Population-based metaproteomics reveals functional associations between gut microbiota and phenotypes

Shuang Liang[1,2,3,4#], Yingying Sun[1,2,3#], Zelei Miao[1,2,3,5#], Bang-yan Li[6#], Ziyuan Xing[1,2,3], Yuting Xie[1,2,3], Enci Cai[2], Sainan Li[1,2,3], Pu Liu[7], Min Yang[2], Menglei Shuai[1,2,5], Wanglong Gou[1,2,3,5], Wenhao Jiang[1,2,3], Youming Wang[8], Huanhuan Gao[1,2,3], Ke Zhang[1,2,5], Jing Yu[7], Xue Cai[1,2,3], Xingbing Wang[8], Yu-ming Chen[6]*, Ju-Sheng Zheng[1,2,3,5]*, Tiannan Guo[1,2,3,9]*

1 Affiliated Hangzhou First People's Hospital, State Key Laboratory of Medical Proteomics, School of Medicine, Westlake University, Hangzhou, Zhejiang Province, China.

2 Westlake Center for Intelligent Proteomics, Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, Zhejiang Province, China.

3 Research Center for Industries of the Future, School of Life Sciences, Westlake University, Hangzhou, Zhejiang Province, China.

4, State Key Laboratory for Managing Biotic and Chemical Treats to the Quality and Safety of Agro-products, Zhejiang Academy of Agricultural Sciences, Hangzhou 310021, China.

5, Zhejiang Key Laboratory of Multi-Omics in Infection and Immunity, Center for Infectious Disease Research, School of Medicine, Westlake University, Hangzhou, China

6, Department of Epidemiology, Guangdong Provincial Key Laboratory of Food, Nutrition and Health, School of Public Health, Sun Yat-sen University, Guangzhou, China

7, Westlake Omics (Hangzhou) Biotechnology Co., Ltd., No. 1 Yunmeng Road, Cloud Town, Xihu District, Hangzhou 310024, Zhejiang Province, China

8, First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui 230001, China.

9, Lead Contact

# These authors contribute equally

*Correspondence: chenyum@mail.sysu.edu.cn (Y.C.); zhengjusheng@westlake.edu.cn (J.Z.); guotiannan@westlake.edu.cn (T.G.)

Highlights

- The first large-scale population-based metaproteomics landscape.
- Metaproteomics efficiently maps microbial functions, exposing individual functional variation.
- 10,714 associations between metaproteomic taxa, microbial functions, or human proteins with 39 phenotypes.
- *Megasphaera elsdenii* protects against elevated blood glucose levels in the host by degrading lactate to produce butyric acid.

Summary

The association between gut microbiota and physiological disorders underscores the crucial need to understand its largely unexplored biological functions. In this study, we analyzed 86,876 microbial proteins and 1,277 human proteins from 1,967 Chinese stool samples. Our results indicate that metaproteomic data captures the microbial biodiversity of a population more effectively than metagenomic data, displaying greater variability in microbial functions and providing reliable assessments of microbiota stability. We identified crucial functions of core microbes and discovered 10,714 associations between metaproteomic taxon, microbial function, or human proteins with 39 phenotypes, including 1,381 features related to Type 2 Diabetes (T2D). Importantly, *Megasphaera elsdenii*'s conversion of lactate to butyric acid appears to lower blood glucose levels, suggesting a protective mechanism against T2D. Our study underscores the power of population-based metaproteomics to offer unique functional insights into the gut microbiota.

Keywords

Metaproteomics; population-based cohort study; human gut microbiome; diabetes; *Megasphaera elsdenii*; butyrate

Introduction

Apart from their well-known roles in food digestion and nutrition absorption, gut microbes are also involved in multiple physiological processes and disorders including metabolic disorders and cancers[1,2]. The gut microbiota serves as intricate communities, and multi-omic technologies are extensively employed to gauge the dynamic fluctuations within these communities. Studies investigating the relationship between gut microbes and factors such as age, sex, diet, geography, heredity, and health status in specific populations have primarily relied on microbial composition analysis using 16S sequencing or metagenomics[3-7]. Metaproteomics has emerged as a powerful tool to explore the functional diversity of the microbiome, as proteins directly govern cellular functions[8,9]. It can reveal taxonomic biomass contribution variations by quantifying taxon-specific proteins[10]. Additionally, metaproteomics also provides information on human enterocyte proteins, facilitating the investigation of host-microbiome interactions[11,12]. Metaproteomics have been successfully employed in studies of inflammatory bowel disease, diabetes, obesity, and colorectal

cancer[13-20]. However, these studies have been restricted to dozens to hundreds of samples primarily due to the relatively long MS time required by the data-dependent acquisition mass spectrometry (DDA-MS) and sophisticated data analysis[21-27]. Therefore, metaproteomics faces challenges in scalability for population-based metaproteome analysis. Due to the highly heterogeneous composition of gut microbiota across individuals[28], more subjects need to be included to obtain confident biological conclusions. Moreover, no study has systematically investigated the unique insights provided by metaproteomics compared to metagenomics.

Data-independent acquisition (DIA) MS coupled with targeted data analysis enables high-throughput profiling of proteomes[29-32]. In particular, DIA leverages DDA to construct a spectral library, facilitating the precise quantification of thousands of proteins with a single injection. Integrating ion mobility with MS provides additional information into ion size, shape, and charge distribution, thereby enhancing ion separation and reducing chemical noise in measurements[33,34]. As a refinement of DIA, parallel accumulation–serial fragmentation combined with data-independent acquisition (diaPASEF) incorporates trapped ion mobility into the ion separation, leading to a significant improvement in proteomic depth[35,36]. In this study, we performed a diaPASEF-based metaproteomic analysis on a cohort comprising thousands of individuals. The distribution of age, gender, lifestyle, and metabolic diseases in this cohort was representative of the middle-aged and elderly Chinese population. Our findings from population-scale gut metaproteomics offer unique functional insights.

Results

Metaproteomic profiling of GNHS cohort

Our study included 1,399 middle-aged and elderly participants from the Guangzhou Nutrition and Health Study (GNHS). Fecal specimens were collected from these individuals, with 568 participants providing samples at baseline and follow-up time points, separated by a median interval of 3.2 years, over a period ranging from 1.4 to 5.5 years. The remaining 831 participants provided samples only once. Altogether, we collected 1,967 fecal samples, each with comprehensive metadata recorded at the time of sampling (Fig. 1A). The metadata comprised 49 clinical and questionnaire-based phenotypes related to lifestyle, demographics, diet, biomedical measures, anthropometrics, metabolic disorders, self-reported diseases, medicine usage, and antibiotic usage (Fig. 1A). Details of these phenotypes are listed in Table S1.

We employed the diaPASEF-based metaproteomic analysis to study the functional profiles of gut microbiota in a cohort of thousands of individuals. A total of 2,514 MS profiles containing 1,967 fecal samples and 547 quality controls were acquired. Our investigation encountered an initial hurdle in the analysis of mass spectral data. This challenge is commonplace in metaproteomic analysis, primarily due to the extensive protein sequence databases inherent to the human gut microbiome, necessitating significant computational resources. Additionally, the integration of ion mobility dimensions in diaPASEF technology introduces further complexity to the data, rendering its interpretation more intricate. To surmount these obstacles, we developed a customized analysis pipeline specifically designed for diaPASEF-based metaproteomics. This pipeline encompasses ddaPASEF-based spectral library generation, database searches tailored to diaPASEF MS profiles, and the annotation of taxonomic and functional information within metaproteomic datasets. For the spectral library generation, we employed a refined three-step

107 iterative database search method[37] alongside protein inference rules based on the maximum
108 parsimony[38] (Fig. 1B, Fig. S1, Fig. S2). The process commenced with partitioning the extensive
109 human gut protein sequence database into ten segments, each dedicated to the target-only
110 proteomic search of a ddaPASEF MS profile. The identified protein sequences obtained from the
111 target-only search were merged to create a "pseudo-metaproteome" database for each ddaPASEF
112 MS profile (Fig. 1B, Fig. S1). Subsequently, a project-specific database (GNHS.fas) was
113 constructed by matching each profile against its corresponding "pseudo-metaproteome," utilizing
114 a target-decoy approach. This database, comprising proteins identified with high confidence (FDR
115 ≤ 0.01) from all ddaPASEF analyses, served to pinpoint project-specific microbial proteins
116 accurately. The final phase involved searching all ddaPASEF profiles against this project-specific
117 database through standard target-decoy search strategies. Subsequently, an in-house protein
118 inference process based on the "maximum parsimony principle" was applied to eliminate the
119 redundant proteins (Figure S2 and methods), resulting in the construction of the definitive spectral
120 library.

121 Employing a three-step iterative search strategy enabled us to notably diminish the size of the
122 protein sequence database. Initially, database segmentation reduced the entry count from
123 10,352,085 to 1,035,209. Subsequent refinement, incorporating metaproteomic data from GNHS
124 population, further reduced this to just 164,695 entries, denoted as "GNHS.fas". This streamlined
125 process facilitated the construction the spectral library within a timeframe of 38 hours. The
126 subsequent protein characterization and quantification via diaPASEF MS-based analysis of all
127 samples required approximately 625 hours, averaging about 15 minutes per MS profile (Fig. 1B).
128 Finally, taxonomic and functional annotations were performed leveraging the clusters of
129 orthologous genes (COG) and KEGG ontology (KO) databases, augmenting our metaproteomic
130 data with significant biological insights.

131 **Data overview**

132 For all the 1,967 fecal samples, we measured 7,731 human-specific peptides belonging to
133 1,277 unique human proteins, 360,180 microbial-specific peptides belonging to 86,876 microbial
134 proteins (Fig. 1C). Of all the microbial proteins, 74,624 are unique microbial proteins and 12,252
135 are microbial protein groups comprising undistinguishable microbial proteins. Here, we refer to
136 the unique microbial proteins and the microbial protein groups as "metaproteins". For the
137 taxonomic characterization, we identified four superkingdoms, 34 phyla, 49 classes, 74 orders,
138 111 families, 235 genera, 373 species, and 146 strains. For the functional characterization, a total
139 of 2,537 COGs and 2,679 KO clusters were annotated (Fig. 1C). To assess the stability of human
140 protein detection, we subjected the fecal samples to multiple washes, ranging from three to five
141 times. Remarkably, a significant majority (86.3%) of these human proteins were consistently
142 detected after multiple washes, underscoring their robustness and stability in the human fecal
143 samples (Fig. S3 A-E). Further Gene Ontology (GO) analysis revealed an enrichment of these
144 proteins in peptidase activity, highlighting their pivotal role in food digestion processes (Fig. S3F).
145 Additionally, 96 human proteins have been shown to interact with gut microbes and are
146 potentially implicated in regulating human diseases [39] (Fig. S3G). Furthermore, to assess data
147 quality over extended acquisition periods, we obtained multiple pairs of technical and biological
148 replicate samples within and among batches. We observed high correlation and closely clustered
149 Bray-Curtis (BC) distance among the 547 QC replicates, which included 95 inter-batch biological

150 replicates, 98 inter-batch technical replicates, 169 intra-batch biological replicates, and 185
151 intra-batch technical replicates (Fig. S4A). The BC distance across these replicates showed no
152 significant batch-to-batch variances (Fig. S4B). Principal Component Analysis (PCA) further
153 validated our results, demonstrating a consistent distribution of samples with minimal batch
154 effects, thus affirming the integrity and reliability of the data (Fig. S4C).

155 In each sample, we identified an average of 477 ± 114 human proteins (mean ± SD),
156 9,761 ± 3,905 metaproteins, 103 ± 18 genera, 140 ± 29 species, 1,304 ± 268 KO, and 1,365 ± 241
157 COGs clusters. The median taxonomic annotation rate per sample was 86.0%, while the median
158 functional annotation rates per sample were 75.2% for KO and 93.9% for COG clusters (Fig. S5
159 A–D). A considerable metaproteome variation in richness was observed on the microbial function
160 and taxonomy levels (Fig. S5E), consistent with findings from a previous population-scale
161 metagenomic study[4]. Based on the metaproteomic data, we gained a fundamental understanding of
162 the taxonomy and functions of the gut microbiome across the GNHS population (Fig. 1 D–E, Fig.
163 S6). The dominant bacterial phyla were Firmicutes, Bacteroidetes, Proteobacteria, and
164 Actinobacteria (Fig. 1D, Fig. S6A), which aligns with the finding of a recent population-scale
165 metagenomic study [40]. Apart from Bacteria, we also identified six microbes from Archaea and
166 Eukaryote (Fungi), as well as seven Viruses (Fig. 1D). In terms of microbiome functions, the
167 predominant COG categories were "G-carbohydrate transport and metabolism", "J–Translation,
168 ribosomal structure and biogenesis", and "C–Energy production and conversion" (Fig. 1E, Fig.
169 S6F). On the other hand, host human proteins were primarily associated with "O–Posttranslational
170 modification, protein turnover, chaperones", "R–General function prediction only", and
171 "Z–Cytoskeleton" (Fig. 1F). This underscores distinct functional profiles between microbial and
172 human proteins within the human gut.

173 **Population-scale comparison on metaproteomics and metagenomics**

174 To emphasize the unique insights gleaned from metaproteomics, we conducted an in-depth
175 comparison between the metaproteomic data and the metagenomic data [41] of the GNHS cohort.
176 Multiple levels, including human protein, metaprotein, COG, KO cluster, and taxonomy were
177 employed to capture the microbial taxonomic and functional diversity within the GNHS cohort.
178 Analysis was conducted on non-redundant samples collected from participants who had not
179 received antibiotics (n = 1385) (Fig. 2A). Our findings showed that, as sample size increased,
180 metaproteomic data reached maximum estimated richness with 333 individuals at the genus level
181 and 302 individuals at the species level. For functional cluster KO, the metaproteome data
182 achieved 99% of the maximum estimated richness with approximately 200 samples and reached
183 the maximum estimated richness with 1,049 individuals (Fig. 2B, Fig. S7A). Additionally,
184 metaproteomic data covered over 99% of the estimated maximum richness at the human protein,
185 metaprotein, and COG cluster levels within 500 samples (Fig. S7A). In contrast, the metagenomic
186 data of the GNHS cohort could only reached 85% of the saturation level at the taxonomic genus
187 level, 81% at the species level, 93% at the KO functional cluster level, and 97% at the COG level
188 with 1,219 individuals (Fig. 2B, Fig. S7B). Previous population-based metagenomic studies have
189 shown that the observed richness is lower than the estimated richness even with sample sizes
190 exceeding 8,000 [40]. We also observed that the diversity of microbial taxa and their functions
191 predicted by metagenomics far exceeds those detected by metaproteomics. Therefore, we infer
192 that metagenomics more accurately represents the entire microbial composition and the spectrum

193 of its genetic functions. Conversely, metaproteomics is more adept at identifying actively living
194 microbes and their actively expressed functions. Despite abundant estimations in gut microbiome
195 composition diversity among individuals, the diversity of expressed functions appears to be
196 relatively conservative and less diverse. Consequently, metaproteomic studies can draw robust
197 conclusions with fewer samples.

198 Subsequently, we examined the correlation within and between metaproteomic and
199 metagenomic datasets, focusing on paired metaproteomic and metagenomic data from 1,041
200 individuals with non-redundant sample collection (Fig. 2A). Apart from the human protein level,
201 the BC distance data of microbial levels exhibited strong correlation with each other within each
202 omic data (> 60% correlation) (Fig. S7C). We also observed statistically significant correlation
203 between the two omic datasets. Notably, the correlation between the taxonomic datasets of the two
204 omics was higher than that of the functional datasets (Spearman rho: 0.34 for the genus, 0.43 for
205 species compared to 0.26 for COG, 0.25 for KO) (Fig. S7C), suggesting taxonomic data are more
206 consistently aligned across omics than functional data. This was further supported by the
207 distribution of abundant genera and KOs, where taxonomic correlation was typically higher. The
208 most abundant taxa in two omics datasets were found to be consistent (Fig. S8 A-B). Among the
209 top 20 genera identified in the metaproteomic data, 19 were also detected in the metagenomic data,
210 exhibiting a strong correlation ($r_{Spearman}$ = 0.74, p = 5e-4) in their relative abundance (Fig. 2C). A
211 similar pattern was observed at the species level ($r_{Spearman}$ = 0.72, p = 0.0052) (Fig. S8C). However,
212 at the KO and COG functional levels, the highly abundant functions are distinct between the two
213 omics datasets (Fig. S8 D-E). Although the highly abundant functions in the metaproteomic data
214 were also detected in the metagenomic data, the quantitative results showed no statistically
215 significant correlation between the two omics datasets (Fig. 2C, Fig. S8F). The observed
216 differences in functional profiles between the two omics data might arise from the inherent
217 measurement focuses: metagenomics quantifies gene copy numbers, whereas metaproteomics
218 assesses protein expression [42,43].

219 Population-based metagenomic studies reported conserved functional category variations
220 between individuals [4,40,44]. Given the quantitative functional discrepancies between
221 metaproteomics and metagenomics, we investigated whether metaproteomics could reflect the
222 functional variation within the population. Analyzing the feces of 1,041 subjects with both
223 metagenomic and metaproteomic data, we observed more pronounced variability in
224 metaproteomics KO and COG categories compared to that in metagenomics (Fig. 2D).
225 Additionally, our analysis revealed that variation in functional categories between baseline and
226 follow-up samples were higher in the metaproteome data than in metagenomic data across 377
227 individuals (Fig. 2D). Thus, we inferred that metaproteomics may offer a more sensitive approach
228 to detecting functional differences. Finally, we examined the long-term stability of the gut
229 microbiome within individuals. Both metaproteomics and metagenomics data showed that
230 individuals displayed stable gut microbiome compositions and functions for over 1.4–5.5 years
231 (Fig. S9 A–B), consistent with a previous metagenomic study [45]. It has been reported that the
232 stability of the microbiome over time inversely correlates with its biodiversity at the initial
233 assessment [45]. In our data, the metagenomic analysis showed a slight negative correlation ($r_{Spearman}$
234 = -0.1) between individual microbiome variation and the baseline taxonomic species richness. No
235 significant correlation could be obtained at metagenomic taxonomic genus, functional KO, or

236 COG levels (Fig. 2E, Fig. S9C). In contrast, metaproteomic analysis revealed significantly
237 negative correlations across all microbial taxonomic and functional levels. Notably, the correlation
238 was stronger at the microbial function level than the microbial taxonomy level ($r_{Spearman}$ = -0.23 for
239 genus vs. $r_{Spearman}$ = -0.48 for KO function), suggesting that functional diversity, as captured by
240 metaproteomics, provides a more reliable measure of microbiota stability over time.

241 In summary, our study highlights the superior capabilities of metaproteomics in revealing the
242 functional dynamics of the microbiota. Through the lens of metaproteomics, we were able to
243 capture a comprehensive view of microbial functional biodiversity within a population of fewer
244 than 500 individuals, identify significant functional variations, and gain deeper insights into the
245 temporal dynamics of microbiome functions.

246 **Core microbial functions and the core species expressing these functions**

247 Core features are believed to play crucial roles in maintaining the health of the gut ecosystem.
248 To identify core features within the GNHS cohort, we conducted a screening for attributes shared
249 across individuals. Features present in more than 80% of both inter-individual (non-redundant
250 samples collected from 1,041 individuals) and intra-individual samples (samples collected at the
251 baseline and follow-up time points from 377 individuals) were categorized as "core". In the
252 metaproteomic data, 262 human proteins, 353 metaproteins, 9 phylum, 56 genera, 62 species, 797
253 KOs, and 947 COGs emerged as core features (Fig. S10A, C–E; Table S2). The metagenomic data
254 revealed 4 phyla, 19 genera, 23 species, 3,227 KOs, and 2,773 COGs as core features (Fig. S10B,
255 F–H; Table S3).

256 Our primary focus is on the core microbial functionalities present within the population and
257 how these functions are distributed among the core microbes. To pinpoint core microbial
258 functionalities, we characterized pathways significantly enriched in core KOs as "core pathways",
259 using a criterion of a Benjamini & Hochberg (BH) adjusted p-value ≤ 0.05 and a fold enrichment
260 value ≥ 2. Additionally, we mapped the uniquely expressed metaproteins of each core species and
261 identified the pathways that were significantly expressed by them (with a BH adjusted p-value ≤
262 0.05 and fold enrichment ≥ 2). This dual analysis allows us to discern which microbial pathways
263 are likely to be expressed and the core microbes responsible for their expression. Our findings
264 revealed 42 core pathways enriched in metaproteomics and 64 in metagenomics, with 39 pathways
265 common to both datasets (Table S4). Notably, core pathways in metaproteomics showed higher
266 fold enrichment values than those in metagenomics (Fig. S11), indicating a greater likelihood of
267 these pathways being actively expressed. We classified pathways with greater fold enrichment in
268 metaproteomics compared to metagenomics as "highly expressed pathways" and pathways with
269 lower fold enrichment in metaproteomics as "moderately expressed pathways". We found that
270 highly expressed pathways mainly involve "housekeeping" functions essential for microbial
271 survival, such as translation, energy metabolism, carbohydrate metabolism, and amino acid
272 metabolism. These functional pathways, significantly enriched within the core microbes and
273 detected in at least 15 core species (Fig. S11, Table S5), exemplify functionalities commonly
274 expressed across the microbial community. This aligns with findings from population-based
275 metatranscriptomic studies, suggesting these housekeeping functions are readily transcribed [46].
276 Furthermore, the enriched pathways occurred in a limited number of core species, involving
277 functions like glycan degradation, streptomycin biosynthesis, drug metabolism related to

278 xenobiotic degradation, as well as taurine and hypotaurine metabolism. Despite being expressed
279 by only a few core microbes, these pathways still exhibited relatively high expression in the
280 population (Fig. S11, Table S5). In contrast, the moderately enriched core pathways within the
281 population are mainly associated with environmental stress response, such as cellular community,
282 cell motility, replication and repair, and membrane transport. These functions, enriched in fewer
283 than three core microbes, exhibit low expression levels across the population (Fig. S11, Table S5).
284 This distinction between highly and moderately expressed pathways provides insight into the
285 functional stability and adaptability of the core microbes within diverse environments.

286 The metabolism-related pathway constituted the largest proportion of highly expressed group
287 in gut microbe community, underscoring the microbial community's active metabolic function.
288 Short-chain fatty acids (SCFAs) and vitamin Bs act as an energy source for intestinal cells and
289 regulate the activity of the host's immune [47-49]. According to our data, a total of 33 microbes were
290 found to express functions related to the metabolism of SCFAs or vitamin B, with 23 of these
291 microbes previously reported to have the capability of secreting these beneficial metabolites (Fig.
292 3A, Table S6). We also found core microbes that contribute to the production of SCFA or vitamin
293 B which have not been reported. Notably, *Frisingicoccus caecimuris* demonstrated a significant
294 capability for propanoic acid production, evidenced by a high fold enrichment value of 42 in the
295 propanoate metabolism pathway. *Clostridium fessum* showed notable activity in butanoate
296 metabolism with a fold enrichment of 18, while *Blautia hansenii* was involved in vitamin B9
297 metabolism, presenting a fold enrichment value of 22 (Fig. 3A and Table S6). To verify our
298 findings, we measured the butyric acid content in feces from 494 individuals. Our analysis
299 revealed a positive correlation between the total relative abundance of 19 specific bacteria and the
300 level of butyrate in the fecal samples, suggesting the potential contributions of the microbiome
301 community to butyrate production (Fig. 3 B-C, $r_{Spearman} = 0.25$, $p = 2.7e^{-8}$).

302 In conclusion, we have delineated the highly expressed biological pathways and identified
303 the core species responsible for these expressions. We also identified potential core microbes
304 capable of producing SCFAs or Vitamin Bs, emphasize the crucial role of metaproteomic analysis
305 in accurately identifying microbial contributors to key metabolic processes.

306 **Associations between phenotypes and metaproteomic data**

307 To explore the relationship between phenotypic variation and the diversity of gut microbiome
308 among individuals, we performed permutational multivariate analysis of variance (PERMANOVA)
309 to assess the association between 49 phenotypes (excluding antibiotic use) and BC dissimilarities
310 at various levels, including human protein, microbial metaprotein, KO, COG, genus, and species
311 levels. Five metadata parameters, including type 2 diabetes medicine (T2D medicine), blood
312 triglyceride (TG) level, Bristol scale, T2D, and tea consumption, were found to exhibit significant
313 association with BC distance variation at all levels (Fig. S12, adjusted p-value ≤ 0.1, Table S7).
314 Among these metadata, the most significant contributor was T2D medicine, followed by TG level,
315 Bristol stool scale, T2D, and tea consumption. These results align with findings from
316 population-based metagenomic studies conducted in various countries and regions, reinforcing the
317 impact of these phenotypes on the gut microbiome's composition and function [4,40,50-52]. We next
318 explored the specific associations between phenotypes and the metaproteomic data using general
319 linear model (GLM) method. A total of 10,714 associations were identified between

metaproteomic data and 39 phenotypes (Table S8). Still, T2D medicine, Bristol scale, T2D, TG, and tea consumption emerged as the primary phenotypes correlated with metaproteomic data across various levels, including the levels of protein (metaprotein and human protein), microbiota function (COG and KO), and taxonomy (genus and species) (BH adjusted $p \leq 0.05$, Fig. 4A).

Some of the associations at the taxon level have been previously reported. For instance, the genus *Anaerotruncus* was positively correlated with constipation. *Megasphaera* showed a positive correlation with T2D [53]. *Evtepia* and *Romboustsia* [54] exhibited a negative correlation with T2D. Tea drinkers displayed increased *Gemmiger* [55] and decreased *Tyzzerella* levels [56]. In addition, at the function levels, we uncovered protein functions of microbes and the host that are significantly associated with these metadata for the first time. At the microbial KO level, enzymes related to carbohydrate metabolism and energy metabolism were significantly increased in diarrhea and significantly decreased in constipation. Amino acid metabolism functions were upregulated in constipated individuals. Diabetic patients showed decreased expression of carbohydrate metabolism and energy metabolism proteins, while the levels of rubrerythrin, electron transfer flavoprotein-ubiquinone oxidoreductase, and lactate dehydrogenase were elevated. At the human protein level, mucin proteins MUC6 and MUC17, proteasome PSA1 and PSB1, as well as succinate dehydrogenase enzymes SDHA and SDHB, were increased in diarrhea individuals. In individuals with constipation, the abundance of mucin proteins as well as carbohydrate degradation related trehalase TREA and alpha-amylase-associated protein SLC3 were decreased. (Fig. 4D).

Aging is the root cause of numerous diseases and functional disorders. The connection between aging and the gut microbiome has become increasingly emphasized [57,58]. Beyond the five phenotypes previously discussed, our data also revealed age-related gut microbes and their functions. At the microbial functional level, there were 58 microbial KOs associated with aging (Table S7). Among these, glycine dehydrogenase, beta-glucuronidase, and proline racemase exhibited the strongest negative correlation with aging, indicating a potential decline in the overall metabolic activity of the gut microbiome with increasing age (Fig. S13A). Conversely, outer membrane proteins of Gram-negative bacteria, such as ompF, ompC, and lamB, showed the strongest positive correlation with aging. Outer membrane proteins play a crucial role in pathogens, aiding bacteria in evading host defense mechanisms, thereby increasing their virulence [59]. Therefore, the increase in outer membrane proteins might suggest that with age, the activity or potential pathogenicity of certain pathogenic strains in the gut microbiome could increase, potentially impacting the host's health. Taxonomically, four genera were associated with aging, including negative associated genus *Phascolarctobacterium* and positive associated genera *Klebsiella*, *Dialister*, and *Veillonella* (Fig. S13B). By connecting these microbial genera with their unique functional expressions, we observed specific changes in the functionality of each microbe. Functions related to energy metabolism, lipid metabolism, and nucleotide metabolism of *Dialister* showed positive correlation with aging. Similarly, energy metabolism functions of *Veillonella* were also positively associated with aging. Additionally, the outer membrane porins primarily expressed by *Klebsiella* were positively correlated with aging, indicating an increased pathogenicity as age advanced (Fig. S13C).

The correlation between metaproteomic data and phenotypes deepen our understanding of the changes in the gut microbiome and their functions under different phenotypic conditions, thus

363 offering new perspectives in disease prevention and health management.

364 **T2D related microbes and functions identified through metaproteomics**

365 T2D has been proven to be related to gut microbiota composition, but the specific functions
366 of microbiota associated with T2D remain incompletely understood. In this study, we analyzed
367 non-redundant samples from 1153 non-T2D and 223 T2D individuals within the GNHS cohort to
368 investigate T2D-associated metaproteomic features (Fig. 5A). A total of 1,381 features including
369 785 metaproteins, 23 human proteins, 268 KOs, 279 COGs, 14 genera, and 12 species were
370 identified significantly associated with T2D (BH adjusted $p \leq 0.05$, Table S8). Among these, 335
371 metaproteins were uniquely expressed by 36 microbial species (Table S9). Consequently, a total
372 of 40 species were identified to be correlated with T2D when considering the combined results of
373 species related to T2D and those corresponding to T2D-related metaproteins. These findings are
374 consistent with previous reports, where 25 species were identified as related to T2D using 16S or
375 metagenomic sequencing (Table S9). Our data further elucidate the significantly expressed
376 metaproteins of these bacteria, with over half of the species-uniquely-expressed metaproteins (176
377 out of 335) involved in carbohydrate, amino acid, lipid, and nucleotide metabolism, as well as
378 energy production. This highlights the critical regulatory roles of these functions in T2D [51,60-62]
379 (Fig. S14). Specifically, the 40 species mainly form 4 distinct microbial community clusters based
380 on their interrelationships (Fig. S14, Fig. S15A). Cluster 1 included 6 species from the
381 *Bifidobacterium* genus, with their abundance or protein expressions positively correlated with
382 T2D. Specifically, these microbial proteins were mainly associated with amino acid transport and
383 carbohydrate metabolism. Cluster 2 including 14 species, with 11 of them belonging to
384 Eubacteriales. These bacteria mainly expressed carbohydrate metabolism related functions that
385 were negatively correlated with T2D. The Cluster 3 was composed of *D. hominis* and *A. histamini*,
386 both belonging to the family Veillonellaceae. They were positively correlated with T2D. Their
387 proteins mainly participated in energy metabolism and amino acid metabolism. The last cluster
388 consisted of two species belonging to the *Megasphaera* genus, positively correlated with T2D.
389 The microbial proteins expressed in this cluster were mainly related to energy metabolism and
390 lipid metabolism (Fig. S15A). To validate the T2D-related features identified in the GNHS cohort,
391 we included an independent cohort, referred to as the FH cohort, consisting of 82 non-T2D
392 individuals and 21 T2D individuals. Through this validation process, we confirmed the association
393 of 45 metaproteins, 1 human protein, 27 KOs, 18 COGs, and 3 species associated with T2D
394 (consistent orientation of associations and BH adjusted $p \leq 0.05$, Table S10). Notably,
395 *Megasphaera elsdenii* and 20 of its expressed metaproteins, predominantly involved in energy
396 production and lipid metabolism, were confirmed to be associated with T2D in the FH cohort as
397 well. These consistent associations suggested a potential significant role for *M. elsdenii* in the
398 development of the disease (Fig. S15B).

399 Next, we explored the performance of metaproteomic data in classifying T2D patients from
400 those without disease. To achieve this, we utilized the T2D-associated features identified through
401 GLM analysis in the GNHS cohort as potential features (Fig. 5A). These features were further
402 refined using the eXtreme Gradient Boosting (XGBoost) algorithm with wrapped recursive feature
403 elimination. Using this approach, we identified 16 metaproteins, 21 human proteins, 11 KOs, 11
404 COGs, 10 species, and 14 genera as the most robust features for machine-learning models at each
405 level (Table S11, Fig. 5A, and Methods). The performance of these features was evaluated by an

internal validation and the independent FH test cohort at each level. Notably, the microbial functional COG and KO levels exhibited strong performance in both the internal and test cohorts (Fig. 5B). The best receiver operating characteristic (ROC) area under the curves (AUCs) reached 0.738 and 0.723 at COG and KO levels in FH cohort, respectively (Fig. 5B). We explored the feasibility of using metagenomic data for T2D classification using the same approach (Fig. S16B). At the species level, the best AUC of the machine learning model achieved 0.62 in the independent validation set, comparable with that of metaproteomics. However, when evaluating microbial functional KO levels, metagenomics only achieved an AUC of 0.5 much lower than that of metaproteomics (Fig. S16C). These findings highlight the advantage of quantitative characterization of microbial expressed functions by metaproteomics in T2D classification. (Table S12, S13)

Combining the outcomes of GLM analysis and machine learning, we carried out an extensive screening to identify the most conclusive features associated with T2D. These features exhibited significant correlation with T2D in both the GNHS and FH cohorts. We identified five pivotal features, consisting of two microbial functions—electron transfer flavoprotein-quinone oxidoreductase fixC (K00313) and glycogen phosphorylase (K00688), two bacterial genera—*Clostridium fessum* and *M. elsdenii*, and one human protein—phospholipase B1 (PLB1) (Fig. 5C). *M. elsdenii,* K00313, and K00688 were elevated in T2D patients, and their relationships with diabetes have been reported [63-65]. While PLB1 and *Clostridium fessum* were first found to be decreased in T2D patients. A co-occurrence network was developed to illustrate the regulatory roles of machine learning and FH cohort validated T2D-related features (Fig. 5D). Notably, *M. elsdenii* emerged as a hub microbe with the highest connectivity in the network. The network specifically highlights up-regulated proteins in the *M. elsdenii,* primarily concentrated in butyrate production through lactate utilization and redox pathways. This emphasizes the critical role of *M. elsdenii* and its functions in T2D. (Fig. 5D).

***M. elsdenii* safeguards against diabetes by producing butyric acid**

Next, we directed our analysis towards *M. elsdenii* to explore its potential role in the development of T2D. Supported by prior research [66,67], our findings showed that *M. elsdenii* utilizes lactate as the energy source to produce butyric acid. Given that butyric acid has been recognized to alleviate diabetes symptoms [68,69], the increased abundance of *M. elsdenii* in T2D patients might initially appear paradoxical. However, our investigation also uncovered significant positive correlations between *M. elsdenii* and its proteins with antidiabetic drug (Fig. S18; Table S9), a notable observation considering about 57% of diabetic patients were under medication treatment. This led us to speculate that the observed positive correlation between T2D and *M. elsdenii* could primarily stem from the medications commonly prescribed to these patients. Analyzing data from three distinct groups—non-diabetic individuals, untreated diabetics, and diabetics on medication, we noted an increasing trend in *M. elsdenii* levels among patients undergoing antidiabetic treatment (Fig. 6A). Corroborating our hypothesis, previous studies have documented elevated levels of *M. elsdenii* in patients receiving antidiabetic treatment, including metformin and acarbose [70,71]. To confirm the direct effect of antidiabetic medication on microbial growth, we conducted *in vitro* experiments by introducing common anti-diabetic drugs, namely metformin, acarbose, and glimepiride, to a pure culture of *M. elsdenii.* The outcomes, measured through both the OD600 values and bacterial cell counts, verified that acarbose and metformin

449 promoted the growth of *M. elsdenii*, with the exception of glimepiride (Fig. 6B).

450 The increase of *M. elsdenii* through drug treatment might suggest its glucose-lowering effect
451 on T2D. To explore this possibility, we conducted an experiment with gnotobiotic mice on a
452 high-fat diet (HFD), treating them with *M. elsdenii* and a control bacterium, *Lactobacillus reuteri*,
453 known for its glucose-lowering effects in mice [72,73]. The mice were subjected to an HFD for 42
454 days, with the first 14 days including bacterial colonization. We monitored their fasting blood
455 glucose (FBG) levels throughout the study (Fig. 6C). Compared to the control group on a chow
456 diet (CD), mice on the HFD exhibited significantly higher FBG levels. However, the group treated
457 with *L. reuteri* showed a consistent reduction in FBG levels compared to those on the HFD after
458 26 days (Fig. 6D). Similarly, mice fed with *M. elsdenii* showed a notable and consistent decrease
459 in FBG levels stating from day 16, as compared to HFD-fed mice (Fig. 6D), demonstrating its
460 glucose-lowering impact on T2D. Our metaproteomic data revealed an increase in proteins related
461 to butyric acid production in *M. elsdenii*. To investigate the mechanism underlying blood glucose
462 levels induced by *M. elsdenii*, we examined the intestinal contents of mice after 42 days to assess
463 the level of butyric acid. We found that HFD mice fed with *M. elsdenii* exhibited significantly
464 higher butyrate level than the other groups (Fig. 6E), suggesting a possible mechanism that *M.*
465 *elsdenii* contributes to glucose regulation in T2D.

466 Therefore, drawing from the metaproteomic data, *in vitro* culture, and mouse experiments,
467 we found that *M. elsdenii* responds to the stimulation induced by common antidiabetic drugs like
468 acarbose and metformin. This stimulation enhanced the production of butyric acid, which
469 subsequently reduces high blood glucose levels and alleviates diabetes conditions (Fig. 6F).

470 **Discussion**

471 In this project, we have developed a metaproteomics analysis pipeline tailored for large-scale
472 population samples. This pipeline demonstrated high-throughput capability and reproducibility.
473 We applied this method in a regional cohort from China, which included over a thousand
474 individuals. Notably, this represents the inaugural application of metaproteomics on such a sizable
475 cohort. The critical bottleneck in metaproteomics analysis lies in aligning microbial proteins with
476 mass spectrometry spectra. Given the limited understanding of the full spectrum of gut
477 microorganisms, there is a lack of a comprehensive database to underpin this process. To
478 overcome this obstacle, numerous studies have utilized paired metagenomes to construct protein
479 databases. This approach facilitates the analysis of metaproteomics data and aids in identifying
480 microbial proteins most pertinent to the metagenomic context. For example, open-access resources
481 like the Integrated Gene Catalog (IGC) [74], derived from integrated metagenomic sequencing data,
482 and the Unified Human Gastrointestinal Protein Database (UHGP) [75], built on sequenced
483 microbial genomes, are increasingly being integrated into metaproteomic analyses. However,
484 relying on metagenomic sequencing to generate a reference database adds extra burden to
485 metaproteomic analysis. Here, we utilized an open-source database instead of paired metagenomes
486 to align the microbial protein sequences with mass spectrometry data. Additionally, the IGC
487 database, derived from metagenomic sequences of 1,070 individuals, notably includes a
488 significant representation from China. This makes it particularly suitable for exploring Chinese
489 cohorts. As a result, we have chosen the IGC database as the foundation of our reference database.

490 Through metaproteomic analysis, we identified 1,277 unique human proteins and 86,876

microbial proteins from 1,967 fecal samples of 1,399 individuals. One of the key advantages of metaproteomics over metagenomics is its capability to detect host proteins. However, accurately determining the origin of these host proteins, whether they arise from shed epithelial cells, host-secreted proteins, or genuine interactions with microbes, is crucial as it directly impacts the scientific significance of our protein detection findings. To address this, we initially employed a differential centrifugation method to remove shed epithelial cells and food residues from the fecal samples before enriching the microbial fraction for metaproteomic analysis. This step effectively eliminated the majority of host proteins derived from shed epithelial cells. Furthermore, we performed multiple rounds of differential centrifugation washing to ensure the comprehensive removal of impurities while verifying the presence of proteins. Across the 3 to 5 washing stages, we consistently detected these host proteins, indicating their stable association with microbial precipitates. Functional annotation analysis of the identified host proteins revealed their predominant involvement in food digestion enzymes, with some confirmed to interact with microbial proteins [39]. Consequently, we believe that the identification of these human proteins provides valuable insights for future investigation into the complex interplay between the host and gut microbiota.

Metagenomics and metaproteomics are vital techniques for exploring microbial species and their proportions through gene or protein analysis. While comparisons between these methods have shown some consistency in classifying high-abundance microbial populations [20,75], there is a lack of research on the advantages of metaproteomics over metagenomics in population studies and disease applications, primarily due to limited sample sizes. In our study, leveraging data from thousands of samples, we found that metaproteomic data tends to be more conservative compared to metagenomic data in discerning population diversity. Metaproteomics can capture the diversity of microbial classification within a population with around 350 samples and detect all microbial functional clusters expressed by the microbes with approximately 1000 samples. Our findings, based on a population of middle-aged and elderly individuals in China, suggest that metaproteomics offers advantages in encompassing the diversity of microbial populations in human populations, particularly in capturing active microbial functions in the gut. Considerations must be given to the limitations of metaproteomics, including the sensitivity of mass spectrometry, which can impact protein detection depth, and the prevalence of highly abundant microbial proteins in the results. While the conclusions regarding high-abundance microbial compositions in metagenomics and metaproteomics are generally congruent, the functional expression of microbial populations exhibits little correlation between the two omics approaches. Across populations or in samples collected repeatedly from the same individual over years, microbial protein expression in metaproteomics displays greater variability compared to gene content detected in metagenomics. Large-scale metagenomic studies have shown that while there is substantial variability in microbial composition among populations, the abundance of gene functions remains relatively consistent [40,44]. The increased functional variability in metaproteomics relative to metagenomics may stem from differences between gene copy number and protein expression. Our study also highlights the benefits of metaproteomics for longitudinal studies of microbial populations within individuals over several years. The correlation between the biodiversity and the stability of microbial composition and function within an individual over time is stronger in metaproteomic data compared to metagenomics, making it a valuable tool for investigating the dynamics of microbial communities in human cohorts. Overall, while metagenomics and metaproteomics

535 provide complementary information, metaproteomics offers unique insights into microbial
536 diversity, functional expression, and longitudinal stability within human populations. The
537 combination of these omics' approaches can enhance our understanding of the complex
538 interactions between host and microbiota in health and disease contexts.

539 In the analysis of core microbial features in the GNHS cohort using metaproteomic and
540 metagenomic data, we observed differences of the core microbial taxa at the genus level and core
541 microbial functional clusters at the KO level between two omics data. Metaproteomics revealed a
542 greater number of core microbial taxa at the genus level (56 in metaproteomics vs 19 in
543 metagenomics) compared to metagenomics, while metagenomics identified fewer core microbial
544 functional clusters at the KO level (797 in metaproteomics vs 3,227 in metagenomics). The core
545 genera identified in the metaproteome mainly encompassed those found in the metagenome, and
546 similarly, the core functions identified in the metagenome largely encompassed those found in the
547 metaproteome. The discrepancy of the core genus between two omics may be attributed to
548 differences in the annotation databases for microbial classification. Utilizing the open-source IGC
549 protein sequence database for metaproteomics could lead to variations in taxonomic annotation
550 compared to paired metagenomic data. However, despite these differences, abundant core genera
551 such as *Prevotella*, *Faecalibacterium*, and *Bacteroides* remained consistent between the two
552 omics data. These genera have been previously identified as core genus in worldwide population
553 studies based on 16S rRNA gene sequencing and metagenomics [5,76], further emphasizing their
554 significance in the gut ecosystem. When comparing microbial functions, the disparities between
555 the core microbial functions identified by the two omics approaches may arise from whether genes
556 are actively expressed as proteins. Beyond transcription and ribosomal proteins, the
557 core-expressed microbial functional groups predominantly encompass metabolic processes,
558 including sugar metabolism, amino acid metabolism, lipid metabolism, and energy production.
559 These findings underscore the significance of intestinal microbes in modulating host metabolic
560 pathways for overall environmental health. Enriching these core microbial functions into pathways
561 revealed that widely expressed microbial functions in the population are mainly related to
562 microbial survival, with significant functional redundancy among microbes. These findings are
563 consistent with the results from a metatranscriptomics study [46]. However, within the core
564 functional pathways, certain functions such as glycan degradation, streptomycin biosynthesis,
565 xenobiotic degradation, as well as taurine and hypotaurine metabolism, are expressed by only a
566 limited number of core microbes. This suggests that these microbes and their functions may play a
567 central and potentially irreplaceable role in the microbial community, indicating their importance
568 in maintaining microbial community stability.

569 We delved deeper into the distinct microbial proteins expressed by each microbe, delineating
570 the independent high-expression functions of each organism. This involved functional annotation
571 and pathway enrichment analysis of these metabolic proteins, with a particular focus on the vital
572 roles of SCFAs and vitamin Bs in microbial contributions to host health. We pinpointed 33 core
573 species potentially involved in the metabolism of SCFAs and vitamin Bs. Remarkably, 23 of these
574 species have been previously documented to harbor these metabolic capabilities, underscoring the
575 utility of metaproteomic data in elucidating microbial contributions to industrial production.
576 However, it's imperative to acknowledge the constraints and limitations inherent in such analyses.
577 Given the phenomenon of cross-feeding in the gut microbiota, where various microbes collaborate

578 to fulfill common metabolic pathways [77], pinpointing specific microbes solely based on pathway
579 enrichment merely indicates their potential involvement in SCFA and vitamin B metabolism.
580 Further research is imperative to determine whether these microbes can autonomously undertake
581 SCFA or vitamin B metabolism, or if they rely on assistance from other microbes.

582 We also investigated the associations between metaproteomic data and 49 phenotypes. We
583 found that T2D medicine, TG level, Bristol scale, T2D, and tea consumption exhibit the strongest
584 associations with microbial features. This finding is consistent with the observations from other
585 population-based metagenomic studies [50,78,79]. We identified the microbes and their functions most
586 closely associated with the five phenotypes and aging. Noteworthy correlations include
587 *Anaerotruncus* with constipation, *Megasphaera* with T2D, *Klebsiella* with aging, and specific
588 changes in microbial functions associated with these phenotypes.

589 Finally, our study centered on examining the microbiota and their functions associated with
590 T2D. While previous metagenomic research has identified various microbiota linked to T2D [80-83],
591 this study aimed to uncover the specific functional contributions of these microbiota to T2D. Our
592 findings revealed that *Bifidobacterium* may influence T2D through amino acid transport and
593 carbohydrate metabolism, while *Megasphaera* could contribute to T2D via energy and lipid
594 metabolism. By employing a combination of GLM and XGBoost machine learning algorithms, we
595 pinpointed five stable microbiota features consistently associated with T2D in both the GNHS and
596 independent FH cohorts. Remarkably, we noted a noteworthy correlation between *M. elsdenii* and
597 T2D, along with its association with medication treatment, consistent with prior research [70,84,85].
598 However, the exact mechanisms underlying how *M. elsdenii* influences the development of T2D
599 remain unclear and necessitate further investigation. Based on our data, we hypothesize that *M.*
600 *elsdenii* may contribute to T2D development through the production of butyric acid. To validate
601 this hypothesis, we conducted experiments on germ-free mice fed a high-fat diet. Our results
602 confirmed that administering *M. elsdenii* prevented abnormal elevation of blood glucose levels in
603 these mice. Furthermore, we observed an increase in butyric acid content in the intestines of mice
604 treated with *M. elsdenii*. In conclusion, our results suggest that medication may elevate the levels
605 of *M. elsdenii*, which in turn helps mitigate intestinal inflammation by producing short-chain fatty
606 acids, thereby potentially addressing elevated blood glucose levels in individuals with T2D.

607 Overall, metaproteomics serves as a valuable tool for uncovering the roles of microorganisms
608 in human health and their patterns of functional expression. Our data serve as a significant
609 resource for investigating microbial functions, thereby enhancing our understanding of how
610 microbes influence human health.

611 **Conclusion**

612 We generated high-quality metaproteomic data from 1,967 fecal samples collected from a
613 Chinese population. Our analysis delved into inter-individual variations of human proteins,
614 microbial taxa, and gut microbiota functions. The findings underscore the effectiveness of
615 population-based metaproteomics in representing actively functional biodiversity, elucidating
616 significant variations in functional clusters, and offering a more reliable indicator of the temporal
617 functional stability of the microbiome. We summarized highly expressed functions of each core
618 microbe and pinpointed core microbes capable of expressing elevated levels of SCFA and vitamin
619 B production. In addition, we investigated the correlation between 49 phenotypes and

metaproteomic data. We outlined the microbial functional characteristics linked to T2D and highlighted how *M. elsdenii* alleviates T2D by generating butyric acid through lactate utilization. Our study offers valuable resources for investigating the functions of the gut microbiota.

Acknowledgments

This study and the coauthors of this study are jointly supported by grants from the National Key R&D Program of China (No. 2021YFA1301600 and 2020YFE0202200), the National Natural Science Foundation of China (82073529, 81903316), and the Westlake Educational Foundation. We thank Westlake University Supercomputer Center for assistance in data analysis and storage, Dr Shu Zhu and his team for their advices.

Author contributions

T. G., J.Z., and Y.C. designed and supervised the project. B.L. Z.M., E.C., M.Y., M.S., W.G., and K.Z. collected the samples and metadata. Shuang L., Y.S., E.C., M.Y., Y.X., Sainan L., Y.W., H.G., K.Z., J.Y., X.C., and X.W. generated the data. Shuang L., Y.S., Z.M., Z.X., P.L., M.S., W.G., and W.J. analyzed the data. Shuang L., Y.S., and T.G. drafted the manuscript with inputs from all co-authors. Shuang L., Y.S., Z.M., and B.L. contribute equally to this work.

Competing interests

T.G. is a shareholder of Westlake Omics Inc. P.L., and J.Y. are currently employees of Westlake Omics Inc. The remaining authors declare no competing interests.

Materials and Methods

The GNHS cohort. Our study population included 1,967 stool samples from 1,399 participants from the GNHS cohort [86] of middle-aged and elderly people (40–75 years old) from the Guangzhou city urban area. Stool samples collected on-site at the Sun Yat-sen University were manually stirred, stored on ice, and transferred to -80 °C within 4 hours. Additionally, 25 physiologically measured or questionnaire-based phenotypic factors[87] were collected along with samples. Specifically, metadata variables excepted antibiotic usage were classified into five phenotypic categories, containing demographics, lifestyle, anthropometrics, biochemical measures, and metabolic disorders. Also, the metadata types can be continuous (*e.g.*, age, body mass index, and blood lipid levels), binary (*e.g.,* sex, disease status), or categorical (*e.g.*, Bristol scale scores). More details, including the definition of metabolic syndrome, type 2 diabetes, and hypertension, as well as the statistics of each phenotype, are provided in Table S1.

The FH cohort. Additional 104 individuals (52–83 years old) were recruited in Guangdong Province, China[54]. From these participants, we collected stool samples and their corresponding metadata as the GNHS cohort does. Detailed metadata information is provided in Table S1.

Batch design. To minimize the batch effects, the 1,967 samples were randomly assigned to 96 batches by individuals' age, gender, and sampling stage (baseline or follow-up). We designed intra-batch and inter-batch biological and technical replicates to monitor reproducibility during sample preparation and mass spectrometry (MS) acquisition. In each batch, two fecal samples were weighed twice and processed as intra-batch biological replicates. The inter-batch biological replicate samples mixed from randomly selected fecal samples were also designed in each batch before sample preparation. Also, after sample processing, peptides of two randomly selected fecal samples were injected twice for MS analysis in one batch as intra-batch technical replicates. The inter-batch technical replicates mixed from randomly selected peptides were also included in each batch. A batch contained 21 samples, two intra-batch biological replicates, two intra-batch technical replicates, one inter-batch biological replicate, and one inter-batch technical replicate. Altogether, we finally allocated 1,967 samples, 98 inter-batch technical replicates, 95 inter-batch biological replicates, 169 intra-batch biological replicates, and 185 intra-batch technical replicates to 96 batches (2,514 MS runs in total).

Sample processing. For each sample, about 200 mg stools were gently resuspended in 500 μL cold phosphate buffer (PBS) and then centrifuged at 500 × *g* and 4°C for 5 min to collect the supernatant. This process was repeated twice to fully collect gut microbes and discard epithelial cells and food residues. Supernatants were combined (about 1.5 mL) and centrifuged at 500 × *g* and 4 °C for 10 min to discard more debris. Finally, supernatants (700 μL) were centrifuged at 18,000 × *g* and 4°C for 20 min to collect microbe pellets. Proteins were then extracted using 250 μL lysis buffer (4% w/v SDS and cOmplete Tablets (Roche) in 50 mM Tris-HCl, pH = 8.0) from the pellets. Microbe pellets added with lysis buffer were boiled at 95 °C for 10 min. After cooling at room temperature, samples were ultrasonicated at 40 Khz (SCIENTZ) for 1 h on ice. Following extraction, the lysis buffer containing sample proteins was centrifuged at 18,000 × *g* for 5 min for debris discarding. The resulting 200 μL supernatants were transferred into a new tube and added with a 5-fold volume of pre-cooled acetone. After vigorous shaking, proteins were precipitated overnight at -20 °C. Proteins were then pelleted by centrifuging at 1,000 × *g* for 3 min

681 and washed with acetone three times at room temperature. The purified protein pellets were then
682 completely re-dissolved in a dissolving buffer (8 mM urea and 100 mM ammonium bicarbonate).
683 BCA protein assay kits were used to detect protein concentrations[55]. To prepare for the protein
684 digestion, protein (50 μg) from each sample was reduced using 10 mM tris (2-carboxyethyl)
685 phosphine (TCEP, Adamas-beta) and then alkylated with 40 mM iodoacetamide (IAA,
686 Sigma-Aldrich). The first round of digestion was performed by adding 0.5 μg trypsin (Hualishi
687 Tech) to the protein mixture, which was incubated for 4 h at 32 °C. Then, another 0.5 μg trypsin
688 was added to the protein mixture, which was then incubated for 16 h at 32 °C. The tryptic peptides
689 were desalted using solid-phase extraction plates (ThermoFisher Scientific, SOLAμ™) and then
690 freeze-dried for storage. Dried peptides were resuspended in 2% acetonitrile, with 98% water and
691 0.1% formic acid (FA), before MS acquisition.

692 ***Verification experiment of human proteins.*** Nine inter-batch biological replicate samples were
693 used for the verification experiment. Firstly, the samples were divided into three groups (group 1,
694 group 2, and group 3), each containing three samples as technical replicates. All samples were
695 gently resuspended in 500 μL cold PBS and then centrifuged at 500 × *g* and 4°C for 5 min three
696 times (group 1), four times (group 2), or five times (group 3) to collect the supernatant. The
697 combined supernatants (about 1.5 mL, 2.0 mL, and 2.5 mL for group 1, group 2, and group 3)
698 were centrifuged at 500 × *g* and 4 °C for 10 min to discard more debris (Fig. S3A). Next, 700 μL
699 supernatants were centrifuged at 18,000 × *g* and 4 °C for 20 min to collect microbe pellets (Fig.
700 S3A). The supernatant after centrifuging was collected into new tubes to detect the proteins that
701 remained in the supernatant. Then, stool residues after microbe enrichment in group 1, group 2,
702 and group 3 were further washed with 500 μL cold PBS and centrifuged at 500 × *g* and 4 °C for 5
703 min to collect the supernatant after microbe enrichment. The following procedures of sample
704 preparation of microbe pellets were followed as what has been described above. The supernatants
705 were directly added with a 5-fold volume of pre-cooled acetone to precipitate proteins. Then the
706 proteins were digested and desalted as described above.

707 ***High-pH reversed-phase fractionation.*** To generate a comprehensive spectral library of stool gut
708 microbial proteins, randomly mixed peptides from 672 samples in GNHS cohort together with
709 peptides from 11 healthy Chinese adults' stool (five males and six females, their ages ranging
710 from 22 to 33 years) were mixed into a library building pool (Fig. S1A). The pooled sample was
711 then fractionated using high-pH reversed-phase liquid chromatography (LC). About 300 μg of
712 purified peptides from the pool were separated using a nanoflow DIONEX Ultimate 3000 RSLC
713 nano System (ThermoFisher Scientific) with an XBridge Peptide BEH C18 column (300 Å, 5
714 μm× 4.6 mm × 250 mm) (Waters) at 45 °C. The mobile phase of buffer A was water with 0.6%
715 ammonia and pH = 10, while buffer B contained 98% acetonitrile and 0.6% ammonia, again with
716 pH = 10. A 120 min gradient from 5% to 35% buffer B with a flow rate of 1 mL/min was applied.
717 As a result, 120 fractions were collected at 1 min intervals and then combined into 30 fractions.
718 Next, the fractions were freeze-dried and re-constituted in 2% acetonitrile, with 98% water and
719 0.1% FA.

720 ***ddaPASEF mass spectrometry acquisition for library generation.*** The fractionated peptides were
721 spiked with iRT (Biognosys)[88] and loaded on a precolumn (5 μm, 100 Å, 5 mm × 300 μm I.D.)
722 and then separated by a nanoElute UHPLC System (Bruker Daltonics) equipped with a 150 mm ×
723 75 μm I.D. fused silica column custom-packed with 1.9 μm 120 Å C18 aqua (homemade) using a

724 95 min gradient at 300 nL/min. Buffer A contained 0.1% FA in water, while buffer B contained
725 0.1% FA in acetonitrile. The gradient flow was first set to 2% buffer B, linearly increased to 22%
726 buffer B during the next 80 minutes, then linearly increased to 35% buffer B for additional 10
727 minutes, and finally increased to 80% buffer B during the next 2 minutes and kept at this level for
728 another 3 minutes. A time-of-flight instrument coupled to a dual Trapped Ion Mobility
729 Spectrometry (TIMS) analyzer (Bruker timsTOF Pro) equipped with a CaptiveSpray
730 nano-electrospray ion source was used for MS acquisition. The TIMS section was operated with a
731 100 ms ramp time and a scan range of 0.6–1.6 V·s/cm^2. One cycle was composed of 1 MS scan
732 followed by 10 PASEF tandem-MS (MS/MS) scans. The MS and MS/MS spectra were recorded
733 for mass-to-charge ratios (m/z) ranging between 100 and 1700. A polygon filter was used to filter
734 out singly charged ions. For all experiments, the quadrupole isolation width was set to 2 Th for
735 m/z < 700 and 3 Th for m/z > 800; between these values, the isolation width was linearly
736 interpolated. The collision energy was ramped stepwise as a function of the increasing ion
737 mobility: 20 eV energy start for the 1/K0 start (0.6 V·s/cm^2) and 59 eV for the 1/K0 end (1.60
738 V·s/cm^2). Eighty-seven MS profiles were acquired from fractionated peptides with repeated
739 acquisitions. We also acquired one MS profile for the peptides of the library building pool. As a
740 result, 88 LC-MS profiles were obtained for the spectral library building.

741 ***diaPASEF mass spectrometry acquisition.*** A tandem system comprising a nanoElute UHPLC
742 System (Bruker Daltonics) and a hybrid TIMS quadrupole time-of-flight mass spectrometer
743 (Bruker timsTOF Pro) was used for MS measurements. For the LC analysis, 300 ng peptides were
744 trapped at 500 Bar on the precolumn (5 µm, 100 Å, 5 mm × 300 µm I.D.) in 0.1 % FA/water and
745 then separated along a 65 min LC gradient at a flow rate of 300 nL/min with a 150 mm × 75 µm
746 ID fused silica column packed with 1.9 µm 120 Å C18. A linear gradient from 5 to 27% buffer B
747 was applied for 50 min, followed by a second one from 27 to 40% buffer B for the next 10 min,
748 and closed by a 2 min step to 80% buffer B, which was used for washing during the last 3 min.
749 Buffer A: 0.1% FA in water; buffer B: 0.1% FA in ACN. The ion mobility range was limited to
750 0.7–1.3 V·s/cm^2, and we applied four windows to each 100 ms diaPASEF scan. Fourteen of these
751 scans covered the doubly and triply charged peptides' diagonal scan line in the m/z ion mobility
752 plane with 28 m/z narrow precursor windows. Other parameters were the same as those set during
753 the ddaPASEF analysis.

754 ***Spectral library construction.*** We used the integrated gut microbiome protein database described
755 in a previous study[18] and named it "IGC+", to quantify microbial proteins. The IGC+ database
756 contains the IGC 9.9 database[74], fungi, virus sequences, and an in-house gut microbial gene
757 catalog. The FragPipe platform (version 15), equipped with MSFragger (version 3.3)[89],
758 Philosopher (version 4.0.0)[90], and EasyPQP (version 0.1.20), were chosen to build the spectral
759 library. A three-step iterative search strategy for metaproteomics [13,27] was used with a few
760 modifications. As detailed described in Fig. S1B, in the first step, we divided the IGC+ database
761 into ten small databases and searched the 88 ddaPASEF files with a target-only version for
762 database search. The matched protein sequences of the ten target-only searches for each
763 ddaPASEF file were combined to reconstruct a simplified database for each MS profile
764 (S1.fas–S88.fas). In the second step, each MS profile was searched against its corresponding
765 simplified database using a target-decoy database search for accurate identifications. Proteins
766 obtained from the 88 searches were aggregated into the concentrated database catalog of the

767 GNHS fecal samples (GNHS.fas). In the third step, the 88 profiles were searched together against
768 the GNHS database combined with a *Homo sapiens* protein catalog (Swiss-Prot, date 20211213)
769 and the 245-contaminant database (MaxQuant, date 20160325) for the peptide to spectrum
770 matches (PSMs) and peptide identifications. For the MSFragger analysis, the precursor and
771 fragment tolerance were set to 20 ppm. Fragment ion series was set to b and y ions. Enzyme
772 specificity was set to 'trypsin', allowing cleavages before proline. Up to two missed cleavages
773 were allowed. The isotope error was set to 0/1/2. The peptide length was set from 7 to 50, and the
774 peptide mass was set from 500 to 5000 Da. The carbamidomethylation of cysteine (57.021464 Da)
775 was set as a fixed modification. The oxidation of methionine (15.99490 Da) and acetyl (Protein
776 N-term) (42.010600 Da) was set as variable modification. The false discovery rate (FDR) was set
777 to 0.01 for all the PSMs, ions, peptides, and proteins in the target-decoy searches. The final output
778 from the iterative database search was imported into EasyPQP to establish the primary spectral
779 library. The retention times of the peptides of each fraction were aligned using the iRT calibration.
780 After obtaining the primary spectral library, we manually assigned peptides to proteins based on
781 the maximum parsimony principle of protein inference (Fig. S2A). The final GNHS fecal spectral
782 library included 506,470 peptides,1,433 unique human proteins with human-specific peptides,
783 121,702 unique microbial proteins with microbiome-specific peptides, and 18,143 microbial
784 protein groups that comprising undistinguished microbial proteins.

785 ***diaPASEF MS data analysis.*** The 2,514 diaPASEF profiles were searched against the final
786 GNHS fecal spectral library using DIA-NN (version 1.8)[91]. The maximum mass accuracy
787 tolerance was set to 10 ppm for both MS1 and MS2 spectra. The quantification mode was set to
788 "Robust LC (high precision)". The protein inference in DIA-NN was disabled. All the other
789 settings were left to their default value. The output matrices, filtered at precursor q value ≤ 1% and
790 global protein q value ≤ 1%, were used for metaproteomic analysis.

791 ***Microbial functional annotation.*** The COGs were mapped by the microbial protein sequences
792 with eggNOG-mapper (version 2.1.5) [92]. The narrowest COG, KOG (Eukaryotic clusters of
793 orthologous genes), and arCOG (Archaeal clusters of orthologous genes) of each protein were
794 extracted. Proteins were also aligned to the KO annotation using GHOSTX from GhostKOALA
795 (version 2.2) [93] with a bit-score threshold of 60. The best hit for each query was selected.
796 Regarding microbial protein groups, we kept the microbial protein groups whose microbial
797 proteins had the same functional annotations. When multiple functions were assigned to the same
798 protein, all the functions were considered.

799 ***Microbial taxonomic annotation.*** Taxonomic annotations were performed using the
800 peptide-centric aligning software Unipept (version 2.2.1, searching date 2021.05.09) [94]. The
801 identified microbial peptides were cut according to the "Advanced missed cleavage handing" rules,
802 and the cut peptides were further filtered to contain more than five and less than 50 amino acids.
803 The filtered peptides were annotated using Unipept with the "Equal I and L" rule. The taxonomic
804 assignment results were then manually filtered to the taxa of bacteria, archaea, fungi, or viruses.
805 All the taxonomic annotations were filtered by at least two distinctive peptides, and their
806 abundances were computed by summing the intensities of all the corresponding peptides. Within
807 the taxonomic tree, the "noname" was added to the name of the narrowest taxon if there were no
808 annotated taxa names in the middle taxon rank.

809 ***Taxonomic and functional profiling of metagenomic samples.*** We also performed metagenomic
810 shotgun sequencing for GNHS participants using the Illumina HiSeq platform (Illumina Inc.). We
811 analyzed metagenomic sequences using a pipeline built as previously described[42]. We trimmed the
812 sequences for quality using PRINSEQ (version 0.20.4) [95], and removed the human genome (*H.*
813 *sapiens*, UCSC hg19) using Bowtie2 tool (version 2.2.5) [96]. Taxonomic profiling was performed
814 using MetaPhlAn2 (version 2.6.0) [97], which classifies metagenomic reads to taxa and yields their
815 relative abundances in each sample. Functional profiling was performed by applying HUMAnN2
816 (version 2.8.1) [98]. This analysis mapped reads to UniProt Reference Clusters (UniRef90) [99] and
817 further grouped them into 9,381 COG groups.

818 ***Analysis of richness, intensity, and BC dissimilarity.*** We mainly used three types of matrices for
819 analysis, the human protein matrix, the microbial taxonomic composition matrices, and the
820 microbial functional matrices. Three parameters were used to assess the differences between study
821 individuals, levels of richness, intensity, and BC dissimilarity between samples. The BC distance
822 was calculated using the function *vegdist* of the package vegan (version 2.5-7) in R.

823 ***Quality control analysis.*** The data reproducibility was evaluated by the Spearman pairwise
824 correlation and the BC distance between replicates. PCAs, performed with the command *prcomp*
825 in R, was used to evaluate the batch effects of this study.

826 ***Associations between matrices.*** The associations between matrices were done using the Mantel
827 test. Specifically, the BC distance dissimilarity of the two matrices was inputted to calculate their
828 Spearman correlation by the function *mantel* in vegan (version 2.5-7).

829 ***Features' accumulation curves and estimation of feature richness.*** The accumulation curves
830 were plotted based on the observed richness of features with increasing numbers of collected
831 samples. The observed richness was calculated in subsets of increasing size composed of
832 randomly chosen samples (500 repetitions for each sample size). The estimated richness was
833 calculated based on the Chao1 index using the function *poolaccum* from the package vegan
834 (version 2.5-7) in R. When the samples reached a saturation number (*i.e.*, the observed richness
835 was equal to the estimated saturation number and remained stable even with further increased
836 sample number), we can assume these samples to cover the full diversity of the studied
837 population.

838 ***GO enrichment analysis.*** The human protein GO biological process enrichment was performed
839 using the package clusterProfiler (version 4.2.0) in R. The package org.HS.eg.db (version 3.14.0)
840 was used to annotate the human proteins. For statistically significant results, both the BH-adjusted
841 *p*-value and the q-value were required to be lower than 0.05.

842 ***Microbial function pathway enrichment.*** The pathway enrichment was performed using the R
843 package MicrobiomeProfiler. For population-level pathway enrichment, the core KOs of
844 metaproteomics and metagenomics were selected and enriched by enrichKO function with
845 parameters minGSSize =10 and maxGSSize = 200. Statistically significant enriched pathways
846 were screened with BH-adjusted p value ≤ 0.05, and fold enrichment value (GeneRatio / BgRatio)
847 ≥ 2.

848 ***Fecal butyric acid content analysis.*** The fecal butyric acid content was detected by targeted
849 metabolomics through ultra-high-performance liquid chromatography-tandem mass spectrometry

850 (UPLC-MS/MS) system (ACQUITY UPLC-Xevo TQ-S, Waters Corp., MA, USA). Detailed
851 information about the measurements could be refered to the previously study (Jiang et al., 2020).

852 ***Covariate analysis.*** The associations between each phenotypic factor and variation of human
853 protein, gut microbiome function, or gut microbial taxon were determined by PERMANOVA on
854 the human protein, metaprotein, KO, COG, species or genus levels (Bray-Curtis dissimilarities
855 using the function *adonis* and the vegan R package, version 2.5-7). The *p*-value of each
856 phenotypic factor was determined using 1,000 permutations and adjusted with the BH method.
857 The phenotypic factors with adjusted $p \leq 0.05$ were selected as covariates.

858 ***Association analysis between individual features and the metadata.*** The generalized linear model
859 (GLM) (implemented with the R package glm, version 4.1.1) was used to identify the association
860 between the abundance of features (metaproteins, human proteins, COGs, KOs, species, and
861 genera) and each phenotype. Only the features with higher than 10% prevalence in the GNHS
862 cohort were used for the GLM analysis. Then the feature abundance data were firstly INT
863 transformed to bring them all on the same scale. To avoid the covariate effects potentially
864 associated with feature abundances, for the phenotypes excluded "Sex", "Age" and "Bristol Scale",
865 we included these three metadata variables as the confounding covariates in the GLM models. In
866 the GLM model for the three phenotypes, we use the other two metadata variables as the
867 confounding covariates. For each *p*-value, multiple testing corrections were performed using the
868 BH adjustment. The associations with q value (adjusted p) ≤ 0.05 were considered statistically
869 significant.

870 ***Machine learning.*** A total of 785 metaproteins, 23 human proteins, 268 KOs, 279 COGs, 12
871 species, and 14 genera that were related to T2D (GLM, $q \leq 0.05$) were used for the
872 metaproteomic-based T2D classification. The GNHS samples were randomly divided into the
873 training set and internal validation set with a 4:1 ratio. Then we started to train the model using the
874 training set. Firstly, we used the eXtreme Gradient Boosting (XGBoost) wrapped recursive feature
875 elimination algorithm to select the features for our model at metaprotein, human protein, KO,
876 COG, species, or genus level, respectively. We then counted the occurrence of each feature in the
877 100 candidate feature sets. We extracted the features with more than 70 occurrences as the final
878 input feature set and feed it into an XGB model. We then fine-tuned the XGBoost model using the
879 Genetic Algorithm hyperparameter searching with an early stop of three generations without AUC
880 improvements. Finally, we tested the trained model in the internal validation set to check the
881 generalization of the model. And an independent cohort (FH cohort) with 104 individuals was
882 introduced for an independent test of the model.

883 ***Co-occurrence Network.*** The linkages between the species and their corresponding functions
884 were generated by their common peptide sequences, while the co-expressed linkages between
885 intra-levels and inter-levels were calculated by the Spearman correlation (BH adjusted p-value $\leq$
886 0.05, $|r| \geq 0.3$ for inter-levels, $|r| \geq 0.8$ for intra-levels).

887 ***Data visualization.*** The R packages ggplot2 (version 4.1.3), ggpubr (version 0.4.0), gghalves
888 (version 0.1.1), Goplot (version 1.0.2), forplo (version 0.1.0) and venn (version 1.10) were used
889 for data visualization. The taxonomic tree was visualized using graphlan (version 1.1.3). The
890 co-occurrence networks were visualized with the interactive platform Gephi (version 0.9.2).
891 Additionally, we used the Fruchterman Reingold layout algorithm to visualize the networks. The

modularity was built using the fast unfolding of communities in large networks. Different node colors represented different module communities. The size of the nodes was determined using the number of links between the nodes with other nodes.

In vitro bacterial growth experiments. Precultures of *M. elsdenii* DSM 20460 were inoculated in anaerobic workstation (5% hydrogen, 10% carbon dioxide, and 85% nitrogen) as single colonies in modified reinforced clostridial medium (mRCM) medium. The medium contained the following components per liter: peptone (10g), beef extract powder (10g), yeast extract (3g), glucose (5g), fructose (10g), sodium chloride (5g), L-cysteine hydrochloride (0.5g), sodium acetate (3g), and resazurin (0.001g). After 24 hours of incubation, the preculture was inoculated into fresh mRCM broth at a concentration of 4% (v/v) and incubated for an additional 15 hours. The broth was supplemented with either 20 mM metformin, 20 mM acarbose, or 50 μm glimpiride in a 96-well microplate with 12 replicates per group. The impact of the drugs on bacterial growth was assessed by measuring the OD600 value of the bacterial culture and counting the number of live bacterial cells using BactoBox V7.2 (SBT Instruments), to ensure consistent results.

Animal trail. Mouse experiments and all associated manipulations were carried out in accordance with the approved protocol by the research ethics committee of GNOTOBIO (permit JTAW20230705), following standard care and husbandry guidelines. Germ-free male C57BL/6 mice were randomly assigned to four ABSL-1 animal facilities with controlled environments. These environments maintained a 12-hour dark-light cycle (7 p.m. to 7 a.m.), 50±10% humidity, a temperature of 22 ± 2 °C, and provided ad libitum access to food and water. Following a 3-day acclimation period, the mice were divided into four experimental groups (n = 7). (1) Control group (CD group): These mice were fed a chow diet (catalog no. D12450J, Research Diets) with 10% of calories from fat, 20% from protein, and 70% from carbohydrates, providing 382 kcal per 100 g. They were administered 0.9% saline solution by gavage six times in the first two weeks. (2) High-fat diet group (HFD group): These mice were fed a high-fat diet (catalog no. D12492i, Research Diets) with 60% of calories from fat, 20% from protein, and 20% from carbohydrates, providing 524 kcal per 100 g. They were administered 0.9% saline solution by gavage six times in the first two weeks. (3) HFD+*M. elsdenii* group: These mice were fed a high-fat diet and administered *M. elsdenii* cells at a concentration of 10^9 CFU by gavage six times in the first two weeks. (4) HFD+*L. reuteri* group: These mice were fed a high-fat diet and administered *L. reuteri* cells at a concentration of 10^9 CFU by gavage six times in the first two weeks. After exsanguination, the mice were euthanized by cervical dislocation, and the intestines were dissected to collect the corresponding contents.

926 **References**

927 1 Cani, P. D. Human gut microbiome: hopes, threats and promises. *Gut* **67**, 1716-1725,
928 doi:10.1136/gutjnl-2018-316723 (2018).
929 2 Shreiner, A. B., Kao, J. Y. & Young, V. B. The gut microbiome in health and in disease.
930 *Curr Opin Gastroenterol* **31**, 69-75, doi:10.1097/MOG.0000000000000139 (2015).
931 3 Asnicar, F. *et al.* Microbiome connections with host metabolism and habitual diet from
932 1,098 deeply phenotyped individuals. *Nat Med* **27**, 321-332,
933 doi:10.1038/s41591-020-01183-8 (2021).
934 4 Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science* **352**,
935 560-564, doi:10.1126/science.aad3503 (2016).
936 5 Kurilshikov, A. *et al.* Large-scale association analyses identify host factors influencing
937 human gut microbiome composition. *Nat Genet* **53**, 156-165,
938 doi:10.1038/s41588-020-00763-1 (2021).
939 6 Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel
940 diseases. *Nature* **569**, 655-662, doi:10.1038/s41586-019-1237-9 (2019).
941 7 Rothschild, D. *et al.* Environment dominates over host genetics in shaping human gut
942 microbiota. *Nature* **555**, 210-215, doi:10.1038/nature25973 (2018).
943 8 Muth, T., Renard, B. Y. & Martens, L. Metaproteomic data analysis at a glance: advances
944 in computational microbial community proteomics. *Expert Rev Proteomics* **13**, 757-769,
945 doi:10.1080/14789450.2016.1209418 (2016).
946 9 Wilmes, P. & Bond, P. L. Metaproteomics: studying functional gene expression in
947 microbial ecosystems. *Trends Microbiol* **14**, 92-97, doi:10.1016/j.tim.2005.12.006 (2006).
948 10 Kleiner, M. *et al.* Assessing species biomass contributions in microbial communities via
949 metaproteomics. *Nat Commun* **8**, 1558, doi:10.1038/s41467-017-01544-x (2017).
950 11 Salvato, F., Hettich, R. L. & Kleiner, M. Five key aspects of metaproteomics as a tool to
951 understand functional interactions in host-associated microbiomes. *PLoS Pathog* **17**,
952 e1009245, doi:10.1371/journal.ppat.1009245 (2021).
953 12 Zhang, X. *et al.* Deep Metaproteomics Approach for the Study of Human Microbiomes.
954 *Anal Chem* **89**, 9407-9415, doi:10.1021/acs.analchem.7b02224 (2017).
955 13 Bassignani, A. *et al.* Benefits of Iterative Searches of Large Databases to Interpret Large
956 Human Gut Metaproteomic Data Sets. *J Proteome Res* **20**, 1522-1534,
957 doi:10.1021/acs.jproteome.0c00669 (2021).
958 14 Gavin, P. G. *et al.* Intestinal Metaproteomics Reveals Host-Microbiota Interactions in
959 Subjects at Risk for Type 1 Diabetes. *Diabetes Care* **41**, 2178-2186,
960 doi:10.2337/dc18-0777 (2018).
961 15 Long, S. *et al.* Metaproteomics characterizes human gut microbiome function in
962 colorectal cancer. *NPJ Biofilms Microbiomes* **6**, 14, doi:10.1038/s41522-020-0123-4
963 (2020).
964 16 Mills, R. H. *et al.* Multi-omics analyses of the ulcerative colitis gut microbiome link
965 Bacteroides vulgatus proteases with disease severity. *Nat Microbiol* **7**, 262-276,
966 doi:10.1038/s41564-021-01050-3 (2022).
967 17 Tran, H. Q. *et al.* Associations of the Fecal Microbial Proteome Composition and
968 Proneness to Diet-induced Obesity. *Mol Cell Proteomics* **18**, 1864-1879,
969 doi:10.1074/mcp.RA119.001623 (2019).

970 18 Zhang, X. *et al.* Metaproteomics reveals associations between microbiome and intestinal
971 extracellular vesicle proteins in pediatric inflammatory bowel disease. *Nat Commun* **9**,
972 2873, doi:10.1038/s41467-018-05357-4 (2018).
973 19 Zhang, X., Li, L., Butcher, J., Stintzi, A. & Figeys, D. Advancing functional and
974 translational microbiome research using meta-omics approaches. *Microbiome* **7**, 154,
975 doi:10.1186/s40168-019-0767-6 (2019).
976 20 Zhong, H. *et al.* Distinct gut metagenomics and metaproteomics signatures in prediabetics
977 and treatment-naive type 2 diabetics. *EBioMedicine* **47**, 373-383,
978 doi:10.1016/j.ebiom.2019.08.048 (2019).
979 21 Aakko, J. *et al.* Data-Independent Acquisition Mass Spectrometry in Metaproteomics of
980 Gut Microbiota-Implementation and Computational Analysis. *J Proteome Res* **19**,
981 432-436, doi:10.1021/acs.jproteome.9b00606 (2020).
982 22 Beyter, D., Lin, M. S., Yu, Y., Pieper, R. & Bafna, V. ProteoStorm: An Ultrafast
983 Metaproteomics Database Search Framework. *Cell Syst* **7**, 463-467 e466,
984 doi:10.1016/j.cels.2018.08.009 (2018).
985 23 Mesuere, B. *et al.* Unipept: tryptic peptide-based biodiversity analysis of metaproteome
986 samples. *J Proteome Res* **11**, 5773-5780, doi:10.1021/pr300576s (2012).
987 24 Muth, T. *et al.* The MetaProteomeAnalyzer: a powerful open-source software suite for
988 metaproteomics data analysis and interpretation. *J Proteome Res* **14**, 1557-1565,
989 doi:10.1021/pr501246w (2015).
990 25 Rooijers, K. *et al.* An iterative workflow for mining the human intestinal metaproteome.
991 *BMC Genomics* **12**, 6, doi:10.1186/1471-2164-12-6 (2011).
992 26 Werner, J., Geron, A., Kerssemakers, J. & Matallana-Surget, S. mPies: a novel
993 metaproteomics tool for the creation of relevant protein databases and automatized
994 protein annotation. *Biol Direct* **14**, 21, doi:10.1186/s13062-019-0253-x (2019).
995 27 Zhang, X. *et al.* MetaPro-IQ: a universal metaproteomic approach to studying human and
996 mouse gut microbiota. *Microbiome* **4**, 31, doi:10.1186/s40168-016-0176-z (2016).
997 28 Shanahan, F., Ghosh, T. S. & O'Toole, P. W. The Healthy Microbiome-What Is the
998 Definition of a Healthy Gut Microbiome? *Gastroenterology* **160**, 483-494,
999 doi:10.1053/j.gastro.2020.09.057 (2021).
1000 29 Gillet, L. C. *et al.* Targeted data extraction of the MS/MS spectra generated by
1001 data-independent acquisition: a new concept for consistent and accurate proteome
1002 analysis. *Mol Cell Proteomics* **11**, O111 016717, doi:10.1074/mcp.O111.016717 (2012).
1003 30 Guo, T. *et al.* Rapid mass spectrometric conversion of tissue biopsy samples into
1004 permanent quantitative digital proteome maps. *Nat Med* **21**, 407-413,
1005 doi:10.1038/nm.3807 (2015).
1006 31 Yue, L. *et al.* Generating Proteomic Big Data for Precision Medicine. *Proteomics* **20**,
1007 e1900358, doi:10.1002/pmic.201900358 (2020).
1008 32 Zhu, Y., Aebersold, R., Mann, M. & Guo, T. SnapShot: Clinical proteomics. *Cell* **184**,
1009 4840-4840 e4841, doi:10.1016/j.cell.2021.08.015 (2021).
1010 33 Kanu, A. B., Dwivedi, P., Tam, M., Matz, L. & Hill, H. H., Jr. Ion mobility-mass
1011 spectrometry. *J Mass Spectrom* **43**, 1-22, doi:10.1002/jms.1383 (2008).
1012 34 Lanucara, F., Holman, S. W., Gray, C. J. & Eyers, C. E. The power of ion mobility-mass
1013 spectrometry for structural characterization and the study of conformational dynamics.

1014 *Nat Chem* **6**, 281-294, doi:10.1038/nchem.1889 (2014).
1015 35 Demichev, V. *et al.* dia-PASEF data analysis using FragPipe and DIA-NN for deep
1016 proteomics of low sample amounts. *Nat Commun* **13**, 3944,
1017 doi:10.1038/s41467-022-31492-0 (2022).
1018 36 Meier, F. *et al.* diaPASEF: parallel accumulation-serial fragmentation combined with
1019 data-independent acquisition. *Nat Methods* **17**, 1229-1236,
1020 doi:10.1038/s41592-020-00998-0 (2020).
1021 37 Jagtap, P. *et al.* A two-step database search method improves sensitivity in peptide
1022 sequence matches for metaproteomics and proteogenomics studies. *Proteomics* **13**,
1023 1352-1357, doi:10.1002/pmic.201200352 (2013).
1024 38 Nesvizhskii, A. I. & Aebersold, R. Interpretation of shotgun proteomic data: the protein
1025 inference problem. *Mol Cell Proteomics* **4**, 1419-1440,
1026 doi:10.1074/mcp.R500012-MCP200 (2005).
1027 39 Zhou, H., Beltran, J. F. & Brito, I. L. Host-microbiome protein-protein interactions
1028 capture disease-relevant pathways. *Genome Biol* **23**, 72,
1029 doi:10.1186/s13059-022-02643-9 (2022).
1030 40 Gacesa, R. *et al.* Environmental factors shaping the gut microbiome in a Dutch population.
1031 *Nature* **604**, 732-739, doi:10.1038/s41586-022-04567-7 (2022).
1032 41 Shuai, M. *et al.* Human Gut Antibiotic Resistome and Progression of Diabetes. *Adv Sci*
1033 *(Weinh)* **9**, e2104965, doi:10.1002/advs.202104965 (2022).
1034 42 Maron, P. A., Ranjard, L., Mougel, C. & Lemanceau, P. Metaproteomics: a new approach
1035 for studying functional microbial ecology. *Microb Ecol* **53**, 486-493,
1036 doi:10.1007/s00248-006-9196-8 (2007).
1037 43 Nayfach, S. & Pollard, K. S. Toward Accurate and Quantitative Comparative
1038 Metagenomics. *Cell* **166**, 1103-1116, doi:10.1016/j.cell.2016.08.007 (2016).
1039 44 Human Microbiome Project, C. Structure, function and diversity of the healthy human
1040 microbiome. *Nature* **486**, 207-214, doi:10.1038/nature11234 (2012).
1041 45 Chen, L. *et al.* The long-term genetic stability and individual specificity of the human gut
1042 microbiome. *Cell* **184**, 2302-2315 e2312, doi:10.1016/j.cell.2021.03.024 (2021).
1043 46 Abu-Ali, G. S. *et al.* Metatranscriptome of human faecal microbial communities in a
1044 cohort of adult men. *Nat Microbiol* **3**, 356-366, doi:10.1038/s41564-017-0084-4 (2018).
1045 47 Jandhyala, S. M. *et al.* Role of the normal gut microbiota. *World J Gastroenterol* **21**,
1046 8787-8803, doi:10.3748/wjg.v21.i29.8787 (2015).
1047 48 Olofsson, L. E. & Backhed, F. The Metabolic Role and Therapeutic Potential of the
1048 Microbiome. *Endocr Rev* **43**, 907-926, doi:10.1210/endrev/bnac004 (2022).
1049 49 Vieira-Silva, S. *et al.* Statin therapy is associated with lower prevalence of gut microbiota
1050 dysbiosis. *Nature* **581**, 310-315, doi:10.1038/s41586-020-2269-x (2020).
1051 50 Aasmets, O., Krigul, K. L., Lull, K., Metspalu, A. & Org, E. Gut metagenome
1052 associations with extensive digital health data in a volunteer-based Estonian microbiome
1053 cohort. *Nat Commun* **13**, 869, doi:10.1038/s41467-022-28464-9 (2022).
1054 51 Zhang, X. *et al.* Sex- and age-related trajectories of the adult human gut microbiota
1055 shared across populations of different ethnicities. *Nature Aging* **1**, 87-100,
1056 doi:10.1038/s43587-020-00014-2 (2021).
1057 52 Zhernakova, A. *et al.* Population-based metagenomics analysis reveals markers for gut

1058 microbiome composition and diversity. *Science* **352**, 565-569,
1059 doi:10.1126/science.aad3369 (2016).
1060 53 Sugitani, Y. *et al.* Mucosa-associated gut microbiome in Japanese patients with functional
1061 constipation. *J Clin Biochem Nutr* **68**, 187-192, doi:10.3164/jcbn.20-93 (2021).
1062 54 Kanwal, A. *et al.* Exploring New Drug Targets for Type 2 Diabetes: Success, Challenges
1063 and Opportunities. *Biomedicines* **10**, doi:10.3390/biomedicines10020331 (2022).
1064 55 Zhong, S. *et al.* Dissection of gut microbiota and metabolites reveals the hypolipidemic
1065 effect of green mulberry leaf tea / black mulberry leaf tea in mice. *Journal of Functional*
1066 *Foods* **111**, doi:10.1016/j.jff.2023.105906 (2023).
1067 56 Liao, W. *et al.* Potential prebiotic effects of nonabsorptive components of Keemun and
1068 Dianhong black tea: an in vitro study. *Food Science and Human Wellness* **11**, 648-659,
1069 doi:10.1016/j.fshw.2021.12.022 (2022).
1070 57 Bosco, N. & Noti, M. The aging gut microbiome and its impact on host immunity. *Genes*
1071 *Immun* **22**, 289-303, doi:10.1038/s41435-021-00126-8 (2021).
1072 58 Nagpal, R. *et al.* Gut microbiome and aging: Physiological and mechanistic insights. *Nutr*
1073 *Healthy Aging* **4**, 267-285, doi:10.3233/NHA-170030 (2018).
1074 59 Sun, J., Rutherford, S. T., Silhavy, T. J. & Huang, K. C. Physical properties of the
1075 bacterial outer membrane. *Nat Rev Microbiol* **20**, 236-248,
1076 doi:10.1038/s41579-021-00638-0 (2022).
1077 60 Heianza, Y. *et al.* Gut microbiota metabolites, amino acid metabolites and improvements
1078 in insulin sensitivity and glucose metabolism: the POUNDS Lost trial. *Gut* **68**, 263-270,
1079 doi:10.1136/gutjnl-2018-316155 (2019).
1080 61 Ma, S. *et al.* Alterations in Gut Microbiota of Gestational Diabetes Patients During the
1081 First Trimester of Pregnancy. *Front Cell Infect Microbiol* **10**, 58,
1082 doi:10.3389/fcimb.2020.00058 (2020).
1083 62 Salazar, J. *et al.* Microbiota and Diabetes Mellitus: Role of Lipid Mediators. *Nutrients* **12**,
1084 doi:10.3390/nu12103039 (2020).
1085 63 Graciano, M. F., Valle, M. M., Kowluru, A., Curi, R. & Carpinelli, A. R. Regulation of
1086 insulin secretion and reactive oxygen species production by free fatty acids in pancreatic
1087 islets. *Islets* **3**, 213-223, doi:10.4161/isl.3.5.15935 (2011).
1088 64 Letchumanan, G. *et al.* Gut Microbiota Composition in Prediabetes and Newly Diagnosed
1089 Type 2 Diabetes: A Systematic Review of Observational Studies. *Front Cell Infect*
1090 *Microbiol* **12**, 943427, doi:10.3389/fcimb.2022.943427 (2022).
1091 65 Treadway, J. L., Mendys, P. & Hoover, D. J. Glycogen phosphorylase inhibitors for
1092 treatment of type 2 diabetes mellitus. *Expert Opin Investig Drugs* **10**, 439-454,
1093 doi:10.1517/13543784.10.3.439 (2001).
1094 66 Hashizume, K., Tsukahara, T., Yamada, K., Koyama, H. & Ushida, K. Megasphaera
1095 elsdenii JCM1772T normalizes hyperlactate production in the large intestine of
1096 fructooligosaccharide-fed rats by stimulating butyrate production. *J Nutr* **133**, 3187-3190,
1097 doi:10.1093/jn/133.10.3187 (2003).
1098 67 Shetty, S. A., Marathe, N. P., Lanjekar, V., Ranade, D. & Shouche, Y. S. Comparative
1099 genome analysis of Megasphaera sp. reveals niche specialization and its potential role in
1100 the human gut. *PLoS One* **8**, e79353, doi:10.1371/journal.pone.0079353 (2013).
1101 68 Gao, Z. *et al.* Butyrate improves insulin sensitivity and increases energy expenditure in

1102 mice. *Diabetes* **58**, 1509-1517, doi:10.2337/db08-1637 (2009).
1103 69 Roshanravan, N. *et al.* Effect of Butyrate and Inulin Supplementation on Glycemic Status,
1104 Lipid Profile and Glucagon-Like Peptide 1 Level in Patients with Type 2 Diabetes: A
1105 Randomized Double-Blind, Placebo-Controlled Trial. *Horm Metab Res* **49**, 886-891,
1106 doi:10.1055/s-0043-119089 (2017).
1107 70 de la Cuesta-Zuluaga, J. *et al.* Metformin Is Associated With Higher Relative Abundance
1108 of Mucin-Degrading Akkermansia muciniphila and Several Short-Chain Fatty
1109 Acid-Producing Microbiota in the Gut. *Diabetes Care* **40**, 54-62, doi:10.2337/dc16-1324
1110 (2017).
1111 71 Gu, Y. *et al.* Analyses of gut microbiota and plasma bile acids enable stratification of
1112 patients for antidiabetic treatment. *Nat Commun* **8**, 1785,
1113 doi:10.1038/s41467-017-01682-2 (2017).
1114 72 Mobini, R. *et al.* Metabolic effects of Lactobacillus reuteri DSM 17938 in people with
1115 type 2 diabetes: A randomized controlled trial. *Diabetes Obes Metab* **19**, 579-589,
1116 doi:10.1111/dom.12861 (2017).
1117 73 Simon, M. C. *et al.* Intake of Lactobacillus reuteri improves incretin and insulin secretion
1118 in glucose-tolerant humans: a proof of concept. *Diabetes Care* **38**, 1827-1834,
1119 doi:10.2337/dc14-2690 (2015).
1120 74 Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat*
1121 *Biotechnol* **32**, 834-841, doi:10.1038/nbt.2942 (2014).
1122 75 Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut
1123 microbiome. *Nat Biotechnol* **39**, 105-114, doi:10.1038/s41587-020-0603-3 (2021).
1124 76 Summers, N. A. *et al.* Metabolic Changes Associated With the Use of Integrase Strand
1125 Transfer Inhibitors Among Virally Controlled Women. *J Acquir Immune Defic Syndr* **85**,
1126 355-362, doi:10.1097/QAI.0000000000002447 (2020).
1127 77 Culp, E. J. & Goodman, A. L. Cross-feeding in the gut microbiome: Ecology and
1128 mechanisms. *Cell Host Microbe* **31**, 485-499, doi:10.1016/j.chom.2023.03.016 (2023).
1129 78 Manor, O. *et al.* Health and disease markers correlate with gut microbiome composition
1130 across thousands of people. *Nat Commun* **11**, 5206, doi:10.1038/s41467-020-18871-1
1131 (2020).
1132 79 Zhang, X. *et al.* Sex- and age-related trajectories of the adult human gut microbiota
1133 shared across populations of different ethnicities. *Nat Aging* **1**, 87-100,
1134 doi:10.1038/s43587-020-00014-2 (2021).
1135 80 Cunningham, A. L., Stephens, J. W. & Harris, D. A. Gut microbiota influence in type 2
1136 diabetes mellitus (T2DM). *Gut Pathog* **13**, 50, doi:10.1186/s13099-021-00446-0 (2021).
1137 81 Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes.
1138 *Nature* **490**, 55-60, doi:10.1038/nature11450 (2012).
1139 82 Ruuskanen, M. O. *et al.* Gut Microbiome Composition Is Predictive of Incident Type 2
1140 Diabetes in a Population Cohort of 5,572 Finnish Adults. *Diabetes Care* **45**, 811-818,
1141 doi:10.2337/dc21-2358 (2022).
1142 83 Zhang, B. *et al.* Global research trends on the links between the gut microbiota and
1143 diabetes between 2001 and 2021: A bibliometrics and visualized study. *Front Microbiol*
1144 **13**, 1011050, doi:10.3389/fmicb.2022.1011050 (2022).
1145 84 Talukdar, R. *et al.* The gut microbiome in pancreatogenic diabetes differs from that of

Type 1 and Type 2 diabetes. *Sci Rep* **11**, 10978, doi:10.1038/s41598-021-90024-w (2021).

85 Zhang, Q. & Hu, N. Effects of Metformin on the Gut Microbiota in Obesity and Type 2 Diabetes Mellitus. *Diabetes Metab Syndr Obes* **13**, 5003-5014, doi:10.2147/DMSO.S286430 (2020).

86 Zhang, Z. Q. *et al.* Association between dietary intake of flavonoid and bone mineral density in middle aged and elderly Chinese women and men. *Osteoporos Int* **25**, 2417-2425, doi:10.1007/s00198-014-2763-9 (2014).

87 Gou, W. *et al.* Interpretable Machine Learning Framework Reveals Robust Gut Microbiome Features Associated With Type 2 Diabetes. *Diabetes Care* **44**, 358-366, doi:10.2337/dc20-1536 (2021).

88 Escher, C. *et al.* Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics* **12**, 1111-1121, doi:10.1002/pmic.201100463 (2012).

89 Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat Methods* **14**, 513-520, doi:10.1038/nmeth.4256 (2017).

90 da Veiga Leprevost, F. *et al.* Philosopher: a versatile toolkit for shotgun proteomics data analysis. *Nat Methods* **17**, 869-870, doi:10.1038/s41592-020-0912-y (2020).

91 Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat Methods* **17**, 41-44, doi:10.1038/s41592-019-0638-x (2020).

92 Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* **47**, D309-D314, doi:10.1093/nar/gky1085 (2019).

93 Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J Mol Biol* **428**, 726-731, doi:10.1016/j.jmb.2015.11.006 (2016).

94 Mesuere, B. *et al.* High-throughput metaproteomics data analysis with Unipept: A tutorial. *J Proteomics* **171**, 11-22, doi:10.1016/j.jprot.2017.05.022 (2018).

95 Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863-864, doi:10.1093/bioinformatics/btr026 (2011).

96 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

97 Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods* **12**, 902-903, doi:10.1038/nmeth.3589 (2015).

98 Franzosa, E. A. *et al.* Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods* **15**, 962-968, doi:10.1038/s41592-018-0176-y (2018).

99 Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282-1288, doi:10.1093/bioinformatics/btm098 (2007).

Figure 1

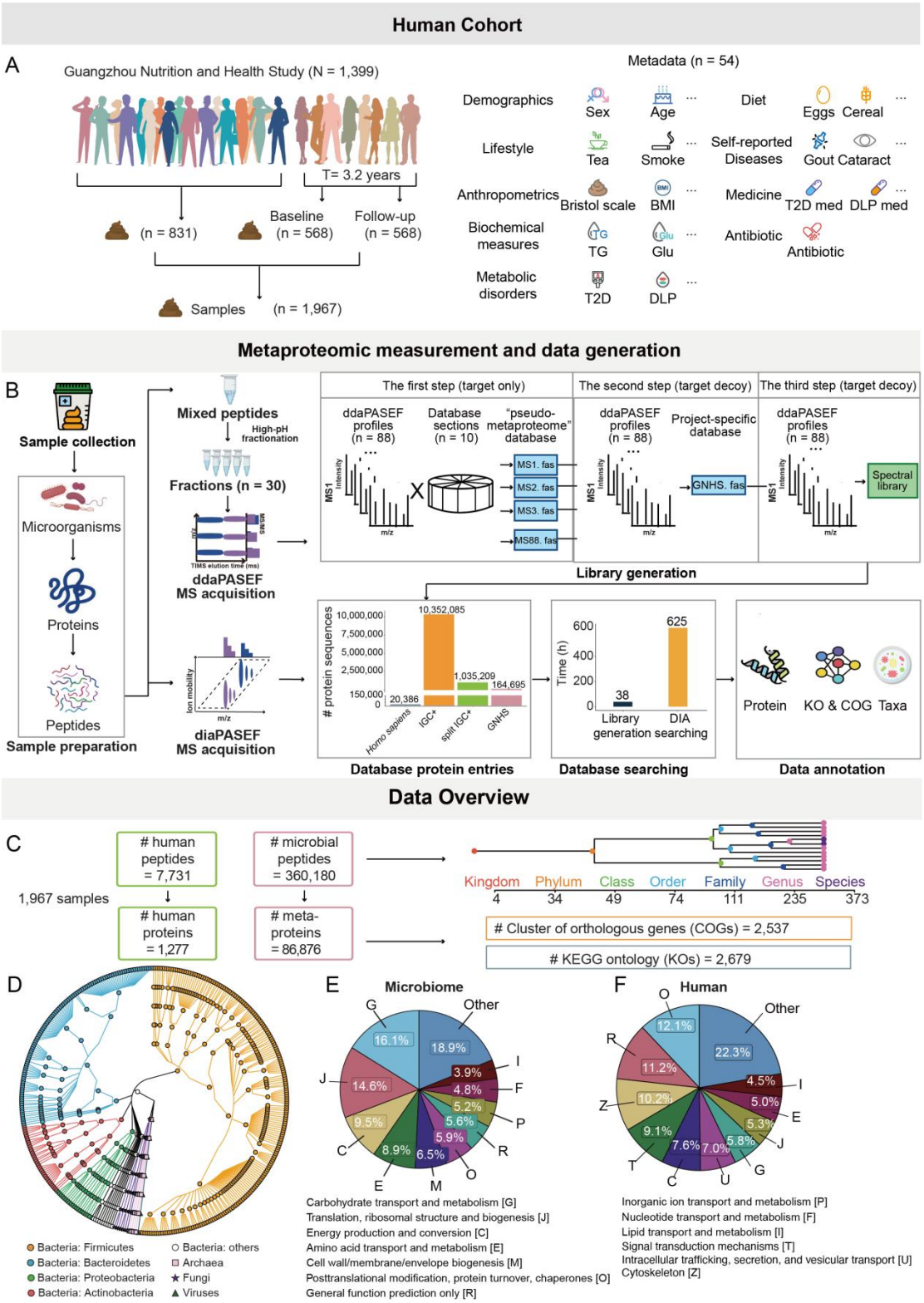


Figure 1. Metaproteomic profiling of the gut microbiome. (A) Population distribution of individuals. Our study covered 1399 individuals from the GNHS cohort. Of these, 831 individuals contributed one sample only, while 568 individuals provided two samples taken approximately 3.2 years apart. This cohort also had 49 available phenotype-related information, including types of

1194 sociodemographics, lifestyle, metabolic disorders, biochemical measures, and anthropometrics. (B)
1195 The experimental strategy of diaPASEF-based metaproteomic profiling of 2,514 MS runs in the
1196 GNHS cohort, of which 1,967 were for our investigations and 547 for quality assessment. (C)
1197 Metaproteomic identifications of 1967 fecal samples at the levels of microbial protein, taxonomy,
1198 and function. (D) Taxonomic tree based on metaproteomic data. The taxonomic levels range from
1199 superkingdoms to species from the inner to the outer circle. (E) The functional distributions of
1200 microbiota. The functional category was annotated using the COG database. (F) The functional
1201 distributions of human proteins. The functional category was annotated using the COG database.

Figure 2

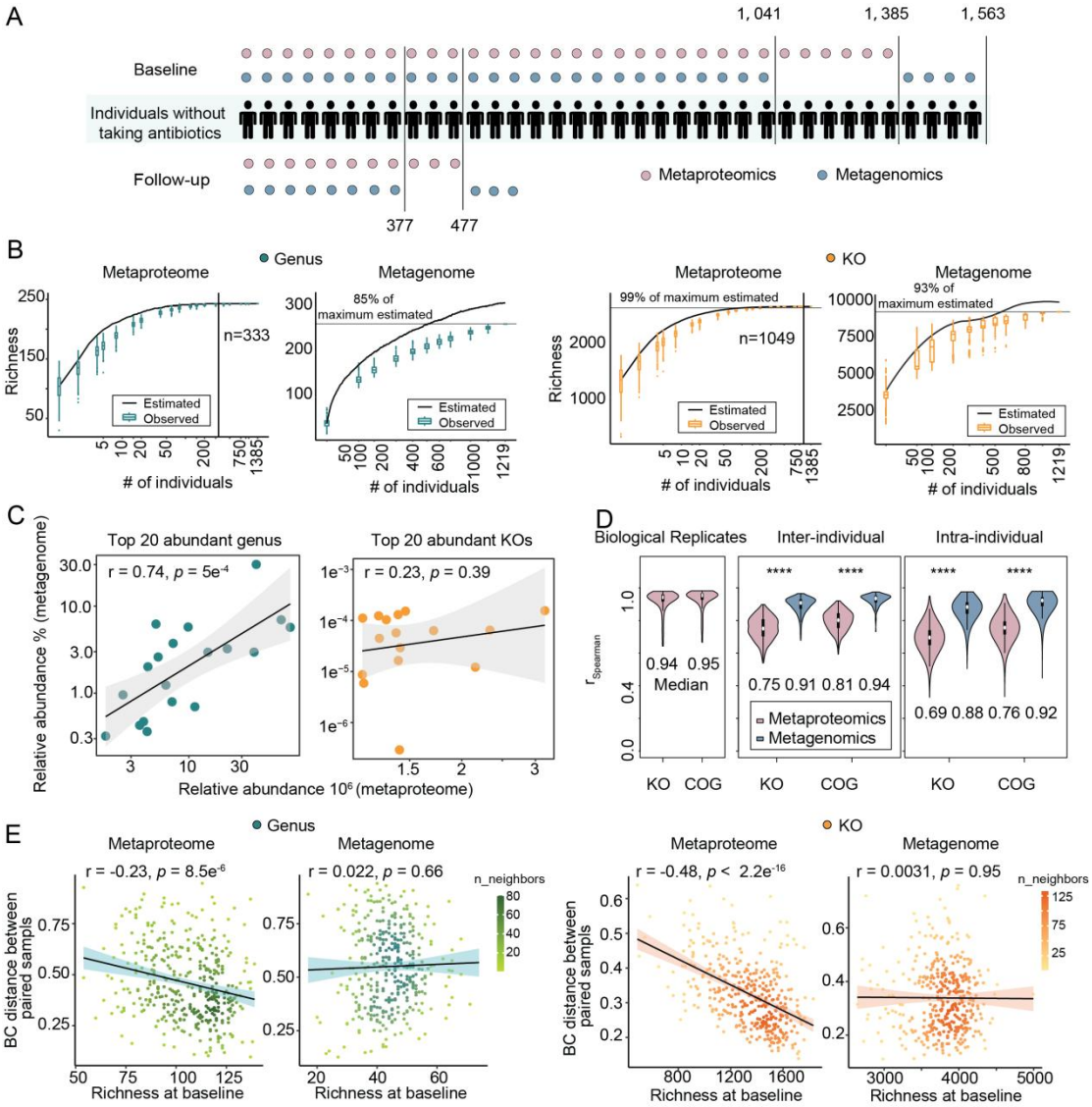
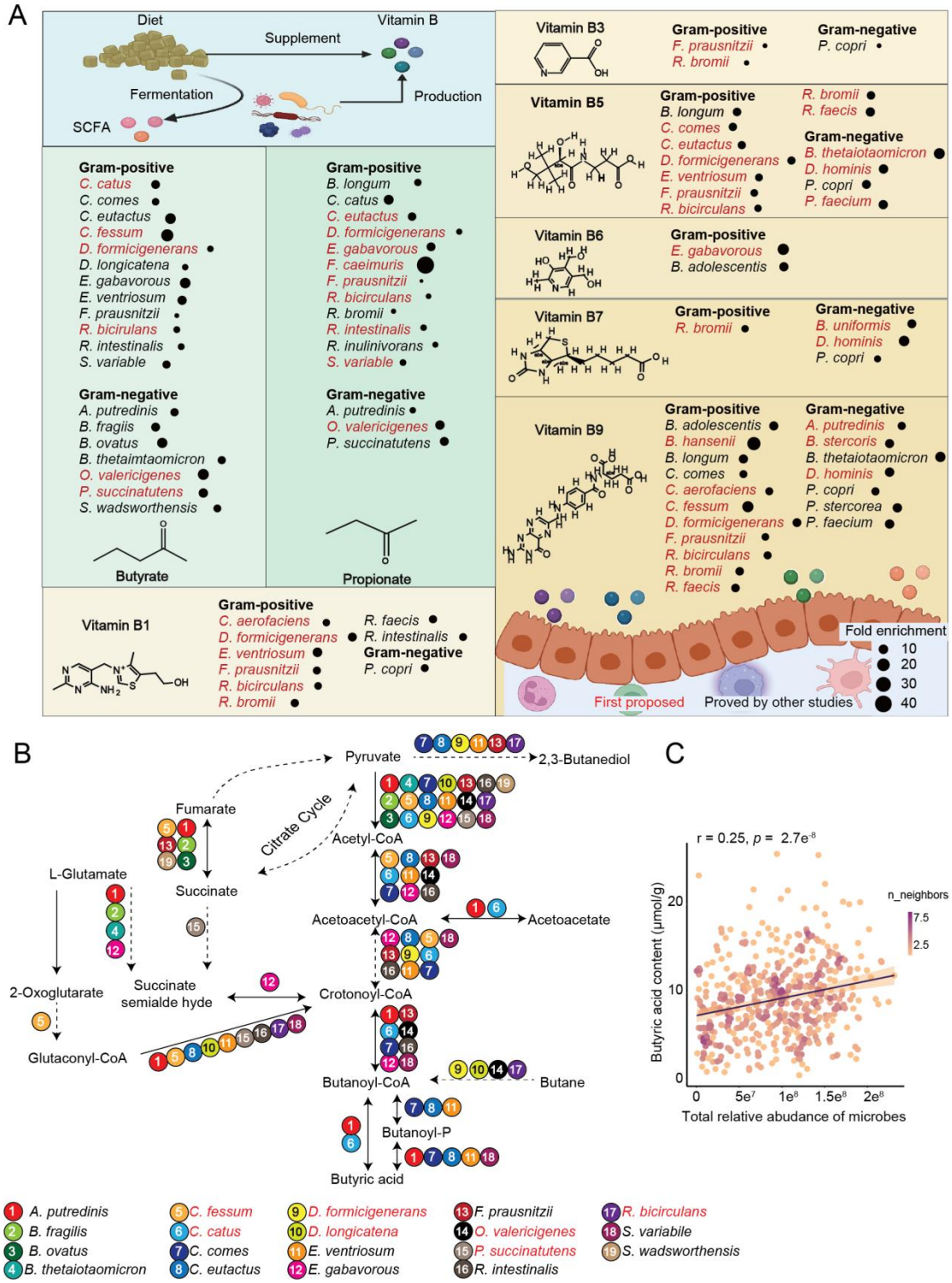


Figure 2. Comparison of metaproteomic and metagenomic data of paired samples. (A)The GNHS cohort collected fecal samples during baseline and follow-up, which were then subjected to metaproteomic or metagenomic analysis. (B) Genus or KO accumulation curves of metaproteomic and metagenomic data. The x-axes represent the cumulative sample number, and the y-axes mean the total richness of each feature. The black curves show the estimated richness; the box plots report the measured richness. The maximum measured richness and its proportion to the maximum estimated richness are indicated by a horizontal line. The vertical lines represent the minimum sample numbers required for full representation of GNHS cohort diversity. (C) Quantitative correlation analysis between the top 20 abundant genera or KOs in metaproteomics and their corresponding values in metagenomics. The x-axes show the relative abundance measured by metaproteomic data, and the y-axes are the quantitative value in metagenomic data. The correlation was calculated using Spearman. (D) The correlation of microbial functions between different samples or two samples collected from the same individual. The statistical analysis was performed with Welch's t-test. *** p value < 0.001. (E) The temporal stability of the

1218 microbial genus or KO functions was associated with corresponding richness at the stage of
1219 baseline in metaproteomic data and metagenomic data. The correlation was calculated using
1220 Spearman.

1221

1222 **Figure 3**



1223

1224 **Figure 3. Metaproteomic core species involved in the metabolism of SCFA and vitamin Bs.**
1225 **(A)** Summary of metaproteomic core species enriched with SCFAs (butyrate and propionate), as
1226 well as vitamin Bs (B1, B3, B5, B6, B7, and B9) metabolisms. The dots represent the fold
1227 enrichment value of pathways in each species. The bacteria denoted in black font are consistent
1228 with the reports, while those in red were firstly found in this study. (B) Metaproteomic core

species highly expressing butyric acid metabolism were summarized, with bacteria capable of detecting the involved enzyme labeled next to the reaction. The bacteria denoted in black font are consistent with reported butyric acid-producing species, while those in red were firstly found in this study. (C) The fecal butyric acid content was found to be correlated with the total abundance of core microbes highly expressing butyric acid metabolic pathways identified in metaproteomics. The correlation was calculated using Spearman.

Figure 4

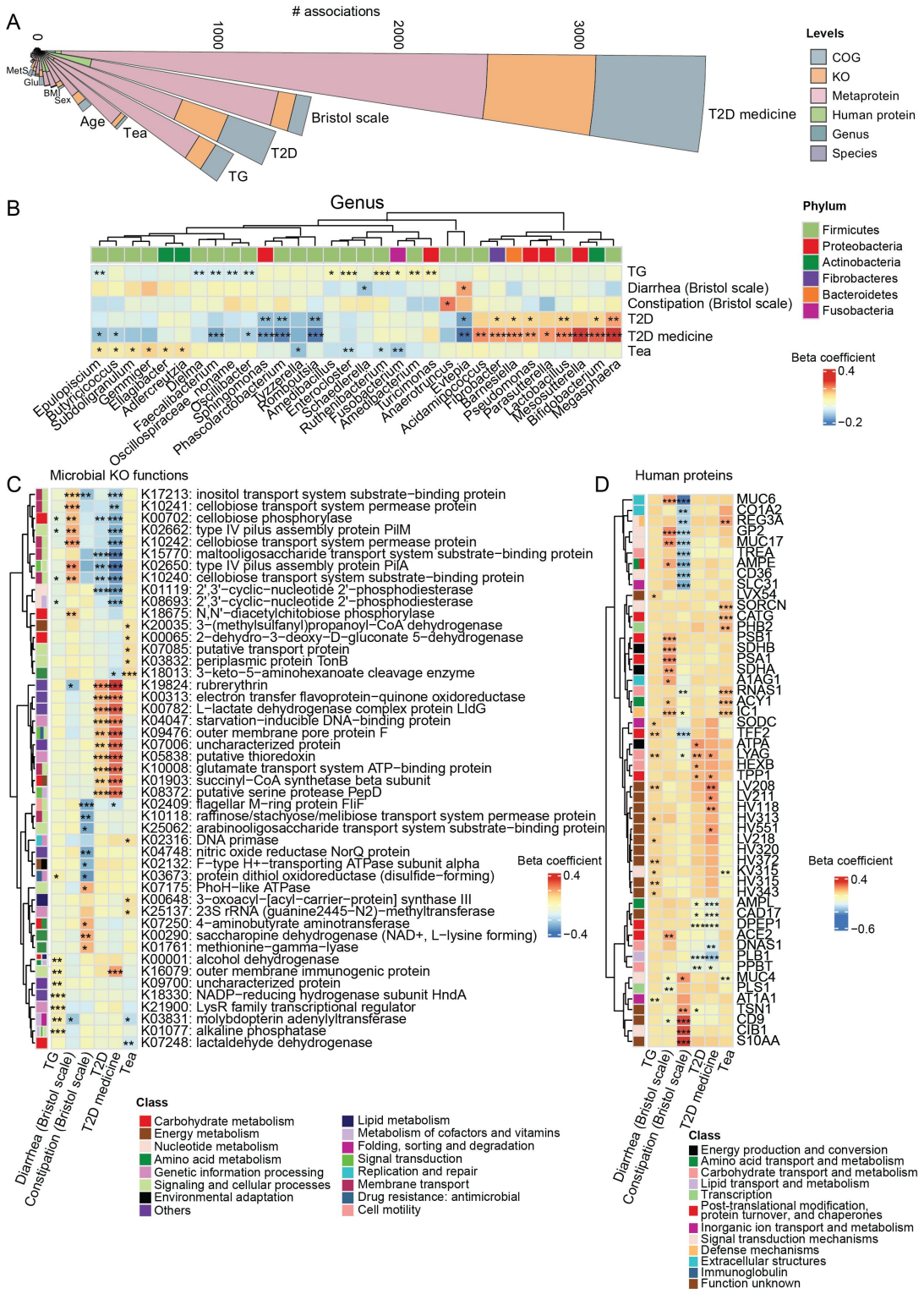
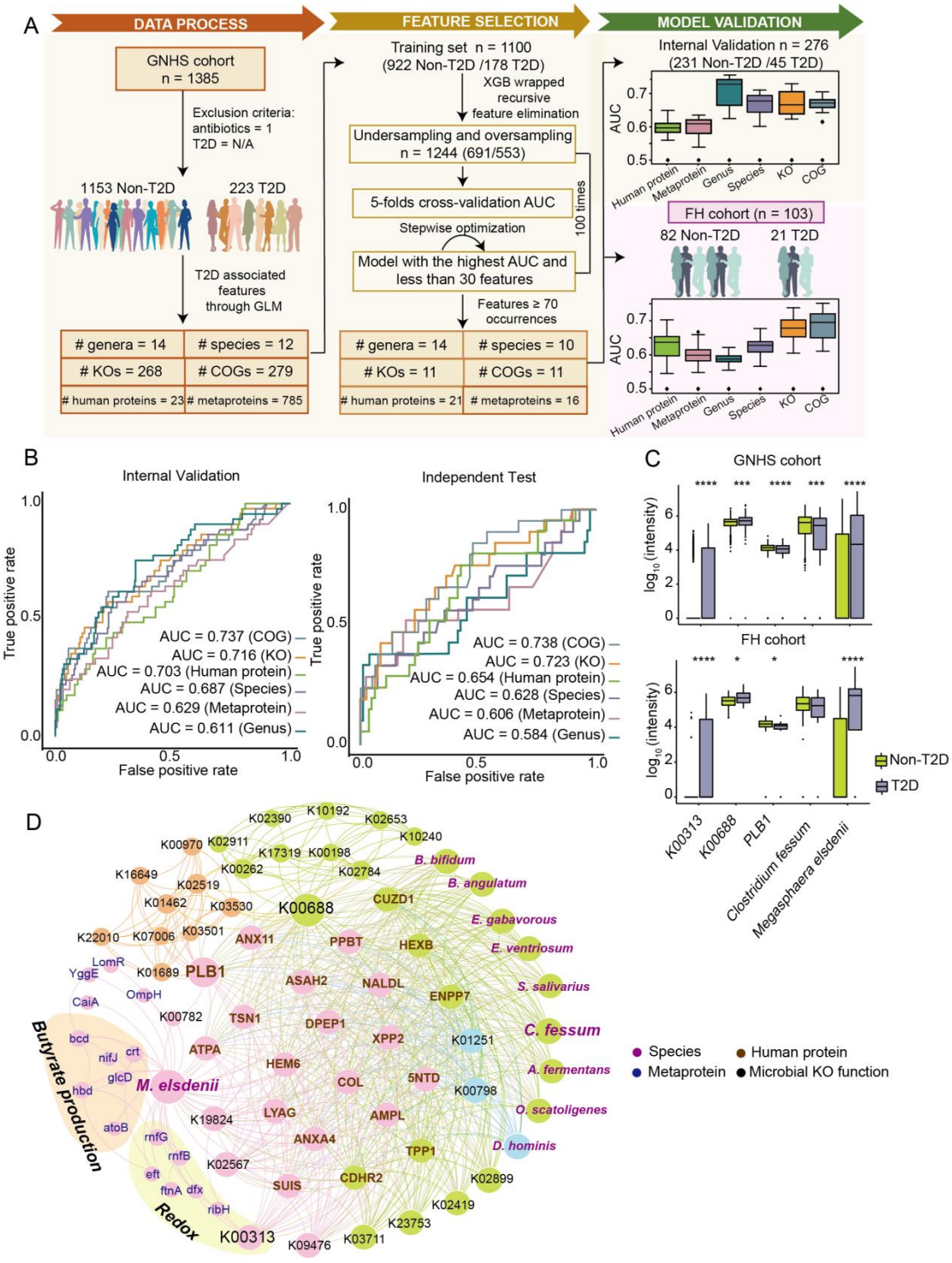


Figure 4. Associations between phenotype and metaproteomic features. (A) Summary of the associations between metaproteomic characteristics and phenotypes. The bars in the graph represent different phenotypes and are sorted according to the number of associations for each level. (B) Metaproteomic genera with the strongest overall correlations with five covariates – Bristol scale, T2D, T2D medicine, TG, and tea. The beta coefficients of each feature and

1242 phenotype were plotted along with their corresponding BH-adjusted p values, where * represents p
1243 < = 0.05, ** < = 0.01, *** < = 0.001. (C) Metaproteomic microbial KO functions with the
1244 strongest overall correlations with covariates. (D) Metaproteomic human proteins with the
1245 strongest overall correlations with covariates.

1246

1247 **Figure 5**

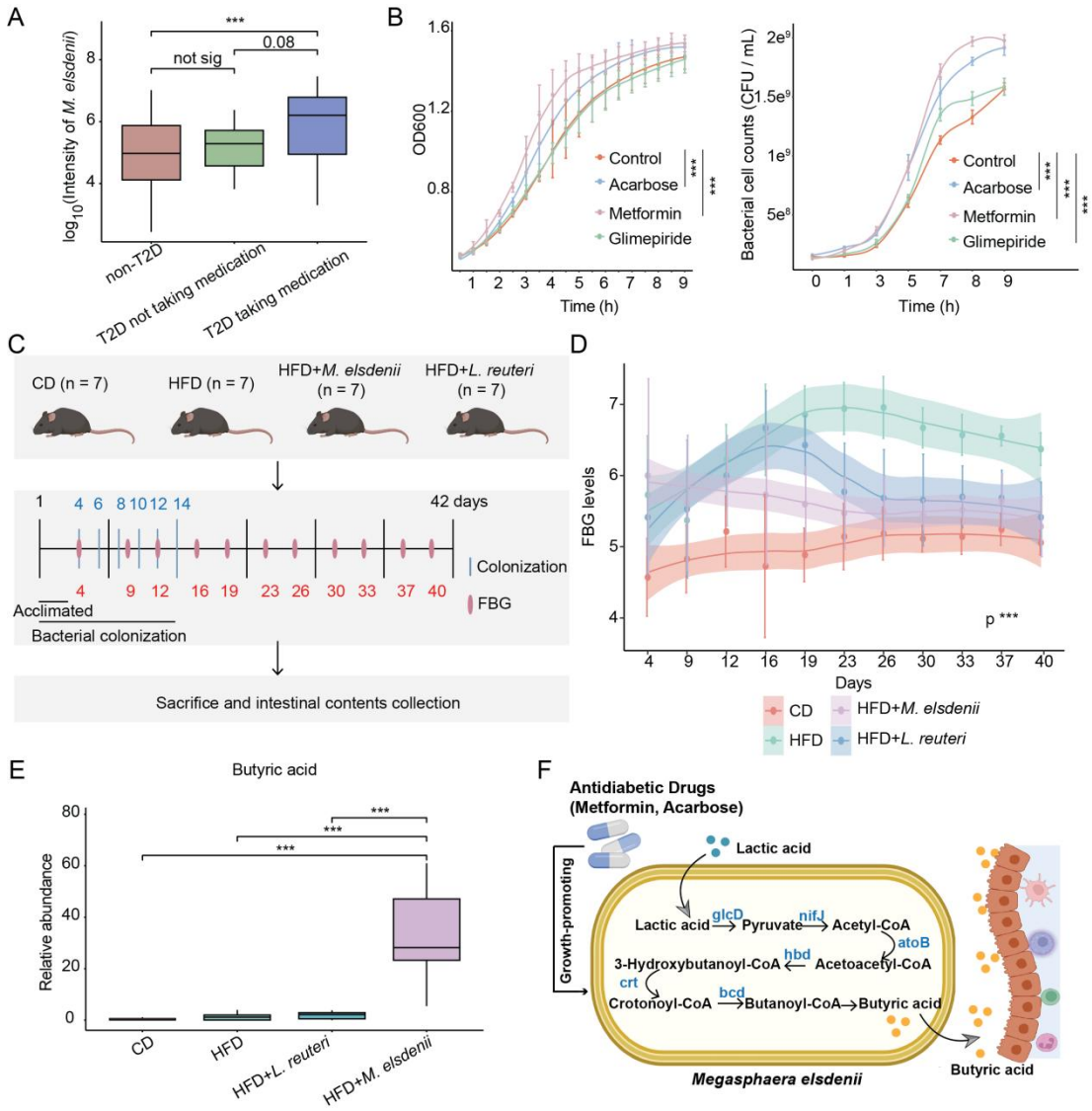


1248

1249 **Figure 5. Metaproteomic functional insights of T2D.** (A) The machine learning workflow for
1250 diabetes classification based on metaproteomic data includes data processing, feature selection,
1251 and model validation. Features with a q-value < 0.05 are filtered using GLM and used as input for
1252 diabetes-related features. Among the 100 machine learning models, features that occur more than
1253 70 times are further selected, and their performance is evaluated in both the internal validation test
1254 and the independent test set at each level. (B) The receiver operating characteristic (ROC) curves

showcasing the best performance for each level in the internal validation cohort and the independent FH cohort. (C) The $\log_{10}$ transformed intensity of the five potential T2D biomarkers in the GNHS cohort and FH cohort. Wilcoxon rank-sum test was used for the differential analyses. * p value < 0.05, *** p value <0.001, **** p value < 0.0001. (D) Networks of the five potential T2D biomarkers and their associated T2D-related features in GNHS cohort. Links between metaproteins and species, as well as KOs and species, were based on their common peptides, while the co-expressed linkages between intra-levels and inter-levels were calculated by the Spearman correlation (BH adjusted p-value < 0.05, $|r| \geq 0.3$ for inter-levels, $|r| \geq 0.8$ for intra-levels).

1264 **Figure 6**



1265

1266 **Figure 6. Biofunction of *M. elsdenii*.** (A) The $\log_{10}$ transformed intensity of M. elsdenii in
1267 non-T2D, T2D not taking medication, and T2D taking medication groups. Wilcoxon rank-sum test
1268 was used for the differential analyses. *** *p* value <0.001. (B) Growth curves of *M. elsdenii* in the
1269 presence or absence of exogenous drugs assessed by measuring the OD600 value of the bacterial
1270 culture and counting the number of live bacterial cells, respectively. Two-way ANOVA with
1271 repeated measurement was applied to calculate the difference between control and drug-treatment
1272 groups. *** *p* value <0.001. (C) Experimental design of germ-free mouse study. Mice were fed
1273 with either a CD or a HFD for 42 days. To investigate the impact of *M. elsdenii* and *L. reuteri* on
1274 blood glucose, mice fed with HFD were monocolonized with these two strains for the initial 14
1275 days of the study. Throughout the experiment, FBG levels were monitored twice weekly. After the
1276 study period, the intestinal contents of the mice were collected to analyze the levels of SCFAs in
1277 the gut. (D) The FBG levels of each mice group during administration. Two-way ANOVA with
1278 repeated measurement was applied to calculate the difference between the HFD and

HFD+FMT(*M. elsdenii*) groups. *** *p* value <0.001. (E) The butyric acid levels in the intestinal contents of each mice group. Wilcoxon rank-sum test was used for the differential analyses. *** *p* value <0.001. (F) The pathogenesis of *M. Eisdenii* in inducing diabetes based on metaproteomic and experimental evidence.