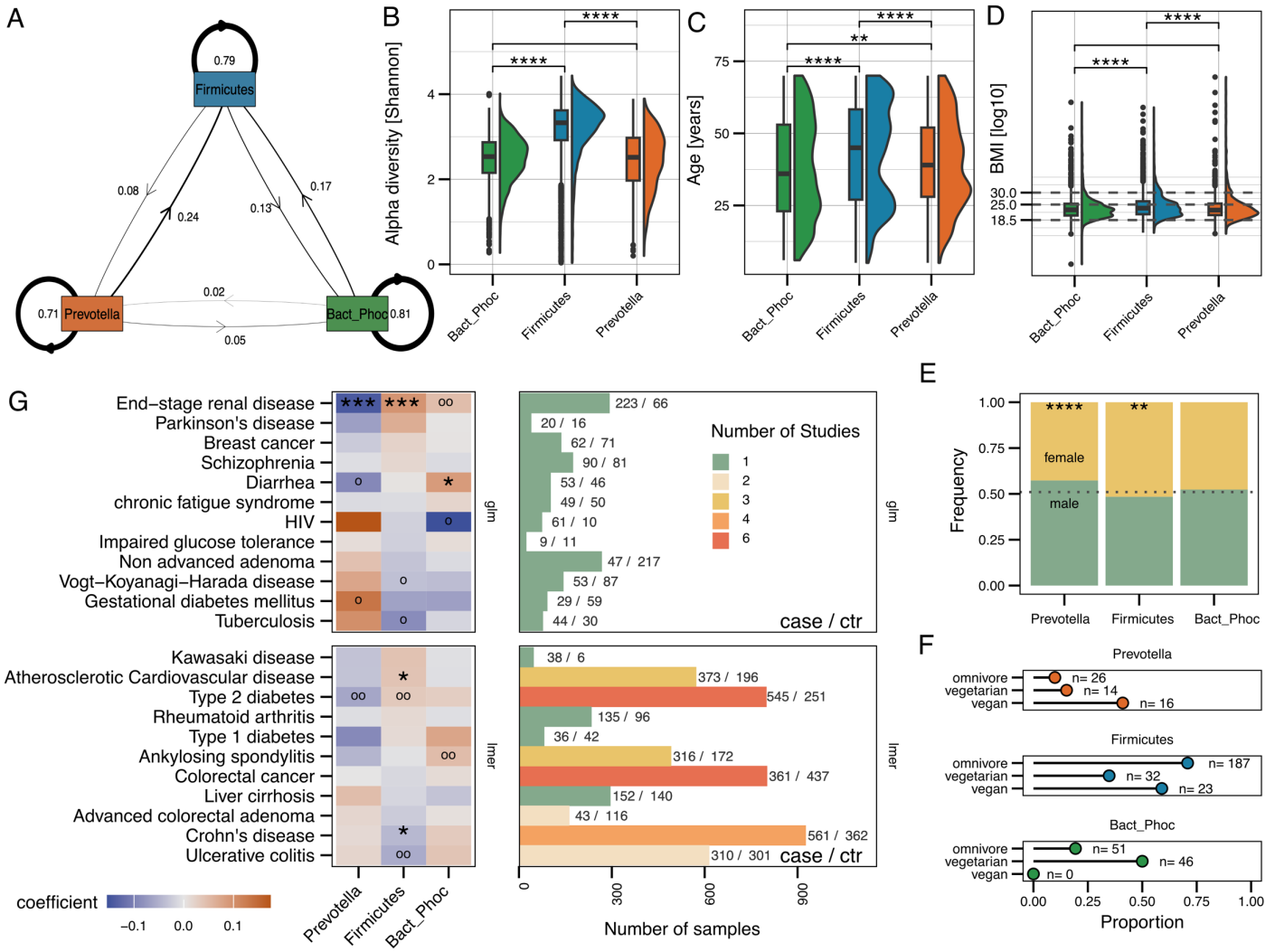


Figure 4: Stability of enterotype and associations with demographic measures and diseases. A. A Markov chain model estimation of the likelihood of transitioning between enterotypes for subjects who provided two or more samples within one year. **B. – D.** Shannon diversity, age, and BMI comparisons between the samples classified into different enterotypes. The cohort was filtered for all samples with information on age (n=7,432) and BMI (n=5,036). Differences between the groups were calculated using the Wilcoxon test. *p <= 0.05, **p <= 0.01, ***p <= 0.001, ****p <= 0.0001 **E.** Distribution of 7,973 subjects with sex information in the metadata. Using a chi-squared goodness-of-fit test, we checked if the observed ratio for each enterotype deviated from the expected ratio in our cohort (0.49:0.51, dotted line)**.** **p <= 0.01, ****p <= 0.0001. **F.** The proportion of samples from subjects reporting a dietary pattern, such as omnivore, vegetarian, and vegan, classified into the enterotypes. The numbers of samples are noted in the plot. **G.** Associations between diseases or symptoms and the probability for the corresponding enterotype were tested using generalized linear models (glm) with additional adjusting for random effects using linear mixed effect models (lmer) to account for longitudinal sampling and study effects where appropriate. A positive coefficient (orange) refers to a positive correlation between enterotype affiliation and disease or condition for a given sample. In contrast, a negative coefficient (blue) means that samples from subjects with this disease or condition are less likely to be classified to this enterotype. For each disease or condition, the models were calculated on available disease and control studies, and the number of cases and controls and the number of studies are reported in the bar plot. ° p <= 0.05, °° p <= 0.01, * q <= 0.05, *** q <= 0.001; q=adjusted p-values using Benjamini-Hochberg correction.

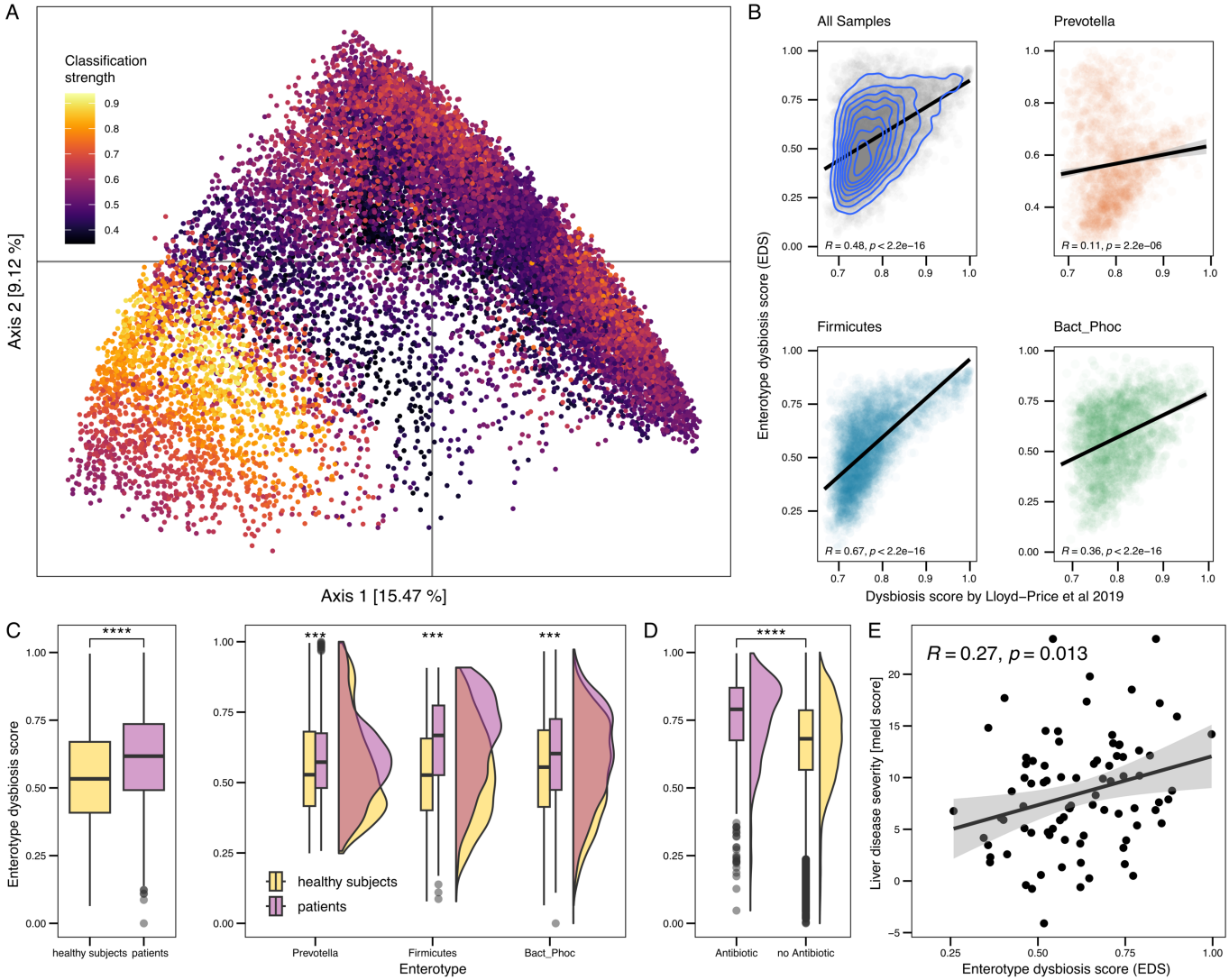


Figure 5: The enterotype dysbiosis score (EDS) is derived from the enterotype classification strength in the FKM clustering. A. Principal coordinates analysis (PCoA) plot based on JSD from genus-level GTDB taxonomic profiles. Each sample is colored based on its enterotype strength. The samples with higher probabilities visualize the "centers" of the enterotypes in the two-dimensional space. **B.** The enterotype strength z-score normalized, reversed, and scaled between zero and one serves as an enterotype dysbiosis score that correlates with an already established one[6]**. C.** The samples from subjects with a reported disease (n = 3,399) showed a higher enterotype dysbiosis score than those from subjects with no reported disease (n = 8,672). *** p<= 0.001, **** p <= 0.0001. **D.** Subjects who reported antibiotic use (n=1,038) showed higher EDS. Although these samples are not included in our global dataset, they were classified alongside it to study the impact of including such samples on enterotype clustering statistics, as shown in Figure S4. **E.** Spearman correlation of the EDS with the MELD (model of end-stage liver disease) score, reported for 181 subjects in Qin et al.[65], as an estimate of liver disease severity.

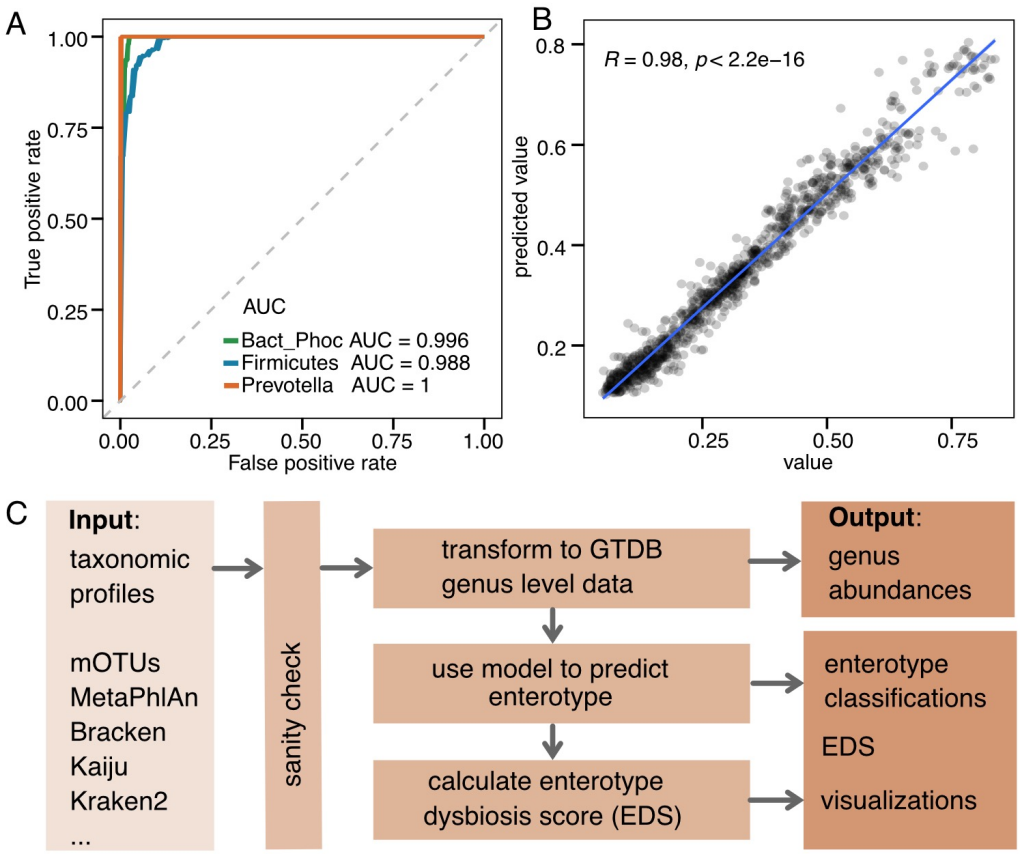


Figure 6: The "Enterotyper" online tool accurately predicts the enterotypes and EDS of independent cohorts. A. The receiver operating characteristic (ROC) curve was estimated on the validation dataset for each enterotype using the default prediction model based on the three FKM enterotypes. **B.** Correlation of predicted value and expected value for the enterotype strength for each sample in the default FKM three enterotypes estimated on the validation dataset. **C.** Schematic and automated workflow for the "Enterotyper" from a user perspective.

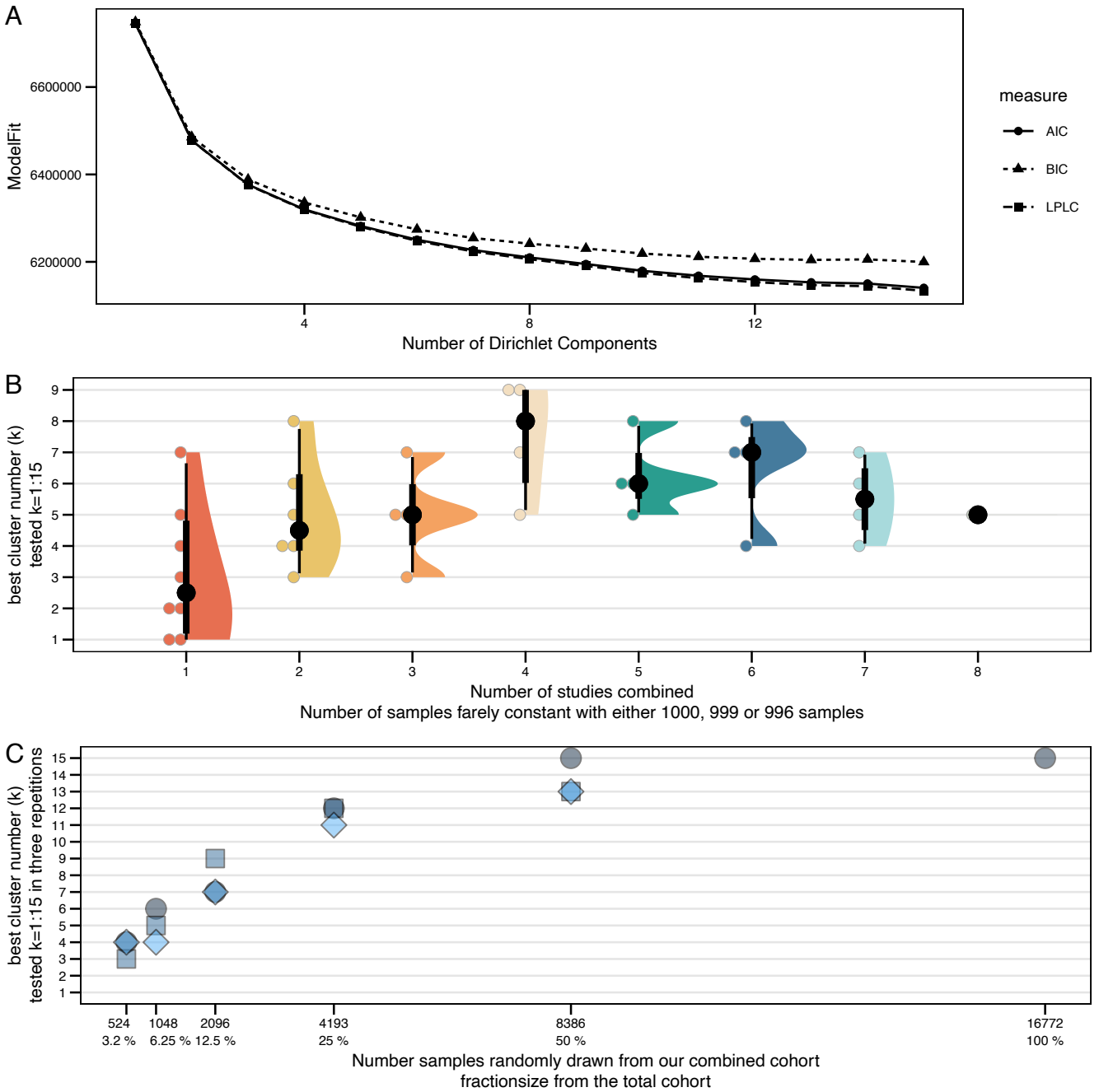


Figure S1: DMM Clustering. A. The DMM clustering on the total cohort showed no local minimum using the Laplace approximation, AIC, or BIC when evaluating the clusters up to 15 clusters. **B.** Benchmarking the DMM clustering with increasing heterogeneity by adding samples from increasing numbers of studies in equal proportions. The number of samples was kept almost the same throughout the benchmarking. **C.** Benchmarking the DMM clustering with an increasing number of samples randomly picked from our cohort to keep a similar level of heterogeneity. The selection process was repeated three times.

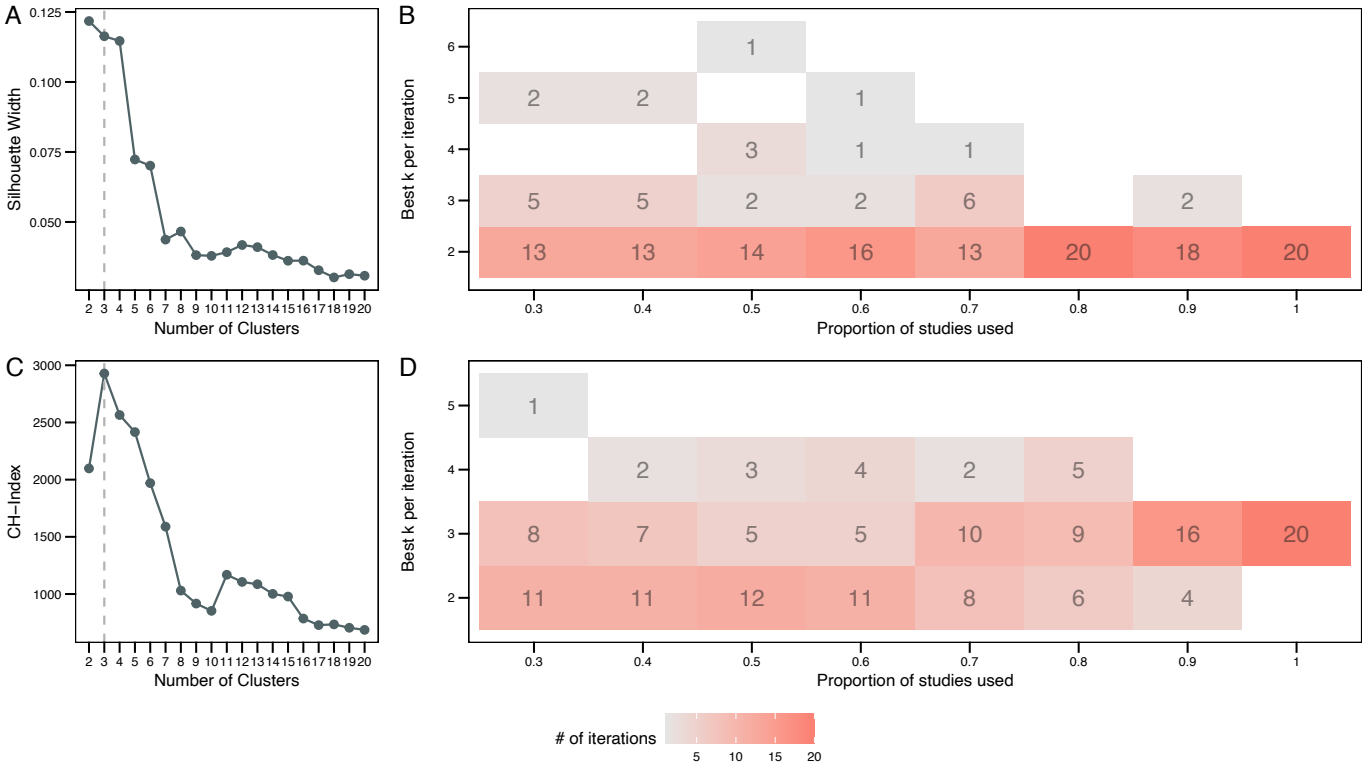


Figure S2: Statistics to determine the optimal number of enterotypes using the PAM clustering. A + C. Silhouette width (A) and CH-index (B) for up to 20 possible clusters **B + D.** Rarefaction analysis showing the optimal number of clusters determined by silhouette width (B) and CH-Index (C) for samples from a subset of studies representing 30% - 100% of the studies. This selection process was repeated 20 times for each percentage (except 100%), and the PAM clustering workflow was performed on the subset. The best cluster number, k, was chosen by selecting the cluster number that resulted in the highest CH-index or silhouette width, respectively.

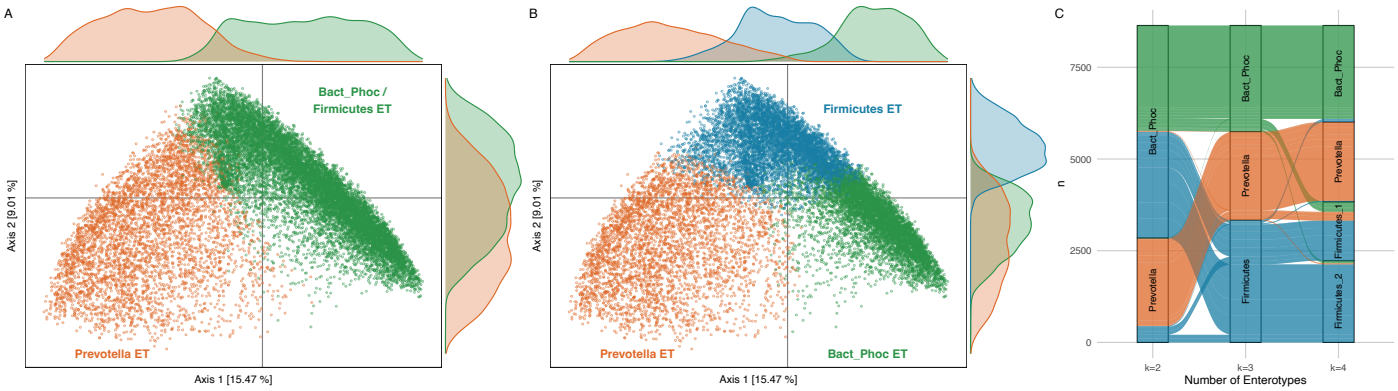


Figure S3: Visualizing enterotype classifications using PAM clustering. A-B. Principal coordinates analysis (PCoA) plots derived from the JSD matrix based on genus-level profiles. Samples are colored according to their enterotype classification, as determined by two enterotypes (A) and three enterotypes (B) using PAM clustering. Density plots illustrate the distribution of points along the first two principal components for each enterotype. **C.** Alluvial diagram depicting the consistency of enterotype classification for each sample across the two-, three-, and four-enterotype PAM clustering models. Each line represents one sample and is colored based on the three-enterotype model

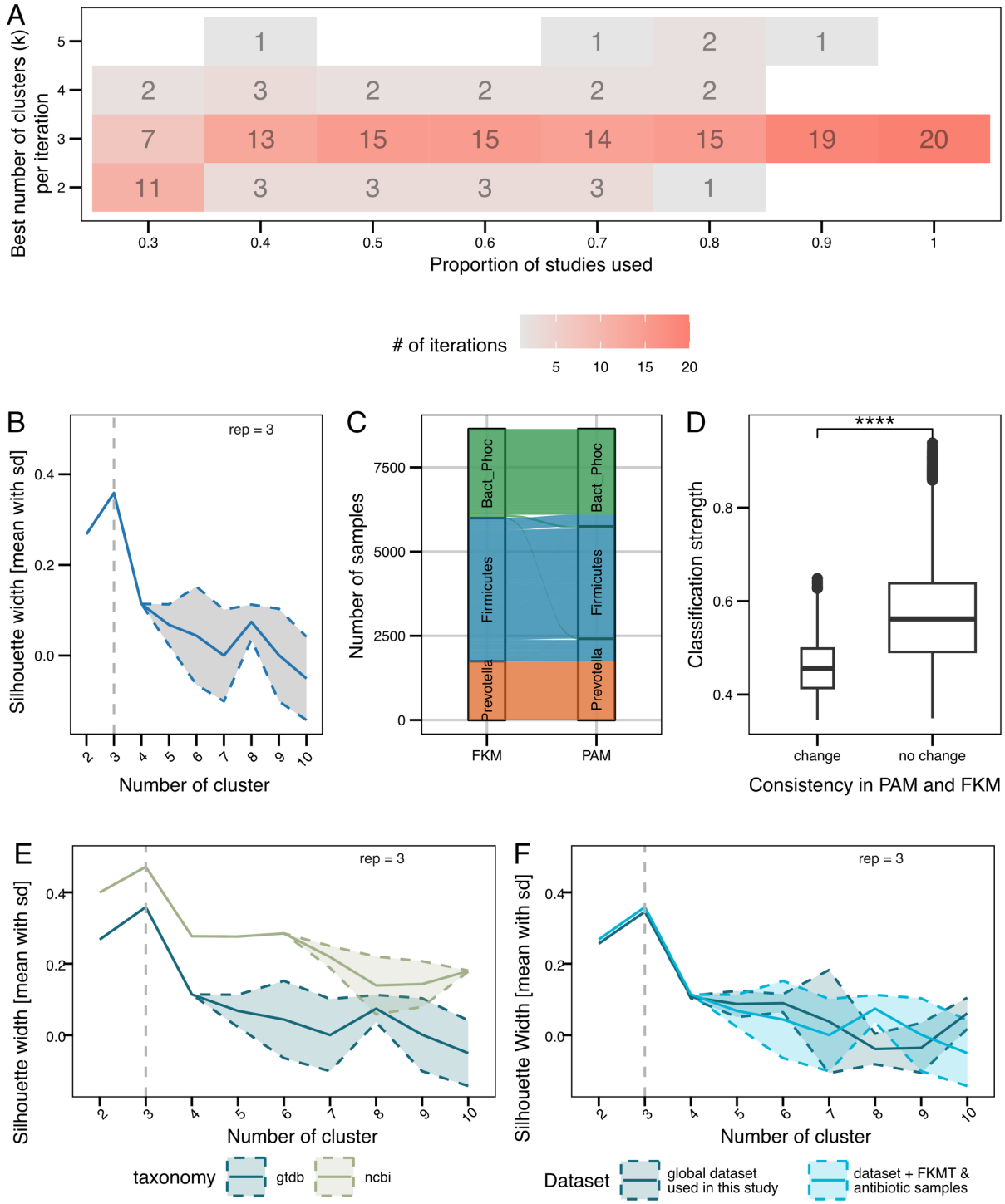


Figure S4: Statistics to determine the optimal number of enterotypes using the FKM clustering. A. Rarefaction analysis showing the optimal cluster number for samples from subsets of studies representing 30% - 100 % of the studies. This selection process and FKM clustering were repeated 20 times for each percentage (selection in 100% cohort not applicable). The best cluster number was chosen by selecting the number that resulted in the highest silhouette width **B.** Silhouette width mean values with confidence intervals for different cluster numbers tested by the FKM algorithm. **C.** Alluvial diagram demonstrating the overlap between enterotype classifications derived from FKM and PAM clustering methods. Each line represents one sample and is colored according to the three-enterotype FKM clustering model. **D.** Classification strength derived from the FKM clustering comparing samples that change their enterotype classification between FKM and PAM clustering and samples that do not. **E.** Silhouette width derived from the FKM clustering on GTDB and NCBI based taxonomic profiles. **F.** Silhouette width derived from the FKM clustering on the global dataset and on the dataset together with samples with associated antibiotic usage or fecal microbiome transplantation (FKM)

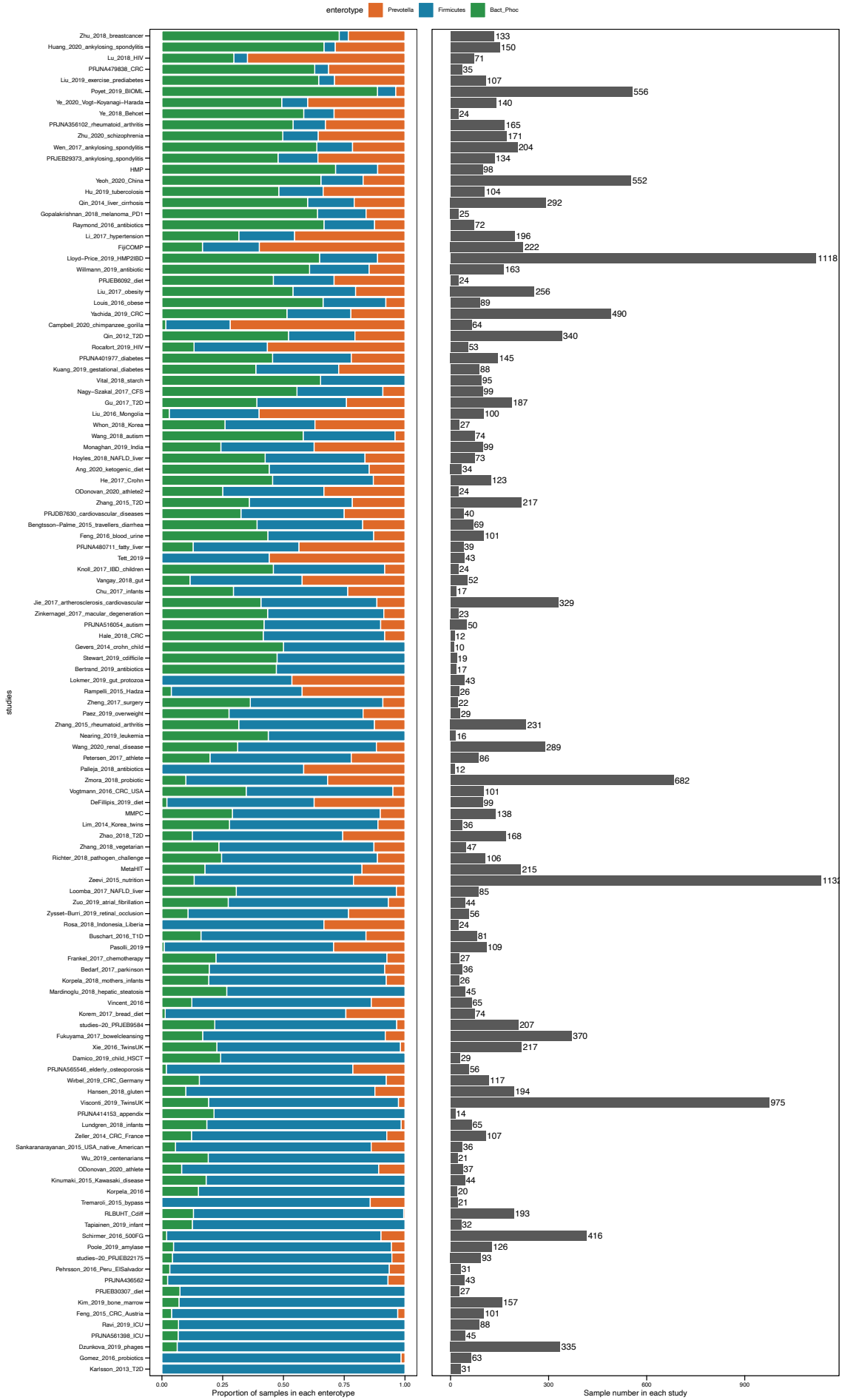


Figure S5: **The proportion of samples from each study classified to particular enterotypes.** The stacked bar plots show the distribution of samples classified according to the respective enterotype. The right grey bar indicates the number of samples in that study.

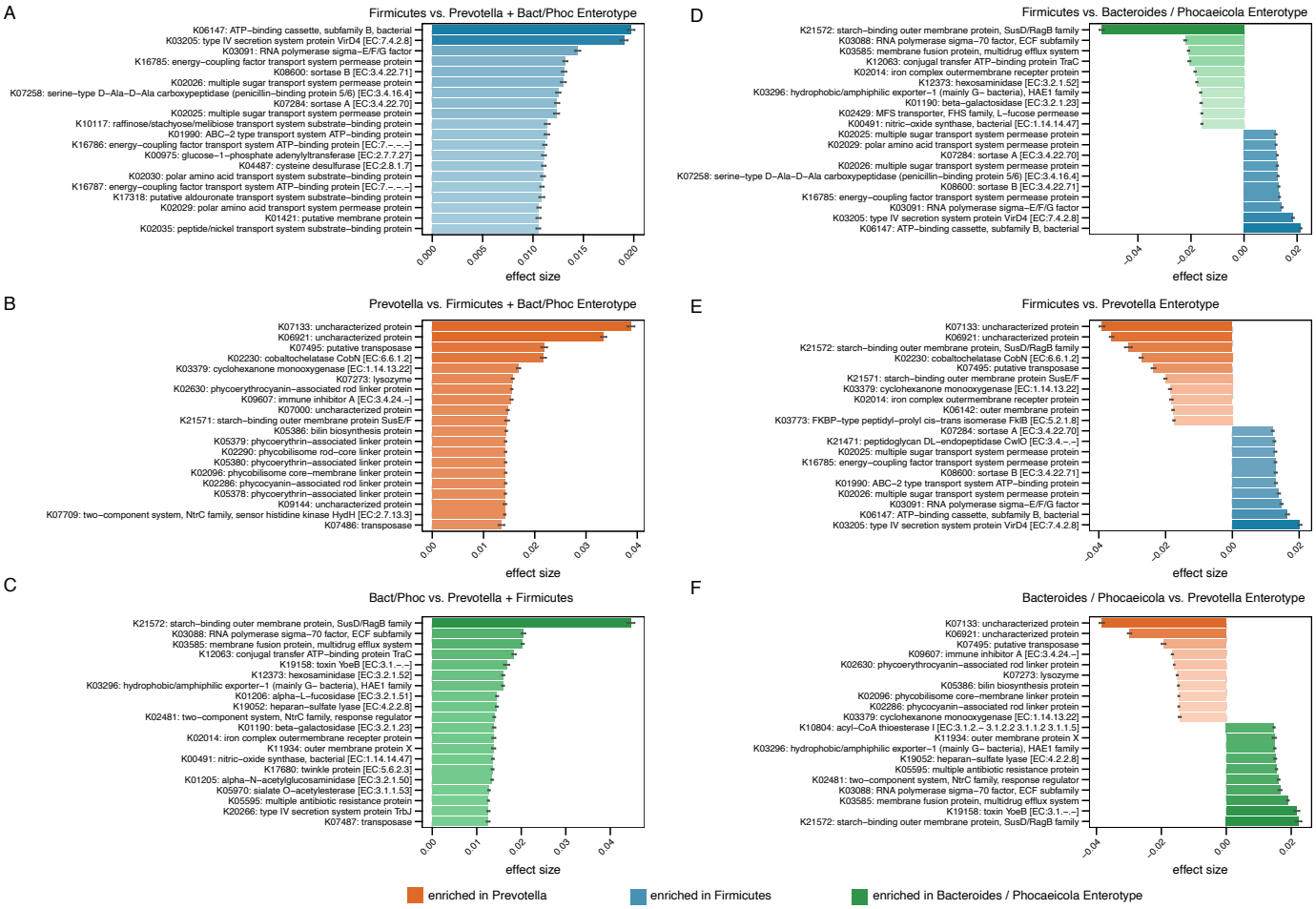


Figure S6: Functional profiles shown as enriched KEGG-orthologs (KO) for each enterotype. Effect sizes with confidence intervals derived from differential abundance analysis using the Wilcoxon test with Benjamini-Hochberg p-value adjustment for false discovery rate on the relative abundances of KOs. We compared the samples classified with each enterotype with those not classified into this enterotype (**A-C**). In addition, and inspired by the sample distribution in the MDS plot (Figure 2D), we compared the KO abundances between the Firmicutes enterotype with the Bacteroides/Phocaeicola (**D**) or Prevotella (**E**) enterotype and directly the Bacteroides/Phocaeicola with the Prevotella enterotype (**F**). The colors of the bars represent the enterotype in which this KO shows an enrichment. Only the top KO with significant changes reported by an adjusted p-value < 0.01 are shown.

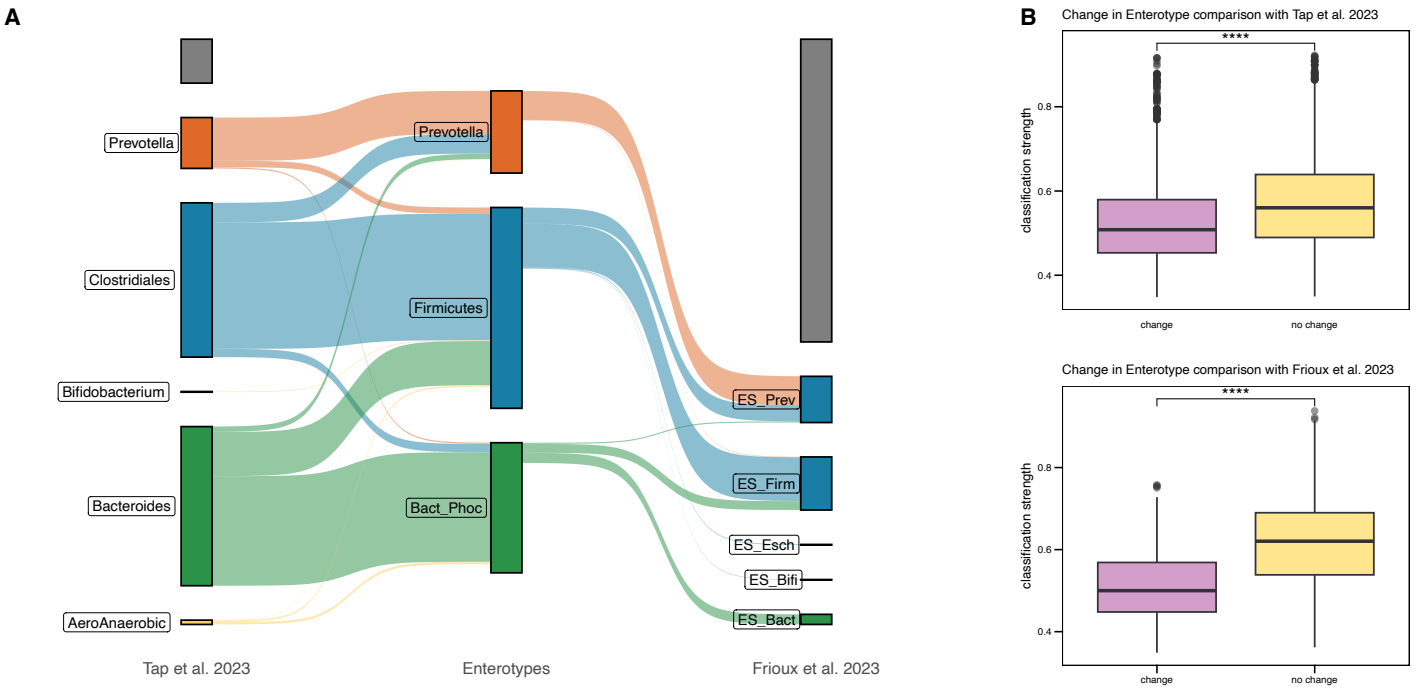


Figure S7: Enterotypes comparison with Tap et al. 2023 and Frioux et al. 2023. A. Comparison of the Enterotype classification from this study with the overlapping samples from Tap et al. 2023 (left, n = 5,668) and Frioux et al. 2023 (right, n = 18,012). The grey bars on the top left indicate the number of samples not at the intersection between Tap and Frioux samples. **B.** Samples that changed their enterotype assignment showed lower FKM-derived classification strength. **** $p <= 0.0001$.

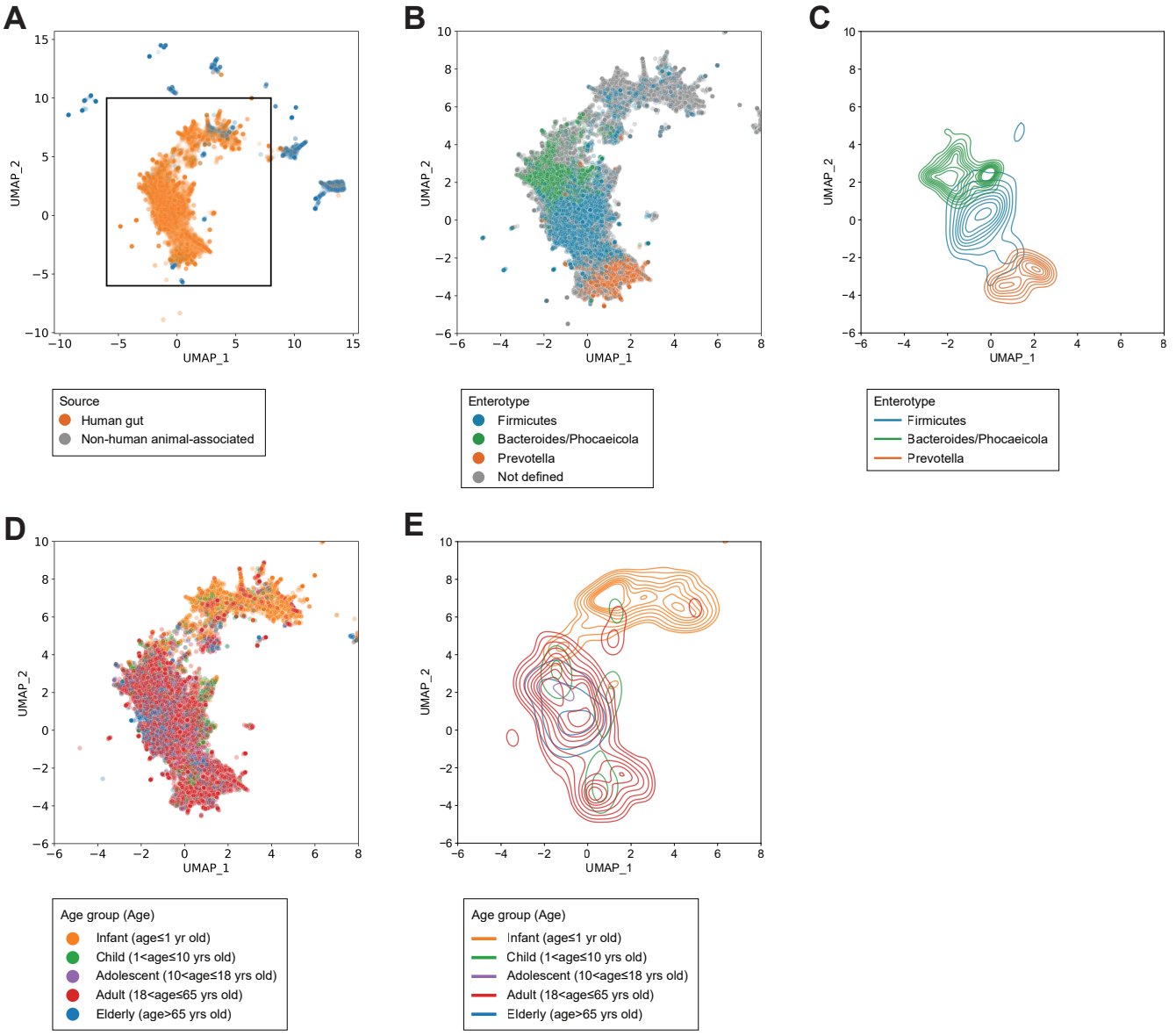


Figure S8: Imposing of enterotypes on UMAP based on the global dataset together with other microbiome sample types A. UMAP plot of 49,762 metagenomic samples. The black square highlighted the UMAP space where the human fecal samples are enriched. **B**. UMAP plot colored by enterotype of the samples. **C**. Contour plot representing the sample density of each enterotype. **D**. UMAP plot colored by the age group of the human fecal samples. **E**. Contour plot representing the sample density of each age group.

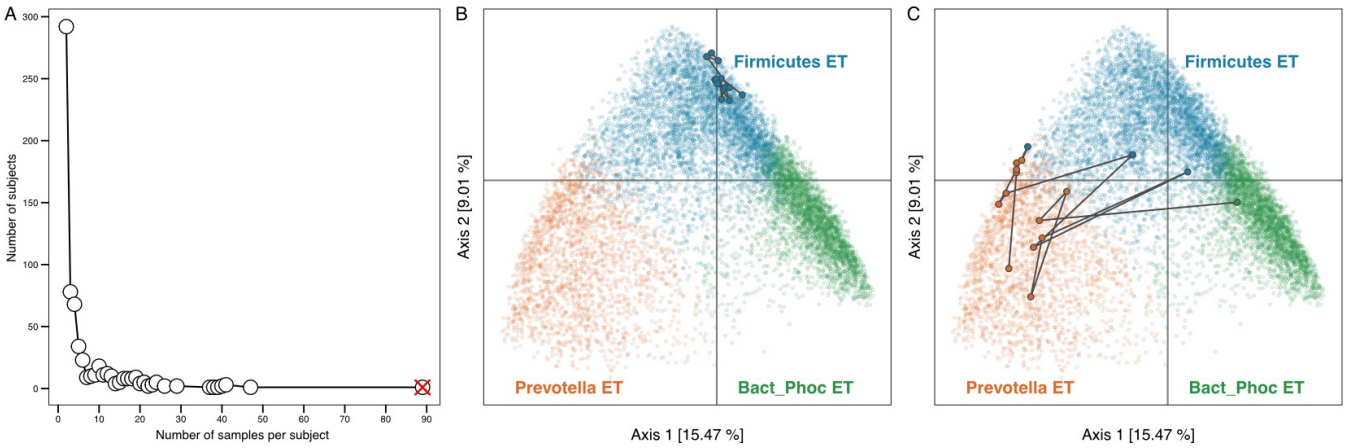


Figure S9: Analysis of enterotype classification stability A. The number of subjects with multiple samples. We excluded one subject with 89 samples, which was excluded as an outlier from the analysis using the Markov chain model (Figure 4). **B & C.** Examples of two opposite cases for longitudinal behavior of the subjects' microbiome composition: a very stable example (B) and an unstable (C) case.

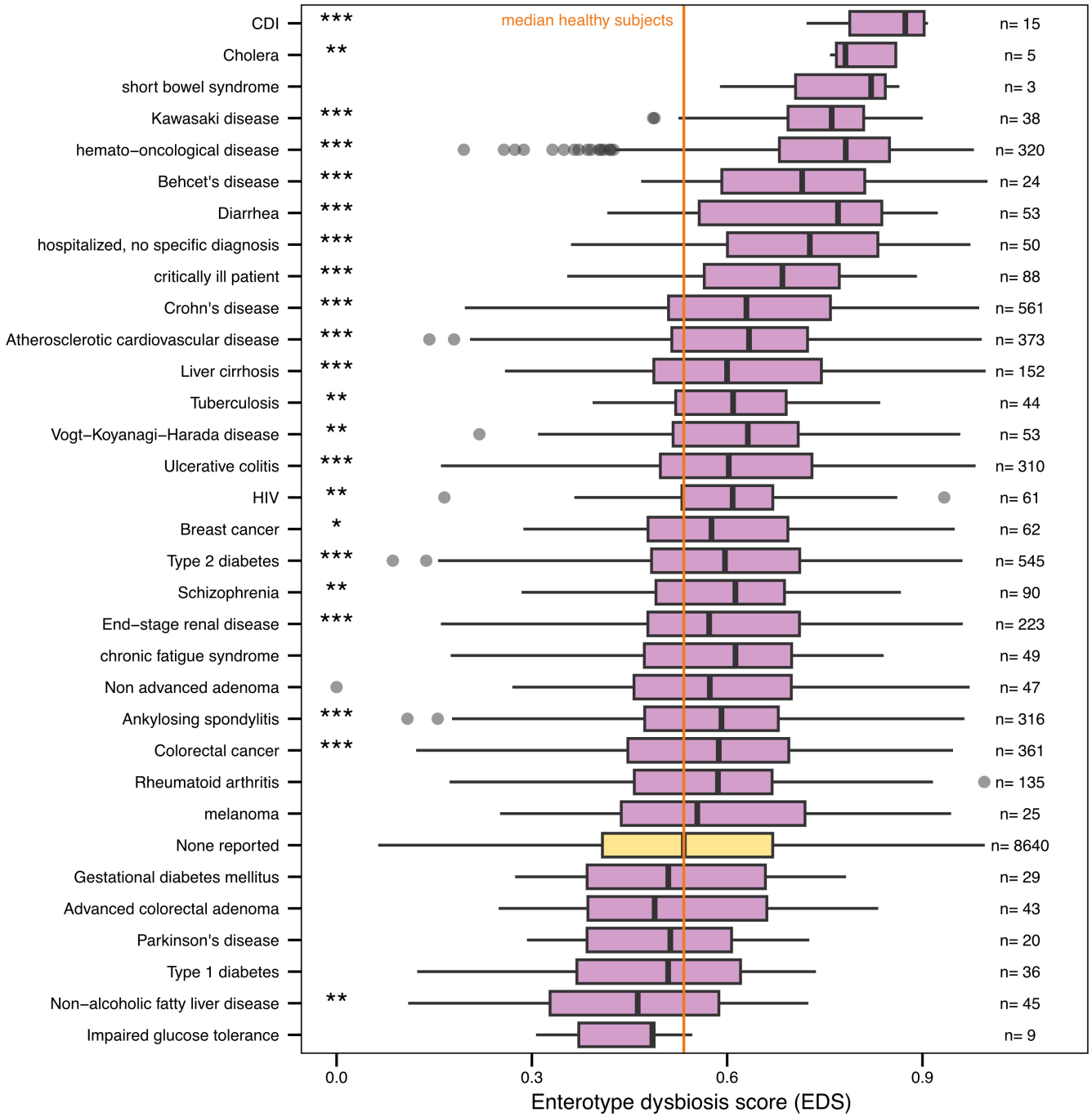


Figure S10: Enterotype dysbiosis score (EDS) by diseases and symptoms. The EDS for all samples associated with a disease or symptom compared to the EDS for samples with reported absence of disease (yellow). The difference in EDS from a disease group to the group without disease was calculated using the Wilcoxon test and adjusted for multiple testing using Benjamini-Hochberg correction. *q <= 0.05, **q <= 0.01, ***q <= 0.001.

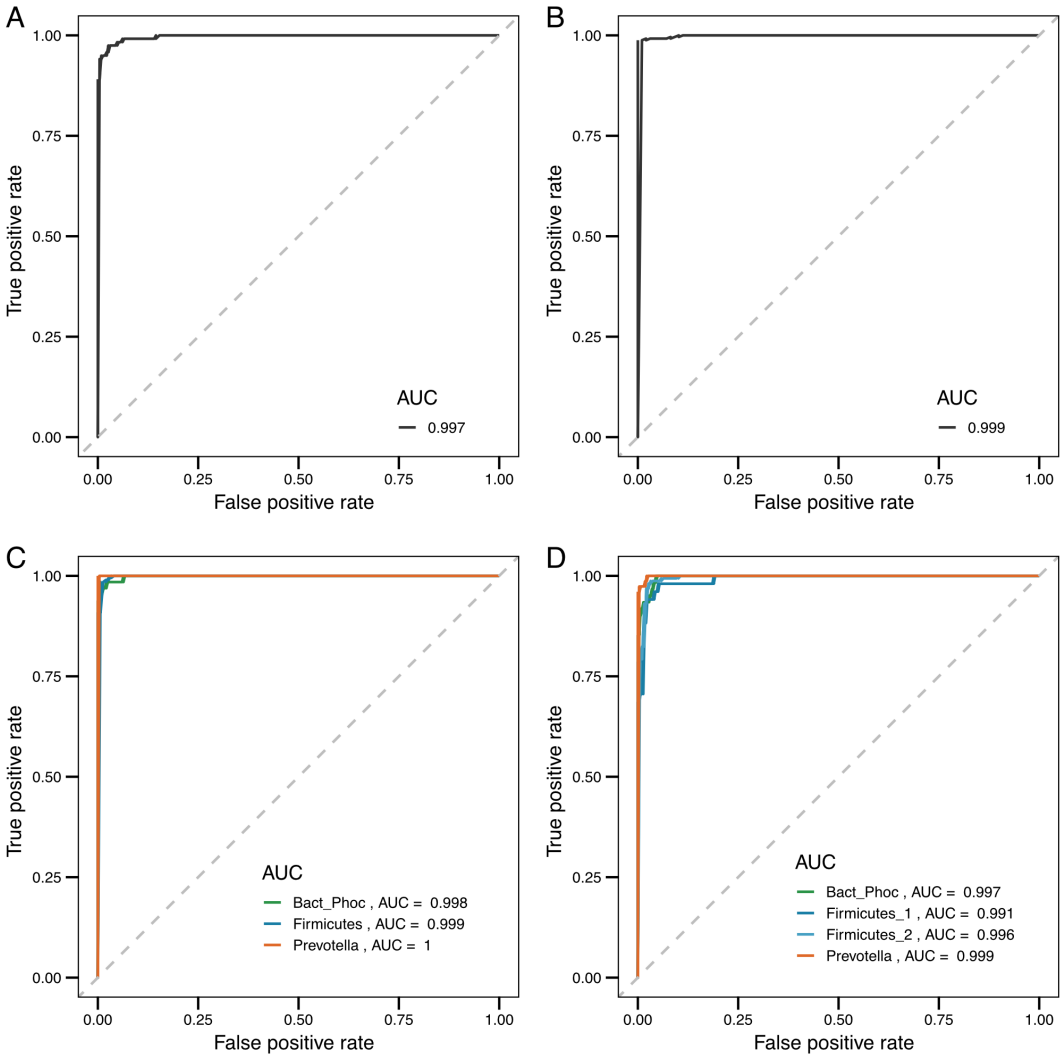


Figure S11: AUCs for the XGBoost models in the "Enterotyper." A. The AUC for the FKM-based enterotype classification prediction for the two enterotypes. **B – D**. AUCs for PAM-based enterotype classification prediction for two (B), three (C), and 4 (D) enterotypes. All AUCs have been calculated based on the prediction of the validation dataset.