Pairing Metagenomics and Metaproteomics to Characterize Ecological Niches and Metabolic Essentiality of gut microbiomes

Short running title: Proteins' ecological and metabolic roles

Tong Wang[1,†], Leyuan Li[2,3,†], Daniel Figeys[3,*], Yang-Yu Liu[1,4*]

[1]*Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA.*
[2]*China Key Laboratory of Medical Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Lifeomics, Beijing 102206, China.*
[3]*School of Pharmaceutical Sciences, Ottawa Institute of Systems Biology, Faculty of Medicine, University of Ottawa, Ottawa, ON K1H8M5, Canada.*
[4]*Center for Artificial Intelligence and Modeling, the Carl R. Woese Institute of Genomic Biology, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA.*

† These authors contributed equally to this work.
* Correspondence should be addressed to Y.-Y.L. (yyl@channing.harvard.edu) or D.F. (dfigeys@uottawa.ca). The physical mailing address of Y.-Y.L. is: Yang-Yu Liu, Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, 181 Longwood Ave, Boston, MA 02115, USA. The physical mailing address of D.F. is: Daniel Figeys, School of Pharmaceutical Sciences, Ottawa Institute of Systems Biology, Faculty of Medicine, University of Ottawa, 451 Smyth Rd, Ottawa, ON K1H8M5, Canada

Abstract

The genome of a microorganism encodes its potential functions that can be implemented through expressed proteins. It remains elusive how a protein's selective expression depends on its metabolic essentiality to microbial growth or its ability to claim resources as ecological niches. To reveal a protein's metabolic or ecological role, we developed a computational pipeline, which pairs metagenomics and metaproteomics data to quantify each protein's gene-level and protein-level functional redundancy simultaneously. We first illustrated the idea behind the pipeline using simulated data of a consumer-resource model. We then validated it using real data from human and mouse gut microbiome samples. In particular, we analyzed ABC-type transporters and ribosomal proteins, confirming that the metabolic and ecological roles predicted by our pipeline agree well with prior knowledge. Finally, we performed *in vitro*

cultures of a human gut microbiome sample and investigated how oversupplying various sugars involved in ecological niches influences the community structure and protein abundance. The presented results demonstrate the performance of our pipeline in identifying proteins' metabolic and ecological roles, as well as its potential to help us design nutrient interventions to modulate the human microbiome.

Keywords

Metagenomics; Metaproteomics; Functional Redundancy; Ecological Niche; Metabolic Essentiality; Gut Microbiome

Introduction

Metagenomic sequencing has enabled the measurement of the genomic content and functional potential of microbial communities at an unprecedented rate, aiding the understanding of their role in host health[1–3] and biogeochemical cycling[4–6]. While various computational approaches based on these genomes quantify interactions within microbial communities[7–12] and analyze the functional redundancy and functional stability of microbial communities[10–12], they focus on potential rather than actual function, as microorganisms only express a subset of genes as proteins[13]. Recent advancements in high-throughput metaproteomics allow us to quantify protein abundances in human gut microbiomes[14], offering insights into gene expression in response to environmental changes when paired with metagenomic data.

From the metabolic perspective, some genes and their encoded proteins are indispensable for cell metabolism under any conditions, and microbial growth halts without these essential functions—aminoacyl-tRNA synthetase[15,16], ribosomal proteins[17–19], and enzymes involved in glycolysis[20,21]. From the ecological perspective, gene expression is influenced by ecological selection, with specific proteins indicating which resources a microbe can utilize and defining its ecological niche. For instance, *E. coli* prefers glucose over lactose due to the repressed expression of lactose-utilizing enzymes, even though it can use both sugars[22,23]. Such specialization of consuming one resource caused by the selective gene expression may reduce the niche overlap with other species and allow microbial coexistence,

as seen with two *E. coli* strains where one expresses acetyl-coenzyme synthetase (Acs)[24,25] to consume acetate produced by the other[26–29].

Understanding the selective expression of microbial genes is an outstanding question in microbiology. Does the behavior of selective expression of microbial genes differ between metabolic function (e.g., essential for microbial growth metabolism) and ecological function (e.g., claiming resources as a niche)? To answer this, we developed a computational method to analyze paired metagenomic and metaproteomic[30–33,14] data, constructing the gene content network (GCN) or protein content network (PCN) --- a bipartite graph that connects microbial taxa to their genes or expressed proteins, respectively (Figs. 1a&b). For each gene and its encoding protein, we compared its gene-level (or protein-level) functional redundancy (FR), revealing each protein family's metabolic or ecological role. Our method, validated with several gut microbiome data, accurately predicts that ABC-type transporters are related to ecological niches[34–36] , and ribosomal proteins are essential[17–19]. Finally, we performed *in vitro* culture experiments using human gut microbiome samples to investigate how oversupplying sugars involved in ecological niches influence community structure and protein expression.

Materials and Methods

***In-vitro* human gut microbiota culture and metaproteomics.** Three healthy individual microbiota samples were collected and biobanked. The frozen microbiome samples were cultured in our optimized culture medium[37] with or without the presence of different sugars in technical triplicates, and were taken at different times for optical density and metaproteomic analyses. For single-strain samples, proteins were extracted with 4% SDS 8M urea buffer in 100 mM Tris-HCl buffer, followed by precipitation and acetone washing. Proteins were digested with trypsin desalted[38] for LC-MS/MS analysis using an Orbitrap Exploris 480 mass spectrometer. For the cultured microbiomes, an automated process extracted and purified proteins, which were then digested, desalted, and quantified using TMT11plex[39], ensuring mixed representation in labeling to avoid bias. Samples underwent a 2-hour LC gradient and were analyzed by mass spectrometry. More details can be found in Supplemental Methods.

Datasets. Metagenomics data from four individual microbiomes were obtained from the previous MetaPro-IQ study[14,33] (accessible from the NCBI sequence read archive (SRA) under the accession of SRP068619) and the same samples were reanalyzed by an ultra-deep

101 metaproteomics approach[14] via the PRIDE partner repository[40] with the dataset identifier
102 PXD027297). Proteomics dataset of the cultured singles strain samples has been deposited to
103 ProteomeXchange Consortium with the identifier PXD037923. Metaproteomic dataset of the
104 RapidAIM-cultured microbiome samples has been deposited to ProteomeXchange Consortium
105 (identifier PXD037925). The metaproteomic dataset of the mouse gut microbiome comprising
106 twenty gut microbes is derived from a previous study[41] that was deposited to
107 ProteomeXchange Consortium with the dataset identifier PXD009535 and to MassIVE with the
108 dataset identifier MSV000082287.
109
110 **Database search and data processing.** Proteomics database searches used FASTA
111 databases of the individual strains downloaded from NCBI and MaxQuant[42] 1.6.17.0 for
112 analysis, without the label-free quantification. Metaproteomic database searches of cultured
113 microbiome samples were performed using MetaLab V2.2[43], MaxQuant option was used to
114 search the TMT dataset against the IGC database of the human gut microbiome. The resulting
115 data table was normalized using R package MSstatsTMT , and missing values were imputed
116 using R package DreamAI[45]. The "fraction" of each taxon-specific protein is computed by
117 dividing the protein intensity by the sum of the intensities of all proteins assigned to the same
118 taxon. The log2 fold change of each protein is obtained by taking log2 of the ratio between its
119 fraction in the treatment group (with added sugars) and its fraction in the control group (without
120 added sugars).
121
122 **Statistics.** To calculate correlation throughout the study, we used Pearson's correlation
123 coefficient. All statistical tests were performed using standard numerical and scientific
124 computing libraries in the Python programming language (version 3.7.1) and Jupyter Notebook
125 (version 6.1).
126
127

Results

Specialist function, niche function, and essential function

Here, we define three types of functions for protein families that we would like to categorize: (1) "*Specialist function*": specialized by only a few taxa and not widely shared within a community. (2) "*Niche function*": arising from ecological competition, widespread among genomes of numerous taxa but selectively expressed under specific ecological conditions. (3) "*Essential function*": metabolically indispensable for and widely shared by many taxa within a microbial community. We emphasize that our definition is not exhaustive; some proteins may display attributes of multiple categories or not align precisely with any single category.

Using a simple hypothetical example of two competing species (Fig. 1a, b), we demonstrated the three function types: (1) The blue protein is a specialist function since it is solely encoded in the pink species' genome; (2)The red protein belongs to a niche function due to its selective expression by the yellow species even though the protein is encoded in the genomes of both species; (3) The green protein is an essential function because both species need it for biomass synthesis. In coexistence, the pink species specializes in the blue resource, avoiding competition with the yellow species for the red resource.

GCN, PCN, and network degree

We can identify the functional types of proteins in this hypothetical case by comparing the structure of the GCN and PCN (Fig. 1a, b). For example, consider the protein responsible for converting red resource to green metabolite (red broken circle in Fig. 1a, b), its degree in the GCN $\mathrm{k_{GCN}} = 2$, while its degree in the PCN $\mathrm{k_{PCN}} = 1$. This degree reduction is due to distinct ecological niches being occupied by two species when they are cocultured. By contrast, the protein responsible for assimilating critical green metabolites (green broken circle in Fig. 1a, b) into biomass does not show a degree reduction ($\mathrm{k_{GCN}} = \mathrm{k_{PCN}} = 2$) because it is essential for microbial growth. Similarly, since the blue protein is only specialized by the pink species, its $\mathrm{k_{GCN}} = \mathrm{k_{PCN}} = 1$. Thus, three function types occupy different regions in the $\mathrm{k_{GCN}}$ vs. $\mathrm{k_{PCN}}$ plot (Fig. 1c).

Quantifying gene- and protein-level functional redundancy of each gene and its encoded protein

However, the network degree does not consider the significant impact of the microbial taxonomic profile, which details the makeup of a microbial community. This profile is represented by $\boldsymbol{p} = (p_1, \dots, p_N)$, where p_i is the relative abundance of taxon-i and $\sum_{i=1}^{N} p_i = 1$. For a given gene and its encoded protein, we can define its gene-level FR ($\mathrm{FR_g}$) and protein-level FR ($\mathrm{FR_p}$) within this sample as

$$\mathrm{FR_g} = \sum_{i=1}^{N} \sum_{j \neq i}^{N} \left(1 - d_{ij}^{\mathrm{GCN}}\right) p_i p_j, \quad (1)$$

and

$$\mathrm{FR_p} = \sum_{i=1}^{N} \sum_{j \neq i}^{N} \left(1 - d_{ij}^{\mathrm{PCN}}\right) p_i p_j. \quad (2)$$

d_{ij}^{GCN} (or d_{ij}^{PCN}) is the distance between taxon-i and taxon-j based on their genomic capacity to express this gene (or the presence of the protein). For simplicity, we assume d_{ij}^{GCN} is binary, i.e., $d_{ij}^{\mathrm{GCN}} = 0$ if and only if both taxa share the potential to express the gene, and $d_{ij}^{\mathrm{GCN}} = 1$ otherwise. $d_{ij}^{\mathrm{PCN}} = 0$ if and only if both taxa have expressed the protein. Here, we define $\mathrm{FR_g}$ and $\mathrm{FR_p}$ for each protein, different from our previous studies where FR was calculated by including all genes or proteins in a microbial community[12,14].

Comparing $\mathrm{FR_g}$ and $\mathrm{FR_p}$ provides deeper insight into proteins' function types. For the red protein in our hypothetical example, $d_{12}^{\mathrm{GCN}} = 0$ and $d_{12}^{\mathrm{PCN}} = 1$ because both species share the potential to express the gene while only the yellow species have expressed it (Fig. 1a, b). As a result, $\mathrm{FR_g} = 2(1-0)p_1p_2 = 2p_1p_2$ and $\mathrm{FR_p} = 2(1-1)p_1p_2 = 0$ (Fig. 1d, e). Following the same analysis, $\mathrm{FR_g} = \mathrm{FR_p} = 2p_1p_2$ for the green protein, and $\mathrm{FR_g} = \mathrm{FR_p} = 0$ for the blue protein (Fig. 1d, e). Different from composition-independent $\mathrm{k_{GCN}}$ and $\mathrm{k_{PCN}}$, $\mathrm{FR_g}$ and $\mathrm{FR_p}$ take the microbial composition into account and thus are more ecologically meaningful. Notably, a more uneven abundance distribution would lead to smaller $\mathrm{FR_g}$ and $\mathrm{FR_p}$ ($p_1 = p_2 = 0.5$ in Fig. 1d; $p_1 = 0.9$ and $p_2 = 0.1$ in Fig. 1e). The influence of relative abundances on FR can be mitigated by using the normalized FR: nFR = FR / TD, where TD $= 1 - \sum_i p_i^2$ (Supplementary Fig. 1; see Supplementary Information for the definition).

Overview of our computational pipeline

Following the idea of comparing $\mathrm{FR_g}$ with $\mathrm{FR_p}$, we developed a computational pipeline to assign the function types (specialist, niche, or essential) to protein families based on the paired metagenome and metaproteome (Fig. 1f). This pipeline starts with DNA sequences from

188 metagenomes and peptides sequences from metaproteomes. Using the "protein-peptide
189 bridge" approach that maps peptides to their taxonomic origins and protein families (i.e.,
190 orthologous protein clusters), it generates the GCN, PCN, and taxonomic profile, from which
191 we compute $\mathrm{FR_g}$ and $\mathrm{FR_p}$. Details about this approach can be found in the Supplementary
192 Information. Finally, based on the scatterplot of $\mathrm{FR_g}$ vs. $\mathrm{FR_p}$, each protein family is categorized
193 into one of the three function types. Note that the computational pipeline assigns the function
194 type without leveraging the known biological functions. Instead, we validate these assignments
195 against the knowledge about biological functions.
196

197 ## Illustration of our computational pipeline using synthetic data

198 To illustrate the pipeline's workflow, we utilized synthetic data generated by a consumer-
199 resource model (CRM). Each niche (or specialist) function is modeled as the consumption of
200 a unique and externally supplied resource (Fig. 2a1), whose loss would make a species unable
201 to consume the corresponding resource (Fig. 2a2, a3). The loss of an essential function is
202 modeled as reducing a species' growth rate by 5% (Fig. 2a4).
203 For each species, each niche (specialist, or essential) function was assigned to the
204 species' genome with probability p_n (p_s, or p_e), respectively (Fig. 2b, left). We set $p_\mathrm{n} = p_\mathrm{e} = 0.7$
205 to ensure that we cannot distinguish niche functions from essential functions only based on
206 their $\mathrm{k_{GCN}}$. We set $p_\mathrm{s} = 0.2 < p_\mathrm{n} = p_\mathrm{e}$ so that specialist functions were assigned to fewer
207 species than niche and essential functions. Species' actual expressed functions were
208 determined by randomly sampling a subset of its potential functions (Fig. 2b, middle). This
209 behavior of sub-sampling was observed when we cultured single microbial strains in different
210 environments (Supplementary Fig. 21). We simulated community dynamics until reaching a
211 steady state, for which we constructed the PCN of the surviving species (Fig. 2b, right; see
212 Supplementary information for technical details). More technical details of CRM are in
213 Supplementary information.
214 In our model with 10,000 species and 20 functions for each of the three function types,
215 each species randomly sampled a subset of potential functions (Fig. 2c, left) to express (Fig.
216 2c, middle). We demonstrated a simulation example with 35 species surviving in the final
217 steady state after the community assembly initialized with 10,000 species (Fig. 2c, right).
218 We applied the taxonomic profile, GCN, and PCN for the surviving 35 species to our
219 computational pipeline, finding that the three modeled protein function types were correctly
220 classified as three clusters (60 out of 60 were correct) by the Gaussian mixture model in both

221 the comparison of network degree (Fig. 2d) and FR (Fig. 2e). We emphasize that the observed
222 three functional clusters arise from community assembly. When we randomly picked 35
223 species (same as the number of surviving species) from the initial pool with equal abundances
224 without assembly, niche functions cannot be distinguished from essential functions (Fig. 2f, g).
225 Even when assigning the same randomly picked 35 species with the same abundances of
226 surviving species, we still cannot differentiate these two functional types (Supplementary Fig.
227 2). Our findings held when varying (1) the number of species and functions or (2) model
228 parameters p_{n}, p_{s}, and p_{e}, affirming the method's robustness in distinguishing function types
229 (Supplementary Figs. 3, 4, and 5).
230

231 ## Three protein functional clusters observed in human gut microbiomes

232 Next, we validated our computational pipeline on real data of human mucosal-luminal interface
233 samples previously collected from the ascending colon of four children[14,33]. Here we focused
234 on the genus level and annotated the identified proteins from metagenomic and metaproteomic
235 data via the clusters of orthologous genes (COGs) database[46,47]. We chose the genus level
236 due to widely shared peptide sequences across species (Supplementary Fig. 6). We searched
237 metagenomic reads and metaproteomic peptides against the integrated gene catalog (IGC)
238 database of the human gut microbiome[48] to generate the GCN and PCN[14], and took the
239 intersected COGs between the two networks. Taxonomic assignment was performed using the
240 "protein-peptide bridge" method as described previously[14]. Our analysis centers on subject
241 HM454, identifying 1,542 intersected COGs in both the GCN and PCN and obtaining a
242 taxonomic profile of 85 genera using MetaPhlAn2[49]. The connectance (i.e., the number of
243 edges divided by the maximal number of possible edges) of the GCN (or PCN) is 0.220 (or
244 0.049), respectively (Fig. 3a, b). The GCN displayed a higher nestedness (nestedness metric
245 NODF[50]=0.667) than the PCN (NODF=0.453). More details about data processing and NODF
246 are in Supplementary Information.
247 The network degree analysis revealed a general decline from GCN to PCN, with 804
248 out of 1,542 COGs having $\mathrm{k_{PCN}} < 0.2\mathrm{k_{GCN}}$ (Fig. 3c). This decline greatly influences FR but
249 does not fully explain why many COGs have $\mathrm{FR_p}\sim 0$ (744 out of 1,542 have $\mathrm{FR_p} < 0.01$ in Fig.
250 3d) and k_{PCN} nonlinearly correlates with $\mathrm{FR_p}$ (Fig. 3e). For example, for L-arabinose isomerase
251 (COG2160), its k_{PCN} (7) is fairly close to k_{GCN} (8), but its $\mathrm{FR_p}$ (0.04) is much lower than $\mathrm{FR_g}$
252 (0.23) since the genus *Blautia* (relative abundance = 22%) did not express L-arabinose
253 isomerase, even if it has this capacity encoded in its genome.

254 Using the Gaussian mixture model fitted on simulated data, we categorized all protein
255 families into three clusters (Figs. 3c&d). Although clusters on real data are not as distinct as
256 on simulated data, the relative positioning of the three clusters (shaded areas in Fig. 3c, d)
257 agrees well with our hypothesis (Fig. 1). The weaker clustering might result from a greater
258 variation in k_{GCN} (or $\mathrm{FR_g}$) for real data (Fig. 3c, d) than that for simulated data (Fig. 2d, e).

259 Some COGs have $\mathrm{FR_p} > \mathrm{FR_g}$ (Fig. 3c, d), contradicting the sub-sampling argument for
260 the gene expression. $\mathrm{FR_p}$ should not exceed $\mathrm{FR_g}$ if the PCN was a proper subgraph of the
261 GCN. This contradiction may stem from limitations in metagenomic sequencing and
262 metaproteomic identification depths, as both metagenomics and metaproteomics require
263 sufficient depth to detect genes or proteins, respectively. We tested how the lower detection
264 capability of metaproteomics or metagenomics influences the FR by varying the protein or
265 gene abundance percentile threshold (PAPT or GAPT), which denotes the percentage of most
266 abundant proteins or genes being kept. As PAPT or GAPT decreases, $\mathrm{FR_p}$ or $\mathrm{FR_g}$ drops
267 respectively (Supplementary Fig. 15). When GAPT decreases, we observed more proteins
268 with $\mathrm{FR_p}$ greater than their $\mathrm{FR_g}$.
269

270 ## Validating three functional clusters observed in human gut microbiomes

271 Our computational pipeline accurately assigns functional clusters for protein families, agreeing
272 with their known biological functions. For example, COG0539 (ribosomal protein S1) was
273 assigned as the essential function, which is essential for translational initiation[17–19,51,52]. Another
274 example is the assignment of COG1116 (ABC-type nitrate/sulfonate/bicarbonate transport
275 system)[34] as a niche function, whose expression has been shown to be selectively enriched
276 for a few microbial species[53].

277 The pipeline's classifications were systematically validated against well-established
278 biological roles of specific protein families: (1) ABC-type transporters are niche proteins due to
279 their connection with ecological metabolic niches[34–36]; (2) Ribosomal proteins are essential
280 proteins because they are indispensable for the microbial growth[54,55]; (3) PTS
281 (phosphotransferase system) proteins are specialist proteins because an evolutionary study
282 has shown that various species within the same genus even possess a different set of PTS
283 proteins[56]. To evaluate our pipeline's performance, we quantified its accuracy in assigning
284 these protein families (ABC-type transporters, ribosomal proteins, or PTS proteins) against the
285 assumed "ground-truth" functions (niche, essential, or specialist functions, respectively).

286 For HM454, our computational pipeline based on the FR_g/FR_p plot correctly categorizes
287 81 of 122 COGs belonging to ABC-type transporters, ribosomal proteins, or PTS proteins. In
288 comparison, when classifying functions based on the k_{GCN}/k_{GCN} plot, 74 COGs were correctly
289 assigned, slightly worse than that based on the FR_g/FR_p plot. Specifically, 26 of 53 COGs
290 belonging to ABC-type transporters are classified as niche functions. The fraction of ribosomal
291 proteins classified to the cluster of essential functions is 83.0%(=44/53). For the PTS proteins,
292 among the identified 16 COGs, 11 are classified as specialist functions.
293

294 ## Alternative clustering and classification methods

295 Alternatively, we explored the unsupervised K-mean clustering with K=3, which captured the
296 three representative functional clusters with their positions agreeing with our expectations
297 (Supplementary Fig. 13). We also designed a supervised classifier based on quadratic
298 discriminant analysis (QDA). QDA, trained on ABC-type transporters, PTS proteins, and
299 ribosomal proteins as the niche, specialist, and essential functions, generated clusters closely
300 resembling those from the Gaussian mixture model (Supplementary Fig. 14). For HM454, the
301 K-mean clustering categorizes 48 of these 122 COGs that are ABC-type transporters,
302 ribosomal proteins, or PTS proteins into clusters respectively representing niche, essential, or
303 specialist functions (i.e., the accuracy is 39.3%). For the QDA classifier, the accuracy is
304 59.0%(=72/122). Thus, we select the Gaussian mixture model as the classification method
305 because of its superior accuracy (66.4%=81/122).
306

307 ## Comparing FR_g with FR_p identifies ecological niches and metabolic essentiality

308 We focused on analyzing ABC-type transporters[34–36] and ribosomal proteins[17–19]. ABC-type
309 transporters are energy-requiring transporter proteins that allow microbes to exploit specific
310 niches like glucose uptake[34–36]. For HM454, we indeed found that k_{GCN} for all ABC-type
311 transporters are much larger than their k_{PCN} (Fig. 4a). Similarly, we also found that their FR_g
312 are much larger than their FR_p, classifying many transporter proteins as niche functions (Fig.
313 4b). Some transporter proteins were classified as specialist functions (blue dots in Fig. 4b) due
314 to the specialization on the gene level, which is carried to the protein level. Some transporter
315 proteins were classified as essential functions (green dots in Fig. 4b). One example is the ABC-
316 type Fe^{3+}/spermidine/putrescine transporter (COG3842), as iron is essential for bacteria to
317 function as a co-factor in iron-containing proteins[57,58].

318 Ribosomal proteins, critical for protein synthesis and microbial growth[54,55], showed little
319 variance between k_{GCN} and k_{PCN} (the mean and the standard deviation of the relative difference
320 $\frac{k_{GCN}-k_{PCN}}{k_{GCN}}$ is 0.09 ± 0.41; Fig. 4e), with most correctly classified as essential (44 out 53 COGs
321 in Fig. 4f). Notably, two ribosomal proteins (L28 and L34), which have been reported as non-
322 essential to microbes such as *E. coli*[17,52,59], were accurately classified as non-essential proteins
323 (red dots in Fig. 4e). Certain specialized ribosomal proteins in microbial genomes continue to
324 be specialized on the protein level and thus were classified as specialist functions.
325 Alternatively, we looked at the distribution of network degrees (Figs. 4c&g) and FR
326 (Figs. 4d&h). For ABC-type transporters, the distribution of k_{PCN} is close to 0 (median of 2),
327 while the median of k_{GCN} is 25. For ribosomal proteins, the distribution of k_{PCN} (median is 12) is
328 similar to k_{GCN} (median is 14). For ABC-type transporters, the distribution of $\mathrm{FR_p}$ is close to 0
329 (with a median ~ 0.01), while the median of $\mathrm{FR_g}$ is around 0.30. For ribosomal proteins, the
330 distribution of $\mathrm{FR_p}$ (median ~ 0.20) is similar to the distribution of $\mathrm{FR_g}$ (median ~ 0.21).
331 The similar patterns are also true for the other three individuals (Supplementary Figs.
332 8-10). Variations in the FR_g /FR_p plot across individuals (Supplementary Fig. 8-10) are likely
333 due to differences in gut environments, diets, and microbial composition. Note that HM503 is
334 an outlier due to its lower diversity (36 genera versus the average of 56.7) (Supplementary Fig.
335 11). Similarly, the Shannon diversity index for HM503 is only 1.41, lower than other individuals
336 (mean and standard deviation are 2.05 and 0.18). The lower diversity in HM503 leads to fewer
337 taxa owning the same function on average, resulting in lower FR_g and FR_p values.
338 Extending our analysis to additional protein families, we discovered that proteins linked
339 to glycolysis, RNA polymerase, and the TCA cycle displayed patterns similar to ribosomal
340 proteins (Figs. 4i&j). By contrast, proteins associated with sugar utilization, glycolyl hydrolase,
341 and aromatic amino acid biosynthesis exhibited patterns more akin to ABC-type transporters.
342 We also analyzed the classification results for the unknown COG functions (with the category
343 'S: Function Unknown'). Out of the 104 unknown COGs in our analysis, the majority are
344 classified as specialist (43/104) or niche functions (40/104), with a smaller portion identified as
345 essential functions (21/104).
346 We also confirmed our results using the KEGG Orthology (KO) annotation[60–63], which
347 has a lower annotation rate (78%) than COG (92%). For HM454, our computational pipeline
348 categorizes 75 of these 126 KOs that are ABC-type transporters, ribosomal proteins, or PTS
349 proteins into clusters respectively representing niche, essential, or specialist functions, similar
350 to the classification accuracy based on the COG. The contrasting difference between ABC-

351 type transporters and ribosomal proteins is well preserved (see Supplementary Fig. 7).
352 Additionally, the distribution of $\mathrm{FR_p}$ shows a dramatic difference across KO groups
353 (Supplementary Fig. 12). Some ecologically strongly selected KO groups such as ABC
354 transporters have small $\mathrm{FR_p}$ (Supplementary Fig. 12). As a comparison, proteins from
355 aminoacyl-tRNA biosynthesis[15,16], glycolysis[20,21], and ribosomes[17–19] have large $\mathrm{FR_p}$ and huge
356 variations within each group (Supplementary Fig. 12).
357

358 ## Validating our method on the mouse gut microbiome

359 In testing our method's feasibility in other microbial communities, we leveraged a
360 metaproteomic dataset from mice gavaged with a synthetic microbiome comprising twenty
361 sequenced bacteria[41]. Since this study lacks paired metagenomes, we used its 16S rRNA gene
362 sequencing data with individual genomes to infer its metagenome. Here we focused on the
363 strain level because peptides of different strains in this simple synthetic gut microbiome can
364 be distinguished. We relied on the comparison between $\mathrm{FR_g}$ and $\mathrm{FR_p}$ to generate the
365 distribution of functional clusters across many protein families for this dataset (Fig. 5n). The
366 results mirrored those of human gut microbiomes, especially the contrasting patterns between
367 ABC-type transporters and ribosomal proteins (Fig. 5).
368

369 ## Response of community and protein abundance to the introduction of sugars

370 After identifying niche functions through our computational pipeline, we explored using
371 nutrients associated with niche functions to manipulate the community structure. In ecology, a
372 niche is often defined as an abiotic and biotic factor that supports the survival of species[9,64–66].
373 Therefore, niche functions are associated with corresponding limiting resources involved in
374 those functions. For example, COG1879 (ABC-type sugar transporter) is categorized as a
375 niche function due to microbial competition for sugars. Here, we leveraged the *in vitro*
376 community and studied how expression levels of ATP-type transporters respond to supplied
377 sugars so that a microbial taxon can achieve a better living strategy.
378 Using the RapidAIM V2.0 approach[67], which replicates the functional profiles of
379 individual gut microbiomes *in vitro*[37], we cultured three individual human gut microbiota
380 samples and used a semi-automated metaproteomics workflow to observe how taxon-specific
381 proteins respond to the presence of glucose, fructose, and kestose (Fig. 6a). Samples were
382 cultured in technical triplicates, and protein abundances were quantified at 0, 1, 5, 12, and 24
383 hours using 11-plex tandem mass tag (TMT11plex)[39] for a total of 189 samples. We analyzed

the Bray–Curtis dissimilarity of metaproteomes over time, finding that more complex sugars induce more pronounced alterations in protein profiles (Supplementary Fig. 16).

To reflect the effect of introduced sugars on protein abundances, we used log2 of fold change in normalized protein abundances/intensities (see Supplemental Methods for details) between the treatment and control group (Fig. 6). We hypothesized that excessive sugars remove the growth limitation on carbon resources, prompting microbes to upregulate other transporters for uptaking more other scarce resources (e.g., nitrogen or amino acids) for better growth (Fig. 6a). We analyzed log2 fold changes of ABC-type transporters 5 hours later after sugar introduction (Figs. 6b-d), with most COGs close to zero. Among seven significantly influenced COGs, COG1126 (ABC-type polar amino acid transport system) is the only one that is revealed to be a niche function. Focusing on COG1126, we found that it is specialized by the genus *Holdemanella*. *Holdemanella* benefits from upregulating COG1126, as the proportion of *Holdemanella* proteins significantly increases from 13.5%(± 0.06%) for the control to 15.8 %(± 0.08%) with the added glucose (p-value = 0.04, Mann-Whitney U test applied).

We observed that adding fructose, glucose and fructose, or kestose alters ABC-type transporters' expression similarly to glucose alone (Fig. 6). The correlation in log2 fold changes of ABC-type transporters between different added sugars is significant (Supplementary Fig. 18). Notably, complex sugars trigger more significant fold changes in microbial protein expression. This pattern persists for metaproteomic measurements 12 and 24 hours later, while the fold changes 1 hour later are less significant (Supplemental Fig. 19; p-value < 0.01 for four sugar-adding scenarios, Mann-Whitney U test applied). Additionally, the overwhelmingly positive log2 fold changes of ribosomal proteins (Supplementary Fig. 20; p-value < 10^{-4} for four sugar-adding scenarios, one-sample Wilcoxon test applied) probably implies faster microbial growth when simple sugars are supplied[68,69].

Discussion

We developed a computational pipeline to classify protein families as specialist, niche, and essential functions by comparing $\mathrm{FR_g}$ with $\mathrm{FR_p}$. This approach supplements traditional methods that test metabolic essentiality by gene knockout[17–19] and identify limiting resources by measuring biomass changes upon resource supplies[70–73]. We first illustrated this method on synthetic data and then validated it using real datasets of human and mouse gut microbiomes. We acknowledge a limitation in our validation process—the reliance on limited available

417 literature. Hence, our classification should be seen as a preliminary framework that is open to
418 refinement as the investigation of protein families' functions improves.

419 Our findings bridge the gap between the ecological niche theory, which posits that each
420 resource (or niche) can only be occupied by one species for steady-state conditions[66,74,75], and
421 the FR revealed by shared functions among microbial genomes[11,12]. We solved this dilemma
422 by showing niche proteins usually have very small $\mathrm{FR_p}$ and large $\mathrm{FR_g}$. Additionally, our
423 ecological framework combines genomic capacity and protein functions together by introducing
424 species with sub-sampled functions. The model framework accounts for selective expression
425 due to different environmental conditions[76–78] or evolved strains with distinct metabolic
426 niches[79,26,27], reconciling phenotype-focused ecological models with genetic data.

427 The observed case of $FR_p > FR_g$ for some COGs could stem from using MetaProIQ, a
428 general microbiome catalog, for metaproteome analysis. Although MetaProIQ facilitates the
429 identification of proteins from various gut microbes, the general search against it may cause
430 anomalies. In contrast, directly searching metaproteomic data against gene calls from the
431 paired metagenome may deliver more accurate identifications. However, this strategy suffers
432 when the metagenomic sequencing is incomplete, leading to undetected proteins due to
433 missing genes. In our human gut microbiome datasets, limited sequencing depth might cause
434 such issues. For the synthetic mouse gut communities with complete genomes for each
435 microbial strain, we matched metaproteome to microbial genomes and did not encounter any
436 case of $FR_p > FR_g$, highlighting the effectiveness of this approach in contexts with complete
437 genomic data.

438 In this work, we only validated our pipeline on gut-related biomes due to the limited
439 accessibility of paired metagenome-metaproteome in other environments. In nutrient-poor
440 environments, where ribosomal proteins are less expressed[80,81], essential proteins like
441 ribosomes may be harder to detect, causing potential detection biases. This aspect warrants
442 further investigation to ensure the robustness and applicability of our method in diverse
443 ecological settings. Other technical limitations can also impact clustering accuracy. Smaller
444 ribosomal proteins L28 and L34 could be detected less frequently in metaproteomics, and post-
445 translational modifications may also result in missed cleavage and identification of peptides.
446 Advances such as metaproteomics-assembled proteomes (MAPs)[82] may improve taxon-
447 specific functional annotations and the accuracy of our clustering outcomes.
448
449

Data and code availability. All code for simulations used in this manuscript can be found at https://github.com/wt1005203/ecological_niches.

Acknowledgements. We thank Dr. Janice Mayne for the help with providing biobanking samples, Dr. Zhibin Ning for the help with running mass spectrometry and Dr. Joeselle Serrana for the help with processing the metagenomic data. D.F. acknowledges grants from Natural Sciences and Engineering Research Council of Canada (NSERC), and the Government of Canada through Genome Canada and the Ontario Genomics Institute (OGI-114 & OGI-149). D.F. acknowledges a Distinguished Research Chair from the University of Ottawa. Y.-Y.L. is supported by grants R01AI141529, R01HD093761, RF1AG067744, UH3OD023268, U19AI095219 and U01HL089856 from the National Institutes of Health, USA; a pilot grant from the Biology of Trauma Initiative of Broad Institute, USA; and the Office of the Assistant Secretary of Defense for Health Affairs, through the Traumatic Brain Injury and Psychological Health Research Program (Focused Program Award) under award no. (W81XWH-22-S-TBIPH2), endorsed by the Department of Defense, USA.

Author contributions. Y.-Y.L. and D.F. supervised the study. T.W. and Y.-Y.L. conceived the project. All authors designed the research. L.L. prepared and curated the empirical data as well as performed all wet-lab experiments. T.W. analyzed all data and developed the ecological model. T.W. wrote the initial manuscript. All authors edited and approved the manuscript.

Competing Interests. The authors declare no competing interests. D.F. co-founded MedBiome Inc., a clinical microbiomics company.

IRB determination/approval statement. The sample collection of human gut microbiomes was approved by the Research Ethics Board of the Children's Hospital of Eastern Ontario (CHEO), Ottawa, ON, Canada. The written informed consent forms were obtained from their parents.

474 **References**

475 1. Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P.
476 M. *et al.* Community structure and metabolism through reconstruction of microbial genomes
477 from the environment. *Nature* **428**, 37–43 (2004).
478 2. Flint, H. J., Scott, K. P., Louis, P. & Duncan, S. H. The role of the gut microbiota in
479 nutrition and health. *Nature Reviews Gastroenterology & Hepatology* **9**, 577–589 (2012).
480 3. Lloyd-Price, J., Arze, C., Ananthakrishnan, A. N., Schirmer, M., Avila-Pacheco, J.,
481 Poon, T. W. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel
482 diseases. *Nature* **569**, 655–662 (2019).
483 4. Paerl, H. W. & Pinckney, J. L. A Mini-review of Microbial Consortia: Their Roles in
484 Aquatic Production and Biogeochemical Cycling. *Microb Ecol* **31**, 225–247 (1996).
485 5. Falkowski, P. G., Fenchel, T. & Delong, E. F. The Microbial Engines That Drive
486 Earth's Biogeochemical Cycles. *Science* **320**, 1034–1039 (2008).
487 6. Louca, S., Parfrey, L. W. & Doebeli, M. Decoupling function and taxonomy in the
488 global ocean microbiome. *Science* **353**, 1272–1277 (2016).
489 7. Levy, R. & Borenstein, E. Metabolic modeling of species interaction in the human
490 microbiome elucidates community-level assembly rules. *PNAS* **110**, 12804–12809 (2013).
491 8. Pacheco, A. R., Moel, M. & Segrè, D. Costless metabolic secretions as drivers of
492 interspecies interactions in microbial ecosystems. *Nature Communications* **10**, 103 (2019).
493 9. Fahimipour, A. K. & Gross, T. Mapping the bacterial metabolic niche space. *Nature*
494 *Communications* **11**, 4887 (2020).
495 10. Louca, S., Jacques, S. M., Pires, A. P., Leal, J. S., Srivastava, D. S., Parfrey, L. W. *et*
496 *al.* High taxonomic variability despite stable functional structure across microbial
497 communities. *Nat Ecol Evol* **1**, 15 (2016).
498 11. Louca, S., Polz, M. F., Mazel, F., Albright, M. B., Huber, J. A., O'Connor, M. I. *et al.*
499 Function and functional redundancy in microbial systems. *Nature Ecology & Evolution* **2**, 936
500 (2018).
501 12. Tian, L., Wang, X. W., Wu, A. K., Fan, Y., Friedman, J., Dahlin, A. *et al.* Deciphering
502 functional redundancy in the human microbiome. *Nature Communications* **11**, 6217 (2020).
503 13. Franzosa, E. A., Morgan, X. C., Segata, N., Waldron, L., Reyes, J., Earl, A. M. *et al.*
504 Relating the metatranscriptome and metagenome of the human gut. *PNAS* **111**, E2329-2338
505 (2014).
506 14. Li, L., Wang, T., Ning, Z., Zhang, X., Butcher, J., Serrana, J. M. *et al.* Revealing
507 proteome-level functional redundancy in the human gut microbiome using ultra-deep
508 metaproteomics. *Nat Commun* **14**, 3428 (2023).
509 15. Ibba, M. & Soll, D. Aminoacyl-tRNA synthesis. *Annu Rev Biochem* **69**, 617–650
510 (2000).
511 16. Parker, D. J., Lalanne, J. B., Kimura, S., Johnson, G. E., Waldor, M. K., & Li, G. W.
512 Growth-Optimized Aminoacyl-tRNA Synthetase Levels Prevent Maximal tRNA Charging.
513 *Cell Syst* **11**, 121-130.e6 (2020).
514 17. Shoji, S., Dambacher, C. M., Shajani, Z., Williamson, J. R. & Schultz, P. G.
515 Systematic chromosomal deletion of bacterial ribosomal protein genes. *J Mol Biol* **413**, 751–
516 761 (2011).
517 18. Dabbs, E. R. Mutants lacking individual ribosomal proteins as a tool to investigate
518 ribosomal properties. *Biochimie* **73**, 639–645 (1991).

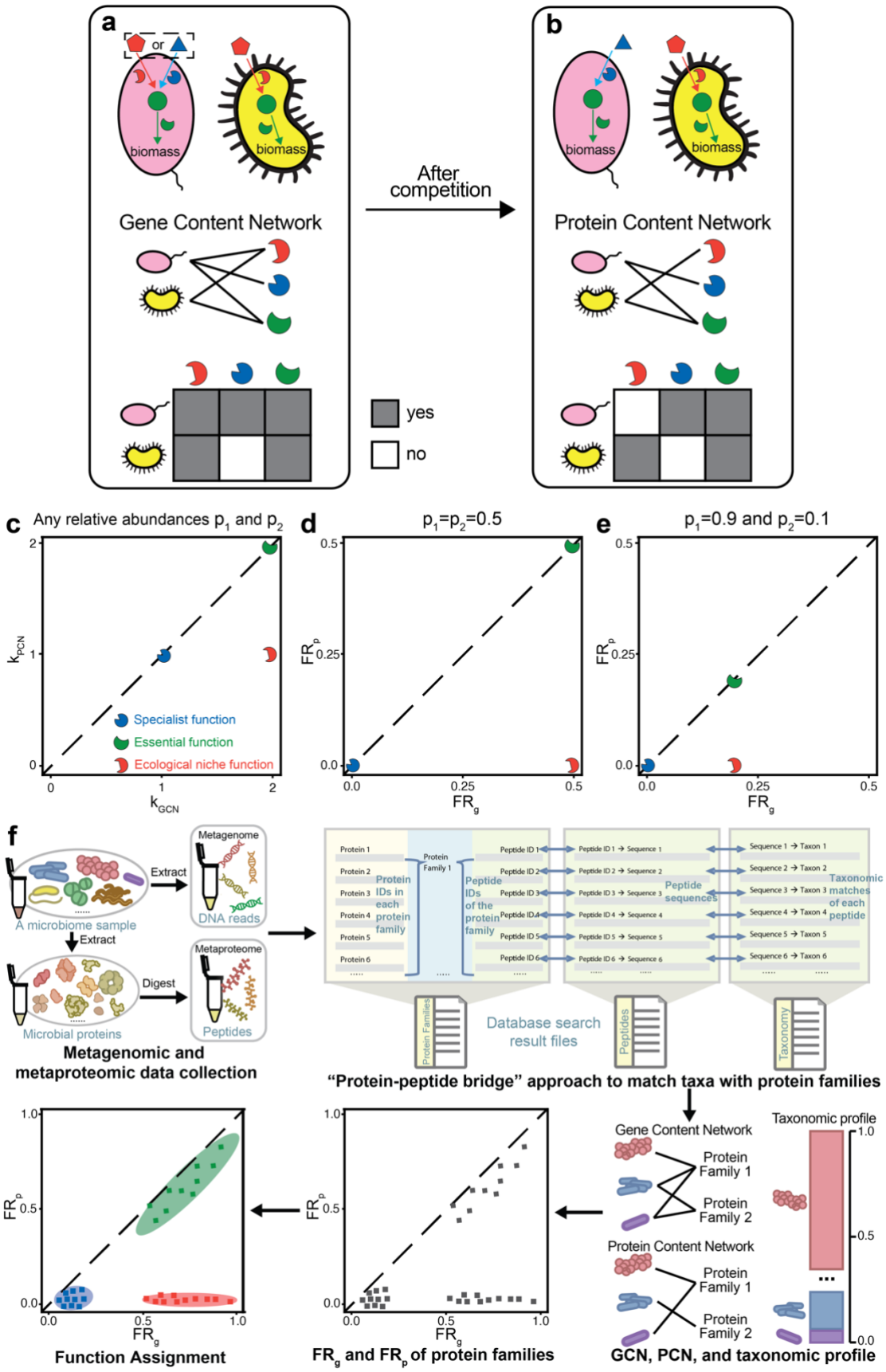
519 19. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M. *et al.*
520 Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio
521 collection. *Mol Syst Biol* **2**, 2006.0008 (2006).
522 20. Romano, A. H. & Conway, T. Evolution of carbohydrate metabolic pathways.
523 *Research in Microbiology* **147**, 448–455 (1996).
524 21. Yan, Y. *Engineering Microbial Metabolism For Chemical Synthesis: Reviews And*
525 *Perspectives*. (World Scientific, 2017).
526 22. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J*
527 *Mol Biol* **3**, 318–356 (1961).
528 23. Okano, H., Hermsen, R., Kochanowski, K. & Hwa, T. Regulation underlying
529 hierarchical and simultaneous utilization of carbon substrates by flux sensors in Escherichia
530 coli. *Nat Microbiol* **5**, 206–215 (2020).
531 24. Kumari, S., Beatty, C. M., Browning, D. F., Busby, S. J., Simel, E. J., Hovel-Miner,
532 G., & Wolfe, A. J. Regulation of Acetyl Coenzyme A Synthetase inEscherichia coli. *Journal*
533 *of Bacteriology* **182**, 4173–4179 (2000).
534 25. Starai, V. J. & Escalante-Semerena, J. C. Acetyl-coenzyme A synthetase (AMP
535 forming). *Cell Mol Life Sci* **61**, 2020–2030 (2004).
536 26. Rosenzweig, R. F., Sharp, R. R., Treves, D. S. & Adams, J. Microbial evolution in a
537 simple unstructured environment: genetic differentiation in Escherichia coli. *Genetics* **137**,
538 903–917 (1994).
539 27. Treves, D. S., Manning, S. & Adams, J. Repeated evolution of an acetate-crossfeeding
540 polymorphism in long-term populations of Escherichia coli. *Mol Biol Evol* **15**, 789–797
541 (1998).
542 28. Lin, H., Castro, N. M., Bennett, G. N. & San, K.-Y. Acetyl-CoA synthetase
543 overexpression in Escherichia coli demonstrates more efficient acetate assimilation and lower
544 acetate accumulation: a potential tool in metabolic engineering. *Appl Microbiol Biotechnol*
545 **71**, 870–874 (2006).
546 29. Kinnersley, M. A., Holben, W. E. & Rosenzweig, F. E Unibus Plurum: genomic
547 analysis of an experimentally evolved polymorphism in Escherichia coli. *PLoS Genet* **5**,
548 e1000713 (2009).
549 30. Penzlin, A., Lindner, M. S., Doellinger, J., Dabrowski, P. W., Nitsche, A., & Renard,
550 B. Y. Pipasic: similarity and expression correction for strain-level identification and
551 quantification in metaproteomics. *Bioinformatics* **30**, i149-156 (2014).
552 31. Ram, R. J., VerBerkmoes, N. C., Thelen, M. P., Tyson, G. W., Baker, B. J., Blake, R.
553 C. *et al.* Community proteomics of a natural microbial biofilm. *Science* **308**, 1915–1920
554 (2005).
555 32. VerBerkmoes, N. C., Denef, V. J., Hettich, R. L. & Banfield, J. F. Systems biology:
556 Functional analysis of natural microbial consortia using community proteomics. *Nat Rev*
557 *Microbiol* **7**, 196–205 (2009).
558 33. Zhang, X., Ning, Z., Mayne, J., Moore, J. I., Li, J., Butcher, J. *et al.* MetaPro-IQ: a
559 universal metaproteomic approach to studying human and mouse gut microbiota. *Microbiome*
560 **4**, 31 (2016).
561 34. Steinsiek, S. & Bettenbrock, K. Glucose transport in Escherichia coli mutant strains
562 with defects in sugar transport systems. *J Bacteriol* **194**, 5897–5908 (2012).
563 35. Fath, M. J. & Kolter, R. ABC transporters: bacterial exporters. *Microbiol Rev* **57**, 995–
564 1017 (1993).

36. Nikaido, H. Maltose transport system of Escherichia coli: an ABC-type transporter. *FEBS Lett* **346**, 55–58 (1994).
37. Li, L., Abou-Samra, E., Ning, Z., Zhang, X., Mayne, J., Wang, J. *et al.* An in vitro model maintaining taxon-specific functional activities of the gut microbiome. *Nat Commun* **10**, 4146 (2019).
38. Zhang, X., Li, L., Mayne, J., Ning, Z., Stintzi, A., & Figeys, D. Assessing the impact of protein extraction methods for human gut metaproteomics. *Journal of Proteomics* **180**, 120–127 (2018).
39. Creskey, M., Li, L., Ning, Z., Fekete, E. E., Mayne, J., Walker, K. *et al.* An economic and robust TMT labeling approach for high throughput proteomic and metaproteomic analysis. 2022.07.30.502163 Preprint at https://doi.org/10.1101/2022.07.30.502163 (2022).
40. Perez-Riverol, Y., Bai, J., Bandla, C., García-Seisdedos, D., Hewapathirana, S., Kamatchinathan, S. *et al.* The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Research* **50**, D543–D552 (2022).
41. Patnode, M. L., Beller, Z. W., Han, N. D., Cheng, J., Peters, S. L., Terrapon, N. *et al.* Interspecies Competition Impacts Targeted Manipulation of Human Gut Bacteria by Fiber-Derived Glycans. *Cell* **179**, 59-73.e13 (2019).
42. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc* **11**, 2301–2319 (2016).
43. Cheng, K., Ning, Z., Zhang, X., Li, L., Liao, B., Mayne, J., & Figeys, D. MetaLab 2.0 Enables Accurate Post-Translational Modifications Profiling in Metaproteomics. *J. Am. Soc. Mass Spectrom.* **31**, 1473–1482 (2020).
44. Huang, T., Choi, M., Tzouros, M., Golling, S., Pandya, N. J., Banfai, B. *et al.* MSstatsTMT: Statistical Detection of Differentially Abundant Proteins in Experiments with Isobaric Labeling and Multiple Mixtures. *Mol Cell Proteomics* **19**, 1706–1723 (2020).
45. Ma, W., Kim, S., Chowdhury, S., Li, Z., Yang, M., Yoo, S. *et al.* DreamAI: algorithm for the imputation of proteomics data. 2020.07.21.214205 Preprint at https://doi.org/10.1101/2020.07.21.214205 (2021).
46. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**, 33–36 (2000).
47. Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
48. Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat Biotechnol* **32**, 834–841 (2014).
49. Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* **9**, 811–814 (2012).
50. Almeida-Neto, M., Guimarães, P., Guimarães Jr, P. R., Loyola, R. D. & Ulrich, W. A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. *Oikos* **117**, 1227–1239 (2008).
51. Sørensen, M. A., Fricke, J. & Pedersen, S. Ribosomal protein S1 is required for translation of most, if not all, natural mRNAs in Escherichia coli in vivo11Edited by D. Draper. *Journal of Molecular Biology* **280**, 561–569 (1998).

610 52. Galperin, M. Y., Wolf, Y. I., Garushyants, S. K., Vera Alvarez, R. & Koonin, E. V.
611 Nonessential Ribosomal Proteins in Bacteria and Archaea Identified Using Clusters of
612 Orthologous Genes. *Journal of Bacteriology* **203**, e00058-21 (2021).
613 53. Bakhti, S. Z. & Latifi-Navid, S. Oral microbiota and Helicobacter pylori in gastric
614 carcinogenesis: what do we know and where next? *BMC Microbiology* **21**, NA-NA (2021).
615 54. Cech, T. R. The Ribosome Is a Ribozyme. *Science* **289**, 878–879 (2000).
616 55. Xue, S. & Barna, M. Specialized ribosomes: a new frontier in gene regulation and
617 organismal biology. *Nat Rev Mol Cell Biol* **13**, 355–369 (2012).
618 56. Barabote, R. D. & Saier, M. H. Comparative Genomic Analyses of the Bacterial
619 Phosphotransferase System. *Microbiology and Molecular Biology Reviews* **69**, 608–634
620 (2005).
621 57. Yilmaz, B. & Li, H. Gut Microbiota and Iron: The Crucial Actors in Health and
622 Disease. *Pharmaceuticals* **11**, 98 (2018).
623 58. Seyoum, Y., Baye, K. & Humblot, C. Iron homeostasis in host and gut bacteria – a
624 complex interrelationship. *Gut Microbes* **13**, 1874855 (2021).
625 59. Bubunenko, M., Baker, T. & Court, D. L. Essentiality of Ribosomal and Transcription
626 Antitermination Proteins Analyzed by Systematic Gene Replacement in Escherichia coli.
627 *Journal of Bacteriology* **189**, 2844–2853 (2007).
628 60. Kanehisa, M. A database for post-genome analysis. *Trends Genet* **13**, 375–376 (1997).
629 61. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic*
630 *Acids Res* **28**, 27–30 (2000).
631 62. Mao, X., Cai, T., Olyarchuk, J. G. & Wei, L. Automated genome annotation and
632 pathway identification using the KEGG Orthology (KO) as a controlled vocabulary.
633 *Bioinformatics* **21**, 3787–3793 (2005).
634 63. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a
635 reference resource for gene and protein annotation. *Nucleic Acids Res* **44**, D457-462 (2016).
636 64. E, H. G. The multivariate niche. *Cold Spring Harbor Symposia on Quantitative*
637 *Biology* **22**, 415–421 (1957).
638 65. Leibold, M. A. The Niche Concept Revisited: Mechanistic Models and Community
639 Context. *Ecology* **76**, 1371–1382 (1995).
640 66. Holt, R. D. Bringing the Hutchinsonian niche into the 21st century: Ecological and
641 evolutionary perspectives. *PNAS* **106**, 19659–19665 (2009).
642 67. Li, L., Ning, Z., Zhang, X., Mayne, J., Cheng, K., Stintzi, A., & Figeys, D. RapidAIM:
643 a culture- and metaproteomics-based Rapid Assay of Individual Microbiome responses to
644 drugs. *Microbiome* **8**, 33 (2020).
645 68. You, C., Okano, H., Hui, S., Zhang, Z., Kim, M., Gunderson, C. W. *et al.*
646 Coordination of bacterial proteome with metabolism by cyclic AMP signalling. *Nature* **500**,
647 301–306 (2013).
648 69. Basan, M., Hui, S., Okano, H., Zhang, Z., Shen, Y., Williamson, J. R., & Hwa, T.
649 Overflow metabolism in Escherichia coli results from efficient proteome allocation. *Nature*
650 **528**, 99–104 (2015).
651 70. Tilman, D. Resource Competition between Plankton Algae: An Experimental and
652 Theoretical Approach. *Ecology* **58**, 338–348 (1977).
653 71. Wandersman, C. & Delepelaire, P. Bacterial iron sources: from siderophores to
654 hemophores. *Annu Rev Microbiol* **58**, 611–647 (2004).
655 72. Hibbing, M. E., Fuqua, C., Parsek, M. R. & Peterson, S. B. Bacterial competition:
656 surviving and thriving in the microbial jungle. *Nature Reviews Microbiology* **8**, 15–25 (2010).

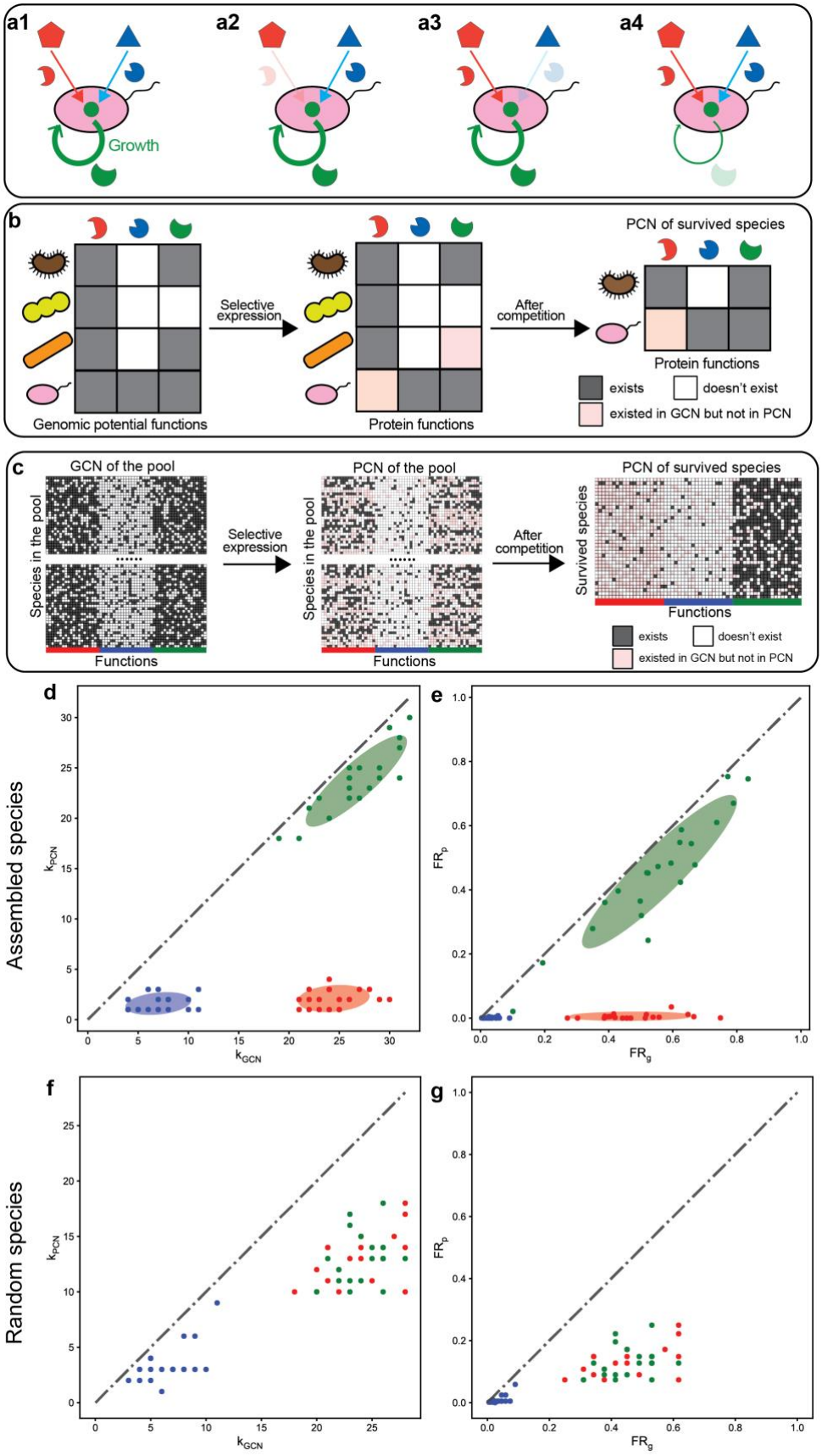
73. Smith, R. L. & Smith, T. M. *Elements of Ecology*. (Benjamin Cummings, San Francisco, 2003).
74. Hutchinson, G. E. The Paradox of the Plankton. *The American Naturalist* **95**, 137–145 (1961).
75. Marsland, R., Cui, W. & Mehta, P. The Minimum Environmental Perturbation Principle: A New Perspective on Niche Theory. *The American Naturalist* **196**, 291–305 (2020).
76. Cappelletti, V., Hauser, T., Piazza, I., Pepelnjak, M., Malinovska, L., Fuhrer, T. *et al.* Dynamic 3D proteomes reveal protein functional alterations at high resolution in situ. *Cell* **184**, 545-559.e22 (2021).
77. Schreiber, F., Littmann, S., Lavik, G., Escrig, S., Meibom, A., Kuypers, M. M., & Ackermann, M. Phenotypic heterogeneity driven by nutrient limitation promotes growth in fluctuating environments. *Nat Microbiol* **1**, 1–7 (2016).
78. Gutierrez-Ríos, R. M., Freyre-Gonzalez, J. A., Resendis, O., Collado-Vides, J., Saier, M., & Gosset, G. Identification of regulatory network topological units coordinating the genome-wide transcriptional response to glucose in Escherichia coli. *BMC Microbiology* **7**, 53 (2007).
79. Goyal, A., Bittleston, L. S., Leventhal, G. E., Lu, L. & Cordero, O. X. Interactions between strains govern the eco-evolutionary dynamics of microbial communities. *Elife* **11**, e74987 (2022).
80. Basan, M., Zhu, M., Dai, X., Warren, M., Sévin, D., Wang, Y. P., & Hwa, T. Inflating bacterial cells by increased protein synthesis. *Molecular Systems Biology* **11**, 836 (2015).
81. Serbanescu, D., Ojkic, N. & Banerjee, S. Nutrient-Dependent Trade-Offs between Ribosomes and Division Protein Synthesis Control Bacterial Cell Size and Growth. *Cell Reports* **32**, 108183 (2020).
82. Armengaud, J. Metaproteomics to understand how microbiota function: The crystal ball predicts a promising future. *Environmental Microbiology* **25**, 115–125 (2023).
83. Zhang, X., Walker, K., Mayne, J., Li, L., Ning, Z., Stintzi, A., & Figeys, D. Evaluating live microbiota biobanking using an ex vivo microbiome assay and metaproteomics. *Gut Microbes* **14**, 2035658 (2022).
84. 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 (2016).
85. MacArthur, R. Species packing and competitive equilibrium for many species. *Theoretical Population Biology* **1**, 1–11 (1970).
86. Goyal, A. & Maslov, S. Diversity, Stability, and Reproducibility in Stochastically Assembled Microbial Ecosystems. *Phys. Rev. Lett.* **120**, 158102 (2018).
87. Goldford, J. E., Lu, N., Bajić, D., Estrela, S., Tikhonov, M., Sanchez-Gorostiaga, A. *et al.* Emergent simplicity in microbial community assembly. *Science* **361**, 469–474 (2018).
88. Marsland III, R., Cui, W., Goldford, J., Sanchez, A., Korolev, K., & Mehta, P. Available energy fluxes drive a transition in the diversity, stability, and functional structure of microbial communities. *PLoS Comput Biol* **15**, e1006793 (2019).
89. Li, Z., Liu, B., Li, S. H. J., King, C. G., Gitai, Z., & Wingreen, N. S. Modeling microbial metabolic trade-offs in a chemostat. *PLoS Comput Biol* **16**, e1008156 (2020).
90. Dubinkina, V., Fridman, Y., Pandey, P. P. & Maslov, S. Multistability and regime shifts in microbial communities explained by competition for essential nutrients. *Elife* **8**, e49720 (2019).

91. Abrams, P. A. Character displacement and niche shift analyzed using consumer-resource models of competition. *Theor Popul Biol* **29**, 107–160 (1986).
92. Ispolatov, I. & Doebeli, M. A note on the complexity of evolutionary dynamics in a classic consumer-resource model. *Theor Ecol* **13**, 79–84 (2020).
93. Arthur, R. M. Species Packing, and What Competition Minimizes. *PNAS* **64**, 1369–1371 (1969).
94. Hardin, G. The competitive exclusion principle. *Science* **131**, 1292–1297 (1960).
95. Atmar, W. & Patterson, B. D. The measure of order and disorder in the distribution of species in fragmented habitat. *Oecologia* **96**, 373–382 (1993).


a
or
biomass
biomass
Gene Content Network
After competition
b
biomass
biomass
Protein Content Network
yes
no
c
Any relative abundances p1 and p2
kPCN
kGCN
Specialist function
Essential function
Ecological niche function
d
p1=p2=0.5
FRp
FRg
e
p1=0.9 and p2=0.1
FRp
FRg
f
Metagenome
Extract
DNA reads
A microbiome sample
Extract
Metaproteome
Digest
Peptides
Microbial proteins
Metagenomic and metaproteomic data collection
Protein IDs in each protein family
Peptide IDs of the protein family
Peptide sequences
Taxonomic matches of each peptide
Protein Families
Database search result files
Peptides
Taxonomy
"Protein-peptide bridge" approach to match taxa with protein families
Gene Content Network
Protein Family 1
Protein Family 2
Protein Content Network
Protein Family 1
Protein Family 2
Taxonomic profile
FRp
FRg
Function Assignment
FRg and FRp of protein families
GCN, PCN, and taxonomic profile

715

Figure 1: **Protein functions involved in determining ecological niches are postulated to have larger discrepancies between the gene-level functional redundancy FR_g and protein-level functional redundancy FR_p.** Here we use a hypothetical example with three representative proteins (3 broken circles with complementary shapes to their substrates) to demonstrate this point. **a,** Schematic of the genomic capacity of two microbial taxa (pink oval vs yellow indented oval). Two resources (red pentagon and blue triangle) are externally supplied to the community. The green metabolite can be transformed from the red or blue resource and further utilized in biomass synthesis. The pink taxon has the capacity of converting either supplied resource into the green metabolite (red and blue arrows), while the yellow taxon can only convert the red resource (red arrow). **b,** Schematic of expressed proteins for two microbial taxa after their competition in the same community. After the competition, the reduced resource conflict (represented by the pink taxon choosing the blue resource as the sole one to consume) can promote their coexistence. Gene content network (GCN) and protein content network (PCN) can be used to capture genomic capacity and expressed protein functions for all taxa. Alternatively, this network can be represented as incidence matrices on the bottom (grey areas imply the existence of edges connecting taxa to proteins). **c-d,** The comparison between k_{GCN} and k_{PCN} or between FR_g and FR_p helps to classify proteins into three protein functional types: specialist function, essential function, and niche function. In the calculation of FR_g and FR_p, we assume equal abundances of the two species, i.e., $p_1 = p_2 = 0.5$. **e,** The comparison between FR_g and FR_p when $p_1 = 0.9$ and $p_2 = 0.1$. **f,** The pipeline of assigning the functions (specialist, niche, or essential) to protein families based on the paired metagenomes and metaproteomes. Each individual's gut microbiome sample was subjected to DNA and protein extraction. Then a protein-peptide bridge approach can be used for generating the GCN based on the metagenome and PCN based on the metaproteome. When matched metagenomes are available, taxonomic and functional annotations of the metagenomes can be used for PCN generation. Based on the generated GCN, PCN, and taxonomic profile, FR_g and FR_p can be computed and used for the function assignment.


a1
a2
a3
a4
Growth
b
Selective expression
After competition
PCN of survived species
Protein functions
Genomic potential functions
Protein functions
exists
doesn't exist
existed in GCN but not in PCN
c
GCN of the pool
PCN of the pool
PCN of survived species
Species in the pool
Survived species
Functions
Functions
Functions
Assembled species
Random species
d
e
f
g
k_{PCN}
k_{GCN}
FR_p
FR_g

743

Figure 2: **Three protein functional clusters (specialist function, essential function, and niche function) considered in the community assembly model form three distinct clusters when the network degree and functional redundancy are compared between the GCN and PCN in model-generated synthetic data. a1-a4,** Three types of functions modeled have different ecological and metabolic roles. The niche function (red proteins) and specialist function (blue proteins) are modeled as abilities to consume externally supplied resources. The role of essential functions (green proteins) is considered as a reduction in the overall growth rate for each missing essential function. **b,** A schematic diagram of the community assembly. Species (ovals and indented ovals) with expressed gene functions selected via the sub-sampling of their genomic capacity. Then all species are co-cultured together to simulate their ecological competition. **c,** A simulation example of the community assembly, and the construction of GCN and PCN for the survived species. **d-e,** The comparison of network degree and functional redundancy respectively based on the GCN and PCN of survived species in the simulation example in panel-c. A Gaussian mixture model with 3 clusters is used to identify 3 protein functional clusters. Ellipses around clusters cover areas one standard deviation away from their means. **f-g,** The comparison of network degree and functional redundancy respectively based on the GCN and PCN of 35 species randomly selected from the 10,000 species in the initial pool. All points/functions are colored red (niche functions), green (essential functions), and blue (specialist functions) according to their types of functions in the model. k_{GCN} (or k_{PCN}) is the network degree of each function in the GCN (or PCN). FR_g (or FR_p) is the functional redundancy of each function on the gene level (or protein level), respectively.

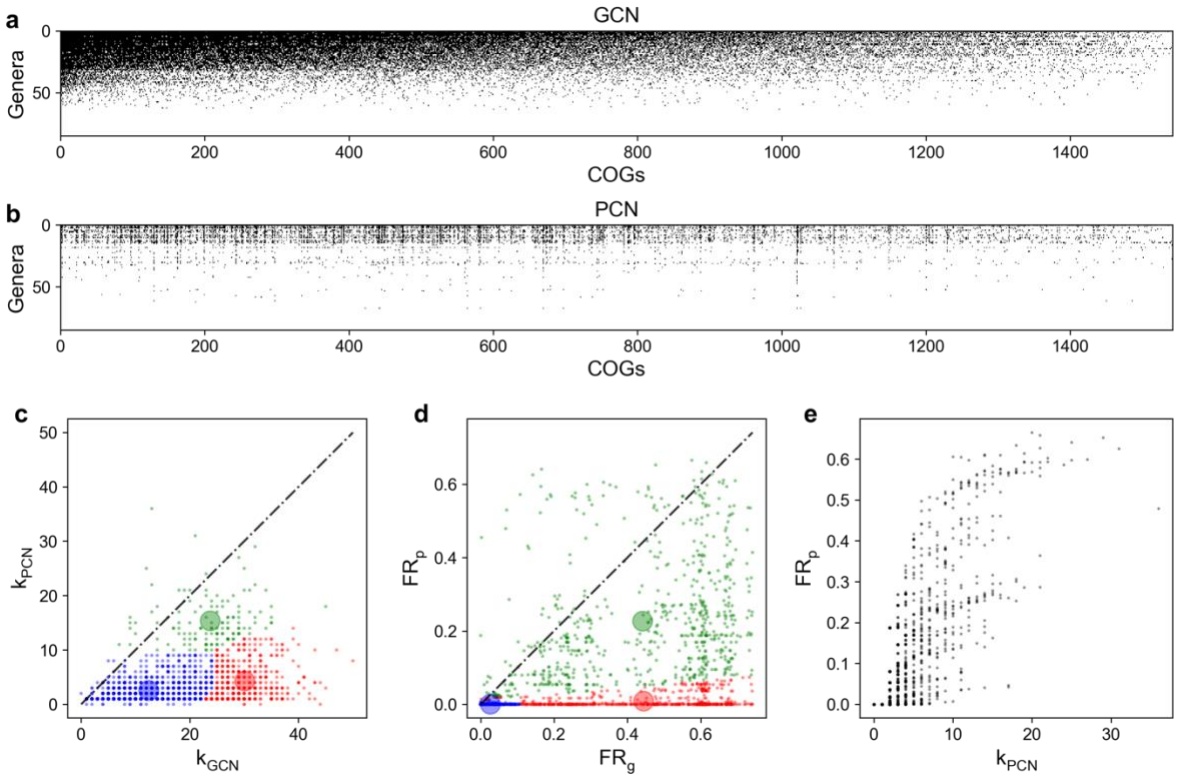


Figure 3: **Real data of the human gut microbiome showing three clusters on the plot that compares FR_g with FR_p.** Metagenome and metaproteome of subject HM454 mucosal-luminal interface samples[33] were used to construct GCN and PCN, respectively. **a,** The GCN shows if a genus owns (or doesn't own) a COG as its genomic capacity, which is colored in black (or white). The GCN matrix is ordered to have decreasing network degrees for both genera and COGs. **b,** The PCN shows if a genus expresses (or doesn't express) a COG as its protein function, which is colored in black (or white). The PCN matrix follows the same order as the GCN. **c,** Differences in network degree for most COGs are large. k_{GCN} is the network degree of each COG in the GCN (i.e. the number of genera owning each COG in the GCN). k_{PCN} is the network degree of each COG in the PCN (i.e. the number of genera owning each COG in the PCN). **d,** FR_g is larger than FR_p for most COGs. Three clusters with three distinct colors (blue, red, and green) are predicted by the Gaussian mixture model with 3 clusters fitted on synthetic data. The transparent large circles represent centroids of three clusters. **e,** The relationship between FR_p and network degree of PCN for COGs is not monotonic.

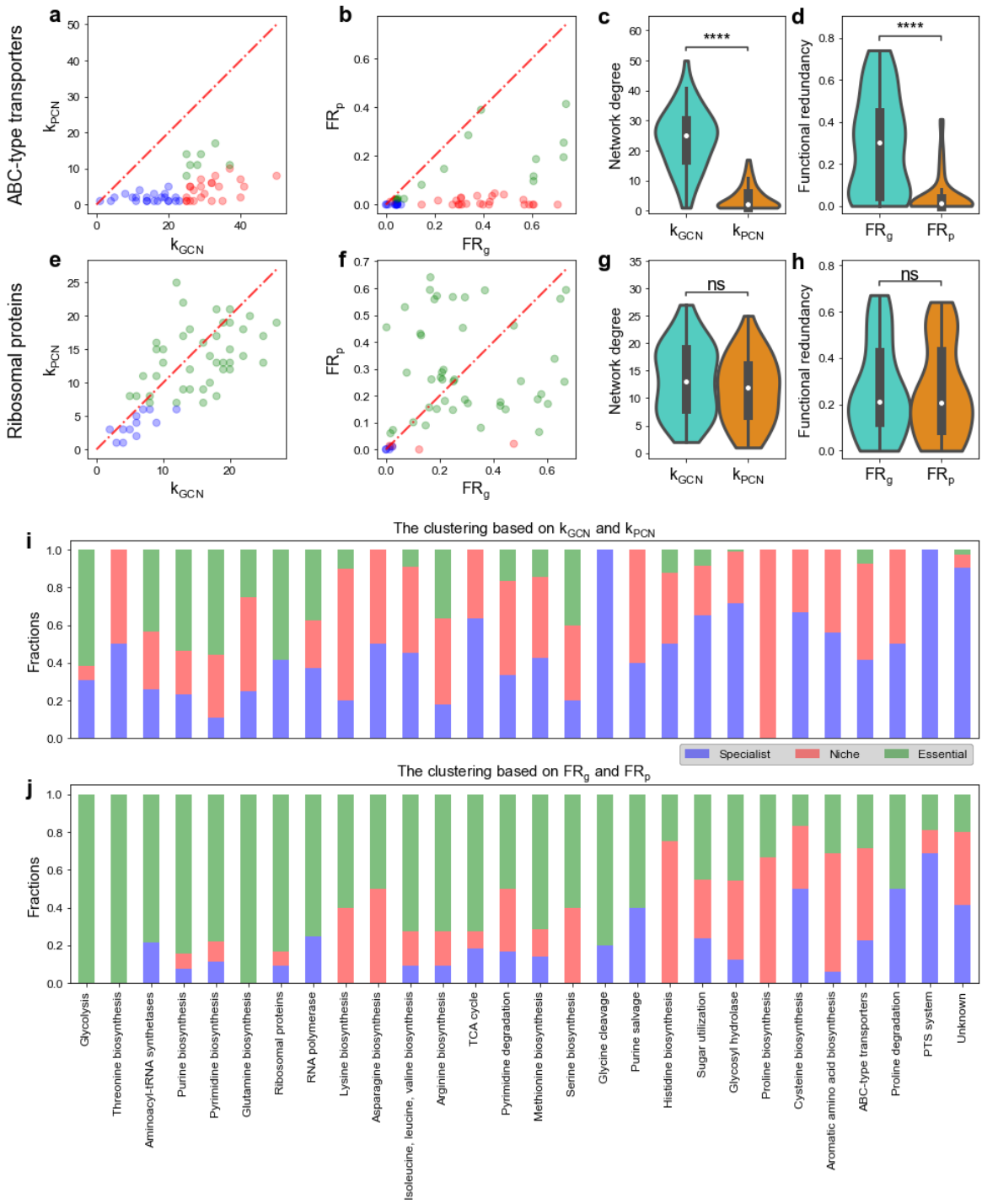
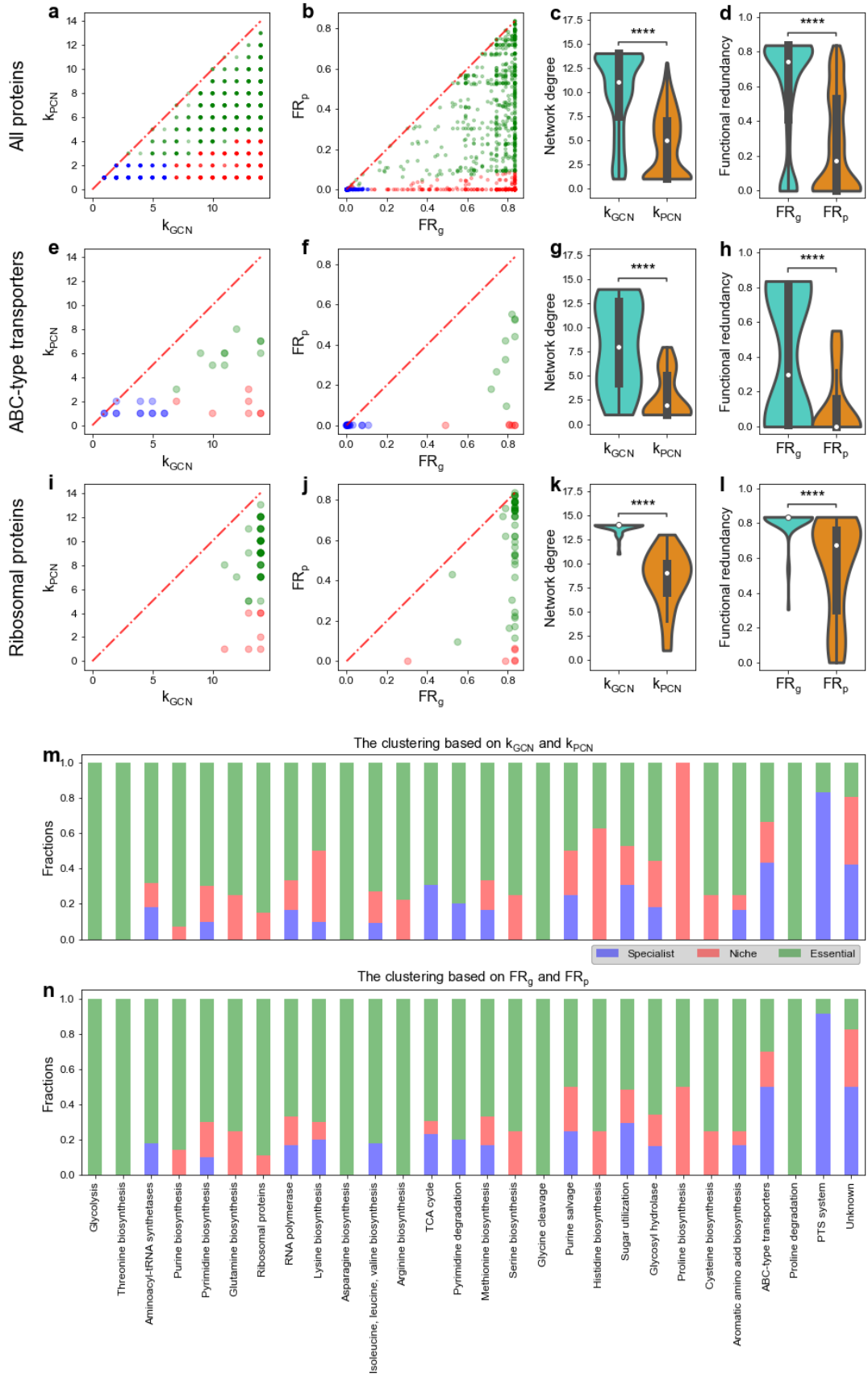


Figure 4: **Comparison of network degree and functional redundancy between the gene and protein level for many protein families including ABC-type transporters and ribosomal proteins from the human gut microbiome. a,** Network degrees in GCN are larger than network degrees in PCN for most ABC-type transporter COGs. k_{GCN} (or k_{PCN}) is the network degree of each COG in the GCN (or PCN). **b,** FR_g is larger than FR_p for most ABC-type transporter COGs. **c-d,** The distribution of network degrees and functional redundancies

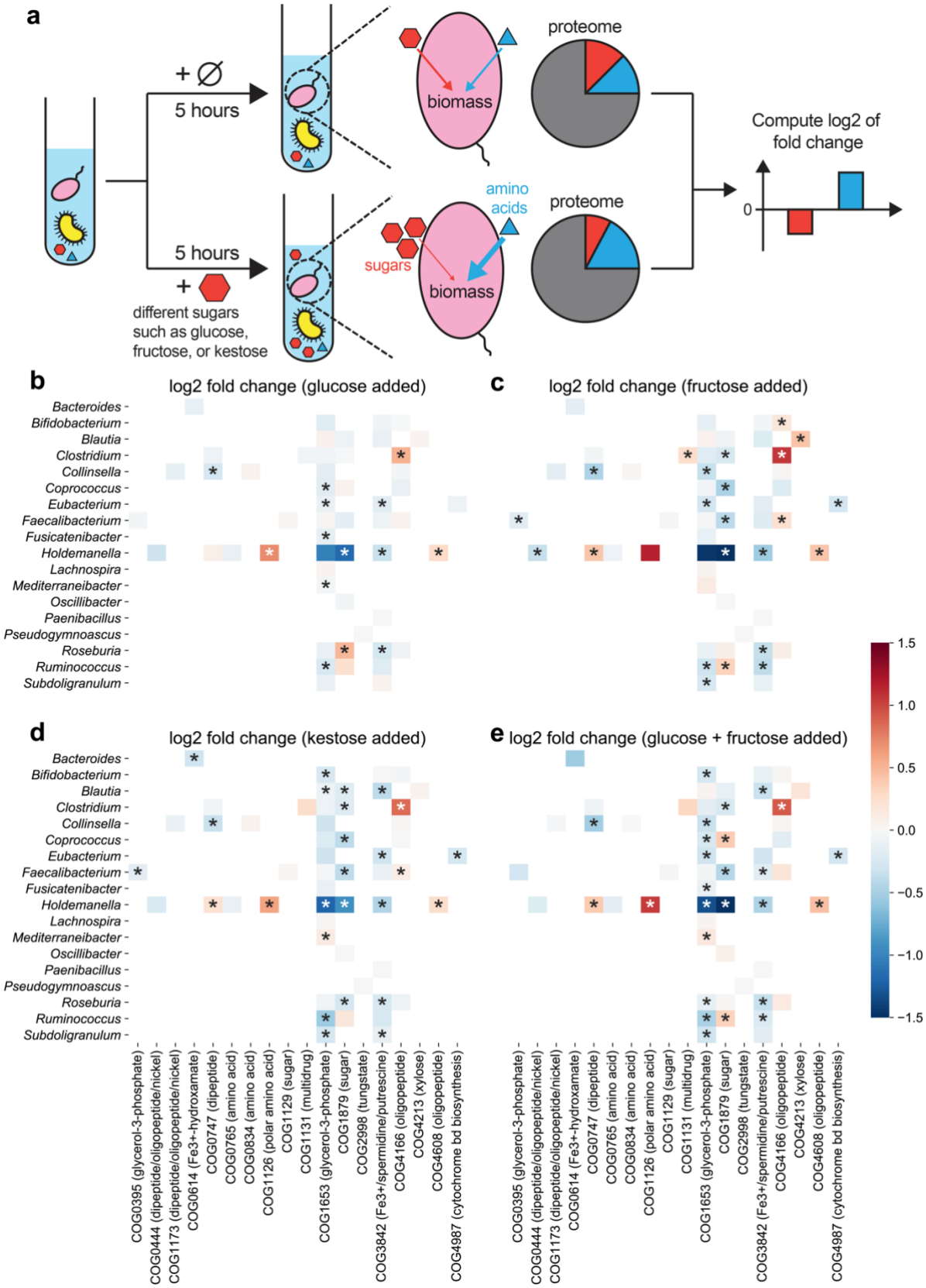
788 (violin plots and boxplots) for ABC-type transporter COGs show a significantly huge reduction
789 from k_{GCN} to k_{PCN} or from FR_g to FR_p. **e,** Network degrees in GCN are comparable with that in
790 PCN for most ribosomal protein COGs. **f,** FR_g is comparable with FR_p for most ribosomal
791 protein COGs. Points in scatter plots are colored by the same colors used in Fig. 3d. **g-h,** The
792 distribution of network degrees and functional redundancies (violin plots and boxplots) for
793 ribosomal protein COGs show no significant reduction from k_{GCN} to k_{PCN} or from FR_g to FR_p. **i,**
794 The fraction of assigned specialist, niche, or essential functions based on comparing network
795 degrees k_{GCN} and k_{PCN} for many protein families. **j,** The fraction of assigned specialist, niche,
796 or essential functions based on comparing functional redundancies FR_g and FR_p for many
797 protein families. In all boxplots, the middle white dot is the median, the lower and upper hinges
798 correspond to the first and third quartiles, and the black line ranges from the 1.5 × IQR (where
799 IQR is the interquartile range) below the lower hinge to 1.5 × IQR above the upper hinge. All
800 violin plots are smoothed by a kernel density estimator and 0 is set as the lower bound. All
801 statistical analyses were performed using the two-sided Mann-Whitney-Wilcoxon U Test with
802 Bonferroni correction between genomic capacity (GCN) and protein functions (PCN). P values
803 obtained from the test is divided into 5 groups: (1) $p > 0.05$ (ns), (2) $0.01 < p \leq 0.05$ (*), (3)
804 $10^{-3} < p \leq 0.01$ (**), (4) $10^{-4} < p \leq 10^{-3}$ (***), and (5) $p \leq 10^{-4}$ (****). Network degree
805 comparison of ABC transporters: $p = 7.11 \times 10^{-16}$. Network degree comparison of ribosomal
806 proteins: proteins: $p = 0.10$. Redundancy comparison of ABC transporters: $p = 2.19 \times 10^{-11}$.
807 Redundancy comparison of ribosomal proteins: $p = 1.00$.


All proteins
ABC-type transporters
Ribosomal proteins
a
b
c
d
e
f
g
h
i
j
k
l
m
n
k_{PCN}
k_{GCN}
FR_p
FR_g
Network degree
Functional redundancy
****
The clustering based on k_{GCN} and k_{PCN}
The clustering based on FR_g and FR_p
Fractions
Specialist
Niche
Essential
Glycolysis
Threonine biosynthesis
Aminoacyl-tRNA synthetases
Purine biosynthesis
Pyrimidine biosynthesis
Glutamine biosynthesis
Ribosomal proteins
RNA polymerase
Lysine biosynthesis
Asparagine biosynthesis
Isoleucine, leucine, valine biosynthesis
Arginine biosynthesis
TCA cycle
Pyrimidine degradation
Methionine biosynthesis
Serine biosynthesis
Glycine cleavage
Purine salvage
Histidine biosynthesis
Sugar utilization
Glycosyl hydrolase
Proline biosynthesis
Cysteine biosynthesis
Aromatic amino acid biosynthesis
ABC-type transporters
Proline degradation
PTS system
Unknown

808

809 Figure 5: **Comparison of network degree and functional redundancy between the gene**
810 **and protein level for many protein families including ABC-type transporters and**
811 **ribosomal proteins from the synthetic mouse gut microbial community. a,** Comparison
812 between network degrees of all COGs in the GCN (k_{GCN}) and network degrees of all COGs in
813 the PCN (k_{PCN}). **b,** Comparison between functional redundancies of all COGs in the GCN (FR_g)
814 and functional redundancies of all COGs in the PCN (FR_p). **c-d,** The distribution of network
815 degrees and functional redundancies (violin plots and boxplots) for all COGs. **e-h,** Comparison
816 between k_{GCN} and k_{PCN}, comparison between FR_g and FR_p, the distribution of network degrees,
817 and the distribution of functional redundancies for ABC-type transporter COGs. **i-l,** Comparison
818 between k_{GCN} and k_{PCN}, comparison between FR_g and FR_p, the distribution of network degrees,
819 and the distribution of functional redundancies for ribosomal protein transporter COGs. **m,** The
820 fraction of assigned specialist, niche, or essential functions based on comparing network
821 degrees k_{GCN} and k_{PCN} for many protein families. **n,** The fraction of assigned specialist, niche,
822 or essential functions based on comparing functional redundancies FR_g and FR_p for many
823 protein families. In all boxplots, the middle white dot is the median, the lower and upper hinges
824 correspond to the first and third quartiles, and the black line ranges from the 1.5 × IQR (where
825 IQR is the interquartile range) below the lower hinge to 1.5 × IQR above the upper hinge. All
826 violin plots are smoothed by a kernel density estimator and 0 is set as the lower bound. All
827 statistical analyses were performed using the two-sided Mann-Whitney-Wilcoxon U Test with
828 Bonferroni correction between genomic capacity (GCN) and protein functions (PCN). P values
829 obtained from the test is divided into 5 groups: (1) $p > 0.05$ (ns), (2) $0.01 < p \leq 0.05$ (*), (3)
830 $10^{-3} < p \leq 0.01$ (**), (4) $10^{-4} < p \leq 10^{-3}$ (***), and (5) $p \leq 10^{-4}$ (****).

831


a
+ ∅
5 hours
5 hours
+
different sugars
such as glucose,
fructose, or kestose
biomass
proteome
amino
acids
sugars
biomass
proteome
Compute log2 of
fold change
0
b
log2 fold change (glucose added)
c
log2 fold change (fructose added)
d
log2 fold change (kestose added)
e
log2 fold change (glucose + fructose added)
Bacteroides
Bifidobacterium
Blautia
Clostridium
Collinsella
Coprococcus
Eubacterium
Faecalibacterium
Fusicatenibacter
Holdemanella
Lachnospira
Mediterraneibacter
Oscillibacter
Paenibacillus
Pseudogymnoascus
Roseburia
Ruminococcus
Subdoligranulum
COG0395 (glycerol-3-phosphate)
COG0444 (dipeptide/oligopeptide/nickel)
COG1173 (dipeptide/oligopeptide/nickel)
COG0614 (Fe3+-hydroxamate)
COG0747 (dipeptide)
COG0765 (amino acid)
COG0834 (amino acid)
COG1126 (polar amino acid)
COG1129 (sugar)
COG1131 (multidrug)
COG1653 (glycerol-3-phosphate)
COG1879 (sugar)
COG2998 (tungstate)
COG3842 (Fe3+/spermidine/putrescine)
COG4166 (oligopeptide)
COG4213 (xylose)
COG4608 (oligopeptide)
COG4987 (cytochrome bd biosynthesis)
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

832

833 Figure 6: **Microbes modify their expression for ABC-type transporters to adapt to added**
834 **sugars.** All heatmaps share the same color bar on the right. **a,** Schematic of *in-vitro* cultures
835 of a collected human gut microbiome. In the treatment group, one sugar is added to the
836 community. Metaproteomic measurements 5 hours later were used to compare the intensity of
837 each taxon-specific protein using the log2 fold change of each protein's fraction (i.e. normalized
838 intensity over each genus) from the treatment group divided by that from the control group.
839 Log2 fold changes of ABC-type transporters were computed 5 hours after (**b**) glucose, (**c**)
840 fructose, (**d**) kestose, or (**e**) glucose and fructose is added. The transported metabolites for
841 each COG are added to the brackets.