Pairing Metagenomics and Metaproteomics to Pinpoint Ecological Niches and Metabolic Essentiality of Microbial Communities

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

[1]*Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA.*
[2]*State Key Laboratory of 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.*

† These authors contributed equally to this work.
* Correspondence: dfigeys@uottawa.ca, yyl@channing.harvard.edu

Abstract

Microbial genomes encode functional repertoire of microbes. However, microbes rely on various proteins to be expressed to carry out specific functions, and the expression of those proteins can be affected by the environment. It remains elusive how the selective expression of a protein depends on whether it is metabolically essential to the microbe's growth, or it can claim resources as an ecological niche. Here we show that by pairing metagenomics and metaproteomics data we can reveal whether a protein is relevant for occupying ecological niches or is essential for microbial metabolism. In particular, we developed a computational pipeline based on the quantification of the gene-level (or protein-level) functional redundancy of each protein, which measures the degree to which phylogenetically unrelated taxa can express (or have already expressed) the same protein, respectively. We validated this pipeline using both simulated data of a consumer-resource model and real data of human gut microbiome samples. Furthermore, for the real data, we showed that the metabolic and ecological roles of ABC-type transporters and ribosomal proteins predicted by our pipeline agree well with prior knowledge. Finally, we performed *in vitro* culture of a human gut microbiome sample and investigated how oversupplying various sugars involved in ecological niches influences the community structure and protein expression. The presented results help us identify metabolic and ecological roles of proteins, which will inform the design of nutrient interventions to modulate the human microbiome.

Introduction

The advance in metagenomic sequencing technology has enabled us to measure the genomic contents and functional potentials of microbial communities at an unprecedented rate, helping us understand how the functionality of microbes influences host health[1–3] and how microbial metabolism in natural environments enables biogeochemical cycling[4–6]. Based on metabolic models inferred from genomes, various computational approaches have been proposed to better quantify inter-species interactions and ecological concepts in microbial communities[7–12]. For example, metabolic networks of microbes have been employed to quantify complementarity and competition indices as a proxy for potential interactions[7]. Also, a nonlinear dimensionality reduction technique has been used to map bacterial metabolic niche space[9]. In addition, functional redundancy and functional stability for microbial communities were analyzed in the past[10–12].

A major limitation of those approaches is that they only rely on metagenomic data, which does not reflect true functional activities but only encodes functional capacity (or potential functions). In reality, at any given time and under any environmental condition, microbes only express a subset of their potential functions as proteins to carry out particular functions[13]. Recently, an ultra-deep metaproteomics approach has been developed to quantify expressed proteins in complex microbial communities, e.g., the human gut microbiome[14]. Pairing metagenomic and metaproteomic data offers the possibility to investigate how each protein is selectively expressed under different environmental conditions.

From the metabolic perspective, it is well known that some genes and their expressed proteins are indispensable for cell metabolism under any conditions, and microbes will not survive or reproduce if those genes are lost or those proteins are not expressed. Indeed, lacking proteins essential to microbial metabolism will cease microbial growth, regardless of ecological competition. For example, the growth of microbes relies on aminoacyl-tRNA[15,16]. Consequently, microbes have to express proteins involved in the aminoacyl-tRNA synthesis due to their metabolic essentiality to microbial growth[15,16].

From the ecological perspective, some proteins are expressed under ecological selection, and the presence of such proteins directly indicates which resources a microbe can utilize so as to thrive, i.e., the ecological niche of this microbe in the microbial community. Different proteins might enable a microbe to utilize different resources or adapt to varying environments. If the function of a protein can simply be performed by another protein, it may

be not necessary to express both proteins at the same time. This is evident in the case of *E. coli*, which prefers glucose over lactose due to the repressed expression of lactose-utilizing enzymes, even though it can use both sugars[17,18]. Such specialization of consuming one resource caused by the selective protein expression may reduce the niche overlap with other species. Another example is Acetyl-coenzyme synthetase (Acs) --- a protein that catalyzes the conversion of acetate into Acetyl-CoA, an essential intermediate in the metabolism[19,20]. The overexpression of Acs in *E. coli* can significantly switch glucose consumption to acetate consumption[21–24]. The glucose specialist (CV103) and acetate specialist (CV101) are two *E. coli* mutants with different metabolic strategies; CV103 does not express Acs while CV101 overexpresses it[21–24]. It has been shown that CV101 can consume acetate produced by CV103, and thus they achieve a coexistence due to the niche partionning[21,22].

How to understand the selective expression of microbial proteins is an outstanding question in microbiology. Does the behavior of selective expression of microbial proteins differ between metabolic function (e.g., essential for microbial growth metabolism) and ecological function (e.g., claiming resources as a niche)? To address this question, in this work we developed a computational method to perform paired metagenomic and metaproteomic[25–28,14] data analysis and revealed whether a protein is essential for microbial metabolism or relevant for occupying ecological niches. In particular, we used the metagenomic data to construct the Gene Content Network (GCN) --- a bipartite graph that connects microbial taxa to their genes (Fig. 1a), and used the metaproteomic data to construct the Protein Content Network (PCN) --- a bipartite graph that connects microbial taxa to their truly expressed protein functions (Fig. 1b). For each protein, we quantified its gene-level (or protein-level) functional redundancy (FR), which is defined as the degree to which unrelated taxa can express (or have already expressed) this protein, respectively. Using synthetic data generated by a consumer-resource model of microbial communities, we found that either the comparison of network degree of a protein (i.e., the number of taxa that own/express the protein) between the GCN and PCN or the comparison between the gene-level and protein-level FR of a protein can reveal its role in metabolic essentiality and ecological niches. Then we applied the same computational pipeline to analyze the real data of human gut microbiome samples to predict metabolic and ecological functions for proteins. We found that the metabolic and ecological roles of ABC-type transporters and ribosomal proteins predicted by our method agree well with prior knowledge. Finally, we performed *in vitro* culture experiments using human gut microbiome samples with and without sugars added to investigate how oversupplying various sugars involved in ecological niches influences the community structure and protein expression.

Results

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

Consider a microbiome sample with taxonomic profile $\boldsymbol{p} = (p_1, \dots, p_N)$, where p_i is the relative abundance of taxon-i and $\sum_{i=1}^{N} p_i = 1$. For a given protein, we can define its gene-level FR ($\mathrm{FR_g}$) within this sample as

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

where d_{ij}^{GCN} is the distance between taxon-i and taxon-j based on their genomic capacity to express this 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 protein, and $d_{ij}^{\mathrm{GCN}} = 1$ otherwise. For the same protein, we can also define its protein-level FR ($\mathrm{FR_p}$) within this sample as

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

where d_{ij}^{PCN} is the distance between taxon-i and taxon-j based on their expression of the protein. Again, we assume d_{ij}^{PCN} is binary, i.e., $d_{ij}^{\mathrm{PCN}} = 0$ if and only if both taxa have expressed the protein, and $d_{ij}^{\mathrm{PCN}} = 1$ otherwise. Note that here we define $\mathrm{FR_g}$ and $\mathrm{FR_p}$ for each protein. This is different from our previous studies[12,14], where FR was calculated by including all genes or proteins in the entire microbial community.

To demonstrate the definitions of $\mathrm{FR_g}$ and $\mathrm{FR_p}$, let's consider a simple community consisting of two coexisting *E. coli* strains CV101 and CV103 with relative abundance p_1 and p_2, respectively[21,22]. For the protein Acs that is required for the acetate consumption, since both CV101 and CV103 own this functional capacity, we have $d_{12}^{\mathrm{GCN}} = d_{21}^{\mathrm{GCN}} = 0$, and $\mathrm{FR_g} = 2p_1p_2$. However, because CV103 does not express Acs and CV101 overexpresses it[21–24], we have $d_{12}^{\mathrm{PCN}} = d_{21}^{\mathrm{PCN}} = 1$, and $\mathrm{FR_p} = 0$. Furthermore, we can compare the network degree of Acs in the GCN and PCN. The network degree of a protein in the GCN (denoted as $\mathrm{k_{GCN}}$) is the number of taxa owning the capacity to express the protein, while the network degree of a protein in the PCN ($\mathrm{k_{PCN}}$) is the number of taxa that have truly expressed the protein. Here, $\mathrm{k_{GCN}} = 2$ and $\mathrm{k_{PCN}} = 1$. Of course, not every protein is ecologically selected. For example, proteins involved in the aminoacyl-tRNA synthesis, critical for the growth of microbes, are not ecologically selected because the loss of ability to synthesize aminoacyl-tRNA inevitably stops the growth of microbes[15,16]. Hence, for each of the proteins involved in aminoacyl-tRNA synthesis, we expect $\mathrm{k_{GCN}} = \mathrm{k_{PCN}}$ and $\mathrm{FR_g} = \mathrm{FR_p}$.

Illustration of our computational pipeline using a hypothetical community

To illustrate our computational pipeline, let's consider a simple hypothetical example with two species (pink oval vs yellow indented oval in Fig. 1a, b). For the pink species to grow, it can either use the red resource (red pentagon in Fig. 1a) or the blue resource (blue triangle in Fig. 1a) and convert either of them to the green metabolite (green circle in Fig. 1a), which can then be assimilated into the cell biomass. For the yellow species, its growth will only occur by transforming the red resource into the green one to fuel the biomass synthesis (Fig. 1a). If the two species are co-cultured in the same environment to compete for externally supplied red and blue resources, an ideal scenario for them to coexist is that the pink species would choose to consume the blue resource, preventing resource competition with the yellow species (Fig. 1b), similar to the niche partitioning observed in the community of two coexisting *E. coli* strains: CV101 and CV103[21,22].

We can capture this hypothetical scenario of selective expression mathematically using the GCN and PCN of this community. The bipartite graph and incidence matrix representations of the GCN (or PCN) are shown in Fig. 1a (or Fig. 1b), respectively. Simply comparing the structure of the GCN and the PCN already offers us some insights into ecological niches and metabolic essentiality. For example, let's consider the protein responsible for converting red resource to green metabolite (this protein is represented as the red broken circle in Fig. 1a, b), its degree in the GCN is $\mathrm{k_{GCN}} = 2$, while its degree in the PCN is $\mathrm{k_{PCN}} = 1$. This degree reduction is due to distinct ecological niches being occupied by two species when they are cocultured. By contrast, the protein responsible for the assimilation of critical green metabolites (green broken circle in Fig. 1a, b) into biomass does not show a degree reduction from the GCN to the PCN, because it is essential for microbial growth.

An ecologically meaningful approach to understanding the selective expression of different proteins would be to systematically compare their respective $\mathrm{k_{GCN}}$ and $\mathrm{k_{PCN}}$ (Fig. 1c), which are independent of microbial compositions; or their respective $\mathrm{FR_g}$ and $\mathrm{FR_p}$ (Fig. 1d), which naturally involve microbial compositions in the calculation. Consider three distinct protein function types: (1) "niche functions" that are under strong ecological competition (e.g., red broken circle in Fig. 1c, d); (2) "specialist functions" that are specialized by a few taxa (e.g., blue broken circle in Fig. 1c, d); and (3) "essential functions" that are metabolically indispensable for many taxa (e.g., green broken circle in Fig. 1c, d). We anticipate that the three function types will occupy different regions in the $\mathrm{k_{GCN}}$ vs. $\mathrm{k_{PCN}}$ plot (or the $\mathrm{FR_g}$ vs. $\mathrm{FR_p}$ plot). Specifically, for essential functions, both their $\mathrm{k_{GCN}}$ and $\mathrm{k_{PCN}}$ (or $\mathrm{FR_g}$ and $\mathrm{FR_p}$) are high.

For specialist functions, both their $\mathrm{k_{GCN}}$ and $\mathrm{k_{PCN}}$ (or $\mathrm{FR_g}$ and $\mathrm{FR_p}$) are low. Niche functions have high $\mathrm{k_{GCN}}$ but low $\mathrm{k_{PCN}}$ (or high $\mathrm{FR_g}$ but low $\mathrm{FR_p}$).

Validate our computational pipeline using a consumer-resource model

Note that previously developed Consumer-Resource models (CRMs) only focus on physiologies of microbes (i.e. phenotypes)[29–31]. Simply put, those models ignored genomic capacity or potential functions, but only considered expressed functions (e.g., how species consume different resources). There was no attempt of building a consumer-resource model of microbial communities that integrates both potential and expressed functions. As a first step toward this direction, we constructed such a model.

We assumed three types of protein functions: niche functions (colored red), specialist functions (colored blue), and essential functions (colored green) in a functional pool. For simplicity, each of the niche (or specialist) functions is modeled as the consumption of a unique and externally supplied resource (Fig. 2a1). To model the difference between niche and specialist functions, we assume they are associated with different numbers of species (i.e., "consumers" in the consumer-resource modeling framework). The former should be associated with much more species than the latter. The loss of a niche or specialist function would make a species unable to consume the corresponding externally supplied resource (Fig. 2a2, a3). The loss of an essential function is simply modeled as the reduction of a species' growth rate (Fig. 2a4). Mathematically, we multiply the intrinsic growth rate of a species by a diminishing factor $\gamma = 0.95$ for each missing essential function.

The key issue in this genome-aware consumer-resource modelling framework is to decide how microbes select a subset of their potential functions to express. To tackle this issue, we first assigned potential functions to each species (Fig. 2b, left). In particular, for each species, each niche (specialist, or essential) function was assigned to the species' genome with probability p_{n} (p_{s}, or p_{e}), respectively. In our simulations, we set $p_{\mathrm{n}} = p_{\mathrm{e}} = 0.7$ to ensure that we cannot distinguish niche functions from essential functions only based on GCN and thus would like to see if they show different patterns after the community assembly. We set $p_{\mathrm{s}} = 0.2 < p_{\mathrm{n}} = p_{\mathrm{e}}$ so that specialist functions were assigned to fewer species than niche and essential functions. Then for each species, we determined its truly expressed functions by randomly sub-sampling a subset of its potential functions (Fig. 2b, middle). For function type-α ($\alpha = 1,2,3$), this was achieved by expressing each potential function with a species-specific and function-type-specific probability $p_{i,\alpha}$ randomly drawn from a uniform distribution $\mathcal{U}(0,1)$.

Since different species have different sub-sampling probabilities, some species will tend to be generalists (or specialists). Similar to all consumer-resource models[29–31], we assume a fixed expression pattern for each species and all resources being supplied so that we don't have to consider the complexity of adaptive expression (such as different expression patterns when different resources are supplied). In the end, we assembled all species in the same community and ran consumer-resource dynamics until the system reached a steady state, for which we constructed the PCN of the survived species (Fig. 2b, right).

We assumed the species pool consists of $N = 10{,}000$ species, and the function pool consists of 20 functions for each of the three function types. We introduced $10{,}000$ species to ensure the number of initial species in the assembly simulation is much larger than the number of functions so that we can assemble a high-diversity community in the end. The GCN of the initial species pool is shown in Fig. 2c (left). For each species, we randomly sub-sampled a subset of potential functions to express (middle panel, Fig. 2c). For each species, its true consumption rates are its maximal consumption rates divided by the number of resources the species can use (see Methods) to prevent the selection of generalist species that consume all resources without a penalty[32,33]. Due to the competitive exclusion principle[34], the maximal number of species survived in the final steady state is 40, because there are 40 unique externally supplied resources ("nutrients") in our model.

In Fig.2c (right), we show a simulation example with 35 species survived in the final steady state. For this assembled steady-state microbial community, we found that the three modeled protein functions types were correctly revealed as three clusters by the Gaussian mixture model in both the comparison of network degree (Fig. 2d) and FR (Fig. 2e). In particular, for niche functions (red cluster in Fig. 2d, e), their mean degree in PCN (2.1) is much lower than that in GCN (24.45), and their mean $\mathrm{FR_p}$ (0.005) is also much lower than their mean $\mathrm{FR_g}$ (0.48). For essential functions (green cluster in Fig. 2d, e), their mean degree in PCN (23.7) is close to that in GCN (26.7), and their mean $\mathrm{FR_p}$ (0.47) is also similar to their mean $\mathrm{FR_g}$ (0.57). For specialist functions (blue cluster in Fig. 2d, e), both their $\mathrm{k_{GCN}}$ and $\mathrm{k_{PCN}}$ (or $\mathrm{FR_g}$ and $\mathrm{FR_p}$) are low.

The three functional clusters revealed by the classification of network degrees and functional redundancies for all modeled protein functions exactly match the three function types in our model. Moreover, the relative positioning of the three functional clusters based on our simulation data agrees well with our hypothesis shown in Fig. 1. This clearly validates our hypothesis that niche-occupying proteins have a larger difference in FR and network degree than metabolically essential proteins.

We emphasize that the three functional clusters observed in the k_{GCN} vs. k_{PCN} (or the FR_g vs. FR_p) plot is highly nontrivial. It is a result of the community assembly. To demonstrate the importance of community assembly, we randomly picked 35 species (same as the number of survived species) from the initial pool with equal abundances (i.e., the relative abundance is 1/35 for each species) without natural selection and found that it is impossible to distinguish niche functions from essential functions (Fig. 2f, g). Interestingly, for essential functions, we noticed that those species survived after the community assembly tend to have much larger FR_p (with mean 0.478) than randomly selected species (with mean 0.132). By contrast, for niche functions, survived species tend to have a smaller FR_p (with mean 0.005) than randomly selected species (with mean 0.133). Similarly, we also computed FR for the same randomly picked 35 species that share the abundances as survived species in the simulation. Again, we cannot differentiate niche functions from essential functions (Supplementary Fig. 1).

We also simulated another community with 100 niche functions, 100 specialist functions, and 100 essential functions. The species pool still consists of $N = 10{,}000$ species. As shown in Supplementary Fig. 2), the results are similar to that for the community with fewer functions (Fig. 2).

Three protein functional clusters observed in human gut microbiomes

After the validation of our computational pipeline using simulated data, we further validated it on real data of human mucosal-luminal interface samples collected from the ascending colon of four children[14,28]. Here we focused on the genus level and annotated the identified proteins from metagenomics and metaproteomics data via the COGs (Clusters of Orthologous genes) database[35,36]. We constructed the GCN and PCN for all the samples following the same procedure as reported in a previous study[14], and took the intersected COGs between the two networks. In the main text, we focus on the analysis and discussion of subject HM454, and similar findings from the other three subjects are shown in Supplementary Figs. 4-6. For HM454, we used MetaPhlAn2[37] to obtain the taxonomic profile, which includes 85 genera with assigned relative abundances. Raw metagenomic reads and unique peptide sequences detected in metaproteomics were searched against an integrated gene catalog (IGC) database of the human gut microbiome[38] to generate the GCN and PCN respectively. Taxonomic assignment was performed using the 'protein-peptide bridge' method as described previously[14]. More details about data processing can be found in Methods. And the number of intersected COGs for the GCN and PCN associated with HM454 is 1,542. The genus- and COG-level GCN

266 and PCN of this microbiome sample are shown in Fig. 3a, b. The connectance (i.e., the number
267 of edges divided by the maximal number of possible edges) of the GCN (or PCN) is 0.220 (or
268 0.049), respectively. The GCN is nested with the nestedness value of 0.667 based on the
269 classical NODF (Nestedness based on Overlap and Decreasing Fill) measure[39] (Fig. 3a; see
270 Methods for details). The PCN has a lower nestedness value of 0.453 for the NODF measure
271 (Fig. 3b).

272 By comparing the network degree and functional redundancy of one COG in the GCN
273 (one column in Fig. 3a) with those for the same COG in the PCN, we can look into how the
274 COG impacts and is influenced by their metabolic essentiality and connection to occupy
275 ecological niches. For example, COG0539 is the ribosomal protein S1, which has been shown
276 to be essential for some microbes[40–44]. For subject HM454, 20 genera have COG0539 in the
277 GCN, while 15 genera have this COG in the PCN, hence $\mathrm{k_{GCN}} = 20$ and $\mathrm{k_{PCN}} = 15$.
278 Additionally, COG0539 has a similar level of functional redundancy in GCN and PCN: $\mathrm{FR_g} =$
279 0.476 and $\mathrm{FR_p} = 0.461$. These results suggest that COG0539 is crucial for microbial
280 metabolism, and not ecologically selected. Another example that falls into a different category
281 (i.e., niche functions) is COG1116, which is the ABC-type nitrate/sulfonate/bicarbonate
282 transport system[45]. For COG1116, we have $\mathrm{k_{GCN}} = 22 \gg \mathrm{k_{PCN}} = 2$; and $\mathrm{FR_g} = 0.388 \gg \mathrm{FR_p} =$
283 0.004, which is evidence for the further specification in transporting nitrate, sulfonate, or
284 bicarbonate across community members on the protein level. Different from the previous
285 examples, some functions are specialized by a few genera on the gene level and thus are still
286 specialized by those genera on the protein level. For example, COG1018 (Ferredoxin-NADP
287 reductase), which has $\mathrm{k_{GCN}} = \mathrm{k_{PCN}} = 1$ and $\mathrm{FR_g} = \mathrm{FR_p} = 0.0$, is classified as a specialist
288 function.

289 To systematically explore the difference between GCN and PCN, we visualized the
290 difference in the network degree (Fig. 3c) and functional redundancy (Fig. 3d) for all COGs. As
291 can be seen in Fig. 3c for comparing network degrees, nearly all COGs are below the black
292 dashed line of $\mathrm{k_{GCN}} = \mathrm{k_{PCN}}$ because the map from the genomic capacity to protein function is
293 a sub-sampling process. The network degrees in PCN for almost all points are less than 10
294 (1,365 out of 1,542) and much less than their corresponding network degrees in GCN (349 out
295 of 1,542 COGs have network degrees less than 10). 804 of 1,542 COGs have a reduction in
296 network degree by more than 80%. Eventually, the major difference in network degree will lead
297 to a significant difference in functional redundancy, although the reduction in network degree
298 from GCN to PCN cannot fully explain why many COGs have $\mathrm{FR_p} \sim 0$ (744 out of 1,542 have

299 $\mathrm{FR_p} < 0.01$ in Fig. 3d). Indeed, the network degrees for COGs in the PCN positively correlate
300 with $\mathrm{FR_p}$, but there is no simple relationship between k_{PCN} and $\mathrm{FR_p}$ (Fig. 3e). For example, for
301 L-arabinose isomerase (COG2160), its network degree in GCN (k_{GCN} = 8) is fairly close to the
302 network degree in PCN (k_{PCN} = 7), but its $\mathrm{FR_p}$ (0.04) is much lower than $\mathrm{FR_g}$ (0.23) since the
303 genus Blautia (which makes up 22% of the subject HM454's total microbial abundance) didn't
304 express L-arabinose isomerase, even if it has this capacity encoded in its genome.
305 We applied the Gaussian mixture model fitted on simulated data to classify all protein
306 functions in the real data and obtained 3 clusters from both the $\mathrm{k_{GCN}}$ vs. $\mathrm{k_{PCN}}$ plot (Fig. 3c) and
307 the $\mathrm{FR_g}$ vs. $\mathrm{FR_p}$ plot (Fig. 3d). Despite that the clustering of protein functions in real data looks
308 weaker than that in simulated data, the relative positioning of the three clusters (shaded areas
309 in Fig. 3c, d) agree well with our hypothesis shown in Fig. 1, as well our simulation results
310 shown in Fig. 2. We suspect that the weakened clustering might be due to (1) the variation of
311 k_{GCN} (or $\mathrm{FR_g}$) in real data (Fig. 3c, d) is much larger than that in simulated data (Fig. 2d, e); the
312 low resolution of the GCN and PCN in the real data (both were constructed at the genus level).
313 Note that some points in Fig. 3c, d are above the diagonal line, contradicting the sub-
314 sampling argument for the gene expression. For instance, we noticed that for the subject
315 HM454, 12 genera have COG0094 in the GCN, while 25 genera have this COG in the PCN.
316 Additionally, COG0094 is even less redundant in the GCN ($\mathrm{FR_g} = 0.166$) than it is in the PCN
317 ($\mathrm{FR_p} = 0.641$). $\mathrm{FR_g}$ should be always larger than $\mathrm{FR_p}$ if the PCN was a proper subgraph of the
318 GCN for COG0094. We believe this contradiction is largely due to the metagenomic
319 sequencing depth and the metaproteomic identification depth. We know that both
320 metagenomics and metaproteomics have depth limitations and require sufficient depth to
321 detect genes or proteins, respectively. More specifically, some proteins detected by the ultra-
322 deep metaproteomics are not found in putative protein sequences annotated from
323 metagenomes. For example, if more proteins were assigned to one COG by the
324 metaproteomics than annotated metagenomes, it indicates the number of taxa that express
325 proteins belonging to the COG is higher than the number of taxa that own the COG. As a result,
326 the network degree of the COG in the GCN is even higher than its network degree in the PCN,
327 making $\mathrm{FR_p}$ of the COG larger than its $\mathrm{FR_g}$ (evidenced by COG0094).
328

329 ## Comparing $\mathrm{FR_g}$ with $\mathrm{FR_p}$ pinpoints ecological niches and metabolic essentiality

330 In order to justify whether or not the FR comparison for many COGs is ecologically or
331 metabolically meaningful, we focus on two types of proteins: ABC-type transporters (under

strong ecological selection because they directly influence the ecological interactions and are influenced by resource availability)[45–47] and ribosomal proteins (under weak ecological selection because of their essentiality)[42–44].

ABC-type transporters are energy-requiring transporter proteins responsible for obtaining and releasing resources in the environment[45–47]. For example, if we consider a particular transporter responsible for the uptake of glucose from the environment, theoretically only top consumers of glucose would have the chance to claim this niche (consumption of glucose) from the ecological standpoint. Consequently, we should expect a specification in glucose consumption on the level of protein functions, even though many species have the capacity to utilize it. For the gut microbiota sample we investigated, we indeed found that k_{GCN} for all ABC-type transporters are much larger than their k_{PCN} (Fig. 4a). Similarly, we also found that $\mathrm{FR_g}$ for all ABC-type transporters are much larger than their $\mathrm{FR_p}$ (Fig. 4b). Many transporter proteins were classified to the red cluster (i.e., the cluster of niche functions) in Fig. 4b. Some transporter proteins were classified to the blue cluster (i.e., the cluster for specialist functions) due to the specialization on the gene level. As a result, such specialization would be carried to the protein level. Some transporter proteins were classified to the green cluster (i.e., the cluster for essential functions) because they have been proven essential for microbes. One example is the ABC-type Fe3+/spermidine/putrescine transporter (COG3842) which has $\mathrm{FR_g} = 0.339$ and $\mathrm{FR_p} = 0.285$. It has been shown that iron is essential for bacteria as it functions as a co-factor in iron-containing proteins in redox reactions, metabolic pathways, and electron transport chain mechanisms[48,49].

Ribosomal proteins are necessary for the growth of all living organisms because, as we know, the ribosome is the place where other proteins are synthesized[50,51]. Since ribosomal proteins are an indispensable part of microbial survival, all abilities of synthesizing such proteins are expected to be expressed. In our data, many ribosomal proteins were classified to the green cluster (i.e. the cluster for essential functions). Moreover, we found that their k_{GCN} were very close to their k_{PCN} (Fig. 4e). In Fig. 4f, we compared $\mathrm{FR_g}$ with $\mathrm{FR_p}$ and found many ribosomal proteins were classified to the green cluster (i.e. the cluster for essential functions), agreeing with our expectation that proteins with high $\mathrm{FR_g}$ and $\mathrm{FR_p}$ are more likely to be essential functions. Interestingly, two ribosomal proteins (L28 and L34) colored red in Fig. 4e have been shown to be non-essential[41,42,52] to microbes such as *E. coli*. Some specialized ribosomal proteins in microbial genomes continue to be specialized on the protein level and thus were classified to the blue cluster (i.e., the cluster for specialist functions).

365 Alternatively, we looked at the distribution of network degrees (Fig. 4c, g) and the
366 distribution of functional redundancy ($\mathrm{FR_g}$ or $\mathrm{FR_p}$ in Fig. 4d, h) for the two protein types to
367 observe their difference. For ABC-type transporters, the distribution of network degrees in PCN
368 is close to 0 (having a median of 2), while the median of network degrees in GCN is 25. For
369 ribosomal proteins, the distribution of network degrees in PCN (median is 12) is similar to that
370 in GCN (median is 14). For ABC-type transporters, the distribution of $\mathrm{FR_p}$ in PCN is close to 0
371 (with a median ~ 0.01), while the median of $\mathrm{FR_g}$ in GCN is around 0.30. For ribosomal proteins,
372 the distribution of $\mathrm{FR_p}$ in PCN (median ~ 0.20) is similar to the distribution of $\mathrm{FR_g}$ in GCN
373 (median ~ 0.21). The same patterns of ABC transporters showing a big reduction (in functional
374 redundancy and network degree) and ribosomal proteins showing little difference are also true
375 for the other 3 individuals (Supplementary Figs. 9-11).
376 We also validated the above results using a different functional annotation method,
377 KEGG Orthology (KO)[53–56]. The annotation rate of proteins involved in PCN of the four
378 individual microbiomes is 78% (much lower than 92% which we had for the COG annotation).
379 The contrasting difference between ABC-type transporters and ribosomal proteins is well
380 preserved (see Supplementary Fig. 7). Additionally, the distribution of $\mathrm{FR_p}$ shows a dramatic
381 difference across KO groups (Supplementary Fig. 8). Some ecologically strongly selected KO
382 groups have small $\mathrm{FR_p}$, while other metabolically essential KO groups show fairly large $\mathrm{FR_p}$
383 and big variations (see Supplementary Fig. 8). For example, almost all proteins in ABC
384 transporters and PTS systems have $\mathrm{FR_p}$ close to zero (Fig. Supplementary Fig. 8), and
385 transporters and PTS systems are well-known as the ecologically selected groups[45–47,57]. As a
386 comparison, proteins from Aminoacyl-tRNA biosynthesis, glycolysis, and ribosomes all have
387 big $\mathrm{FR_p}$ and huge variations across different proteins within the group (Supplementary Fig. 8).
388 In the past, the metabolic essentiality has been demonstrated for Aminoacyl-tRNA
389 biosynthesis[15,16], glycolysis[58,59], and ribosomes[42–44].
390

391 ## The response of community and protein expression to the introduction of sugars

392 In ecology, a niche is often defined as an abiotic and biotic factor that supports the survival of
393 species[9,60–62]. Therefore, niche functions are associated with corresponding limiting resources
394 involved in those functions. For example, COG1879 (ABC-type sugar transport system,
395 periplasmic component, contains N-terminal xre family HTH domain) which is categorized as
396 a niche function owing to its high $\mathrm{FR_g}$ of 0.486 and low $\mathrm{FR_p}$ of 0.041 for the subject HM454, is

397 associated with widely competed sugars in microbial communities. After inferring niche
398 functions such as ABC-type transporters by our computational pipeline, we wonder if it is
399 possible to influence the community structure by externally supplying more limiting resources
400 involved in the niche functions. To demonstrate this, we resort to the *in vitro* community and
401 are interested in how the community structure and expression of proteins involved in niche
402 functions respond to supplied limiting sugars. Specifically, we would like to see how proteins
403 relevant to ecological niche functions within one taxon change their expressions to achieve a
404 better living strategy for the taxon.

405 We used rapid assay for individual microbiome (RapidAIM)[63], which maintains the
406 functional profiles of individual gut microbiomes *in vitro*[64], to culture three different individual
407 human gut microbiota samples, and used metaproteomics to observe how taxon-specific
408 expression of proteins in the niche functional cluster respond to the presence of glucose,
409 fructose and kestose (Fig. 5a). Samples were cultured in technical triplicates, and were taken
410 at 0, 1, 5, 12, and 24 hours of culturing for optical density and metaproteomic analyses. 11-
411 plex tandem mass tag (TMT11plex) was used for metaproteomic quantification[65] for a total of
412 189 samples. To reflect the effect of introduced sugars on protein expression levels, we used
413 log2 of the ratio between normalized protein abundances/intensities (see Methods for details)
414 in the treatment and that in the control group (i.e. log2 of fold change in Fig. 5). We
415 hypothesized that the excessive supply of sugars renders carbon resources no longer limited
416 and instead microbes start to compete for other resources in relatively short supplies compared
417 to carbon resources such as nitrogen resources or amino acids because microbes need all
418 those resources proportionally (Fig. 5a). Therefore, microbes might have to over-express
419 proteins to uptake more non-carbon limiting resources to achieve better growth.

420 To understand how each taxon interacts with the environment and how introduced
421 sugars modulate the interaction, we focused on log2 fold changes of ABC-type transporters 5
422 hours later whose expression levels reveal rates for transporting nutrients (Fig. 5b-d). When
423 glucose is supplied in an excessive amount, log2 fold changes of most COGs are close to
424 zeros except for COG1126 (ABC-type polar amino acid transport system, ATPase component),
425 COG1653 (ABC-type glycerol-3-phosphate transport system, periplasmic component),
426 COG1879 (ABC-type sugar transport system, periplasmic component, contains N-terminal xre
427 family HTH domain), and COG4166 (ABC-type oligopeptide transport system, periplasmic
428 component). Many pronounced changes happen to the genus *Holdemanella* and it is
429 interesting to note that *Holdemanella* reduces the expression of transporters for importing
430 sugars (COG1879) and an intermediate in the glycolysis glycerol-3-phosphate (COG1653)

when glucose is added. Instead, it increases the expression of COG1126 which transports polar amino acids. This strategy benefits *Holdemanella* because the fraction of proteins from *Holdemanella* over all proteins in the community increases from 13.5% in the control to 15.8% with the added glucose. We also measured log2 fold changes of ABC-type transporters when fructose, glucose and fructose, or kestose is added and their overall patterns (Fig. 5c-e) are similar to the pattern when glucose is added (Fig. 5b). The correlation in log2 fold changes of ABC-type transporters between different added sugars is significant (Supplementary Fig. *12*). Similar fold changes of ABC-type transporters were observed for metaproteomic measurements 12 hours, and 24 hours later (Supplemental Figs. 14-15), while the fold changes 1 hour later are still fairly small (Supplemental Fig. 13). We also attempted to look at how ribosomal proteins respond to sugar supplies (Supplementary Fig. 16). Overall, log2 fold changes of ribosomal proteins are overwhelmingly positive, which probably implies a faster growth for microbes when simple sugars are supplied[32,33]. Therefore, we demonstrated that the sugars associated with the niche function (i.e., the sugar transport system) can be introduced to influence gene expression and modulate the community structure.

Discussion

Understanding the functions of proteins in the metabolism and how they are influenced by various ecological interactions is important to fully characterize ecological niches in a given microbial community. Typically, to check if a protein is metabolically essential, one has to knock out the gene in one microbial species that codes for the protein to check how the growth rate of the species reduces[42–44]. A usual way to determine a limiting resource often that is utilized by a protein follows: modify resource supplies and see how the total biomass changes[66–69]. Here, to complement those traditional experimental methods, we proposed a simpler computational method that can identify metabolic and ecological functions of proteins via the comparison of their $\mathrm{FR_g}$ and $\mathrm{FR_p}$, as well as their $\mathrm{k_{GCN}}$ and $\mathrm{k_{PCN}}$. We validated this computational method using both model-generated synthetic data and real data for human gut microbiomes. Also, when we selected two types of proteins (ABC-type transporters and ribosomal proteins in the real data representing niche functions and essential functions, respectively), most predicted protein functional clusters of the two types of proteins fell into the niche function cluster and the essential function cluster, respectively. Besides these two protein types, we were able to generate a list of $\mathrm{FR_p}$ and $\mathrm{FR_g}$ for all COGs (see Supplemental Data 1-4), which is useful for understanding the metabolic and ecological functions of proteins.

The presented results help us reconcile the conflict between the niche theory in ecology[62,70,71] and the observed functional redundancy[11,12]. The traditional niche theory is grounded in the competitive exclusion principle, stating that a resource (or niche) cannot be occupied by two species (or more than two species) for the steady-state conditions[62,70,71]. As a result of competition, organisms within the same community develop different surviving strategies to minimize their competition. One interesting example is the repetitive established coexistence between two evolved *E. coli* strains, even though a single clone of *E. coli* is initiated and maintained in a glucose-limited continuous or serial culture[21,72,73]. Cross-feeding between two evolved *E. coli* strains can be established when one bacterial strain consumes overflow metabolites like acetate excreted by the other bacterial strain[21]. Hence, the two strains avoid competition by specification on different resources (glucose and acetate). However, the picture from the niche theory clashes with the observed functional redundancy in microbial communities because the functional redundancy implies that many species own the same functions in their genomes[11,12]. We solved this dilemma by pointing out that proteins related to occupying ecological niches usually have very low $\mathrm{FR_p}$ and large $\mathrm{FR_g}$. Therefore, if we apply this concept in reverse, then large $\mathrm{FR_g}$ and small $\mathrm{FR_p}$ could help us to pinpoint niche functions.

There is a long-standing gap between the ecological model which considers the protein functions of organisms and the data analysis of genomic data to give ecological insights. Ever since Robert MacArthur proposed a community model in 1970 to consider how different consumers compete exclusively for renewing resources[74], many extensions of this model were proposed to include more complex ecological factors such as cross-feeding interactions[75–78] and multiple essential nutrients[79]. Almost all of them focus on the phenotype of microbes because only functions of expressed proteins are relevant for the consumption and production of nutrients in the ecosystem. Due to the lack of metaproteomic data, many computational approaches attempting to generate ecological implications rely on the over-complete inferred protein capacity derived from genomes[7,9–12]. To reconcile this gap, we built an ecological framework with the genomic capacity and protein functions together by introducing species with sub-sampled functions. The model framework is useful for explaining the difference between genomic capacity and protein functions. The selective expression can be considered as the same microbe with different expressions under different environments[80–82] or evolved strains from the same species that have distinct metabolic niches observed in evolutionary experiments of microbes[83,21,22].

It is worth noting that the assumption of the trade-off between generalists and specialists (represented by assuming that the total proteome is relatively constant) is very

498 important. In our model, this assumption is achieved by considering true consumption rates in
499 PCN as maximal consumption rates in GCN divided by the number of resources. The
500 importance of this trade-off lies in the fact that it forces the niche partitioning among species.
501 In the absence of this assumption, there is no pattern of redundancy difference since
502 generalists can always out-compete specialists. This trade-off makes sense because typically
503 the total proteome budgets for microbes have been observed to be relatively fixed[32,33].
504
505
506

507 # Methods

508 **In-vitro culture of single gut bacterial strains with added sugars.** Five gut commensal
509 bacterial strains, *Bacteroides vulgatus* ATCC 8482, *Bacteroides ovatus* ATCC 8483,
510 *Bacteroides uniformis* ATCC 8492, *Blautia hydrogenotrophica* DSM 10507, *Escherichia coli*
511 DSM 101114 were cultured with or without added sugars (glucose, sucrose and kestose). The
512 base culture medium without sugar added were modified based on the Yeast Casitone Fatty
513 Acids (YCFA) broth, containing 10.0 g/L casitone, 2.5 g/L yeast extract, 45 mg/L $MgSO_4 \cdot 7H_2O$,
514 90 mg/L $CaCl_2 \cdot 2H_2O$, 450 mg/L K_2HPO_4, 450 mg/L KH_2PO_4, 900 mg/L NaCl, 1.0 mg/L
515 resazurin, 4.0 g/L $NaHCO_3$, 1.0 g/L L-Cysteine-HCl, 10 mg/L Hemin, 1.90 mL/L acetic acid,
516 0.7 mL/L propionic acid, 90 μL/L iso-butyric acid, 100 μL/L n-valeric acid, 100 μL/L iso-valeric
517 acid, 0.02 mg/L biotin,0.02 mg/L folic acid, 0.05 mg/L thiamine-HCl, 0.05 mg/L riboflavin,
518 0.001 mg/L vitamin B12, 0.05 mg/L aminobenzoic acid. The pH was adjusted to between 6.7-
519 6.8, and autoclaved media were pre-reduced in an anaerobic chamber overnight. 5 g/L of
520 different sugars (glucose, sucrose, and kestose) were added to the base medium as treatment
521 groups. Master tubes of single bacterial strains were first cultured on Tryptic Soya Agar (TSA)
522 containing 5% sheep blood using the streak plate method. A single colony was picked from
523 each agar plate and inoculated into the base culture medium to culture for 24 hours, before
524 inoculating 100 μL of each culture into 10 mL of four different media: base medium without
525 sugar added, with glucose added, with sucrose added and with kestose added. After culturing
526 for 24 hours, optical density at 600 nm was tested in technical triplicates for each sample.
527 Cultured microbial cells were purified by washing with phosphate buffered saline (PBS) buffer
528 three times, and the resulting microbial pellets were stored at -80 ºC for proteomics analysis.
529

530 **In-vitro human gut microbiota culture with added sugars.** Three healthy individual
531 microbiota samples were collected and biobanked using our live microbiota biobanking
532 protocol[84]. The study was approved by the Ottawa Health Science Network Research Ethics
533 Board at the Ottawa Hospital, Ottawa, Canada (# 20160585–01 H). The frozen microbiome
534 samples were thawed at 37 ºC and cultured in our optimized culture medium[64] with or without
535 the presence of different sugars (10 mM glucose, 20 mM fructose, 10 mM glucose + 20 mM
536 fructose, or 10 mM kestose). Samples were cultured in technical triplicates, and were taken at
537 0 hr, 1hr, 5 hr, 12 hr and 24 hr of culturing for optical density and metaproteomic analyses.
538 After culturing, 96-well deep well plates were first centrifuged at 3,000 g for 45 min under 4 °C.
539 Then the pellets were washed in 4 °C phosphate buffered salin (PBS) buffer and centrifuged
540 at 3,000 g for 45 min again, before pelleting and removing culture debris three times using 300
541 g, 4 °C , 5 min centrifugation. Microbial suspensions were then centrifuged at 3,000 g, 4 °C
542 for another 45 min. The purified cell pellets were stored at -80 °C before protein extraction.
543
544 **Protein extraction, digestion and LC-MS/MS analysis.** For single strain samples, proteins
545 were extracted with 4% SDS 8M urea buffer in 100 mM Tris-HCl buffer and precipitated
546 overnight at -20 ºC, before being purified by washing with ice-cold acetone three times.
547 Quantified proteins were then reduced and alkylated before being digested using trypsin (50:1
548 protein-to-trypsin ratio) for 24 hours at 37 ºC and were desalted using reverse phase beads[85].
549 Proteomic samples were analyzed using an Orbitrap Exploris 480 mass spectrometer
550 (ThermoFisher Scientific Inc.) coupled with an UltiMate 3000 RSLCnano liquid
551 chromatography system following a 1-hour gradient of 5 to 35% (v/v) acetonitrile (v/v) at the
552 flow rate of 300 L/min. MS full scan was performed from 350 - 1400 m/z with a resolution of
553 60,000, followed by an MS/MS scan of 12 most intense ions, a dynamic exclusion repeat count
554 of one, exclusion duration of 30 s, and resolution of 15,000. Metaproteomics samples of the
555 cultured individual microbiomes were prepared using a semi-automated approach. Briefly,
556 samples were lysed in a buffer containing 8 M urea, 4% SDS in 100 mM Tris-HCl (pH = 8.0) to
557 extract microbial total proteins. The proteins were purified by a double-precipitation procedure
558 in 50%:50%:0.1% (v/v/v) acetone: ethanol: acetic acid solution. Protein digestion and desalting
559 steps were performed using an automated liquid handler (Hamilton Nimbus-96). Briefly, 100
560 μg proteins were dissolved in 100 μL 6 M urea in 100 mM Tris-HCl (pH 8) buffer, before being
561 reduced by 10 μL 0.1 M dithiothreitol (DTT) solution under 56 °C for 30 minutes and alkylated
562 by 10 μL 0.2 M iodoacetamide (IAA) solution in dark, 25 °C for 40 minutes. Samples were each
563 added 1000 μL 100 mM Tris·HCl buffer containing 2 μg/mL trypsin (trypsin:proteins = 1:50) for

564 a 24-hour digestion under 37 °C, before being desalted using an automated pipeline based on
565 reverse-phase (RP) desalting columns. 11-plex tandem mass tag (TMT11plex) was used for
566 metaproteomic quantification for a total of 189 samples. An even mixture of all samples was
567 used as the reference channel in each 11-plex. Samples were scrambled before labeling with
568 TMT11plex, so that each labeled sample contains samples from different individuals, different
569 time points and different treatments to avoid any bias that may be induced between analyses.
570 TMT-labelled samples were analyzed using an Orbitrap Exploris 480 mass spectrometer
571 (ThermoFisher Scientific Inc.) coupled with an UltiMate 3000 RSLCnano liquid
572 chromatography system following a 2-hour gradient of 5% to 35% solvent B (80% acetone
573 nitrile, 0.1% formic acid, v/v).
574
575 **Datasets.** Metagenomics data corresponding to the ultra-deep metaproteomic analysis of the
576 four individual microbiomes were obtained from the previous MetaPro-IQ study[14,28] (accessible
577 from the NCBI sequence read archive (SRA) under the accession of SRP068619) and the
578 same samples were reanalyzed by an ultra-deep metaproteomics approach[14] (accessible
579 through the ProteomeXchange Consortium (http://www.proteomexchange.org) via the PRIDE
580 partner repository[86]). Proteomics dataset of the cultured singles strain samples has been
581 deposited to ProteomeXchange Consortium via the PRIDE partner repository. Metaproteomic
582 dataset of the RapidAIM-cultured microbiome samples has been deposited to
583 ProteomeXchange Consortium via the PRIDE partner repository.
584
585 **Database search and data processing.** Proteomics database searches were performed by
586 combining FASTA databases of the individual strains downloaded from NCBI. The databases
587 were combined for performing database search using MaxQuant[87] 1.6.17.0, with the label-free
588 quantification option turned off. Metaproteomic database searches of cultured microbiome
589 samples were performed using MetaLab V2.2[88], MaxQuant option was used to search the TMT
590 dataset against the IGC database of the human gut microbiome. The resulting data table was
591 normalized using R package MSstatsTMT[89], and missing values were imputed using R
592 package DreamAI[90]. The "fraction" of each taxon-specific protein is computed by dividing the
593 protein intensity by the sum of intensities of all proteins assigned to the same taxon. The log2
594 fold change of each protein is obtained by taking log2 of the ratio between its fraction in the
595 treatment group (with added sugars) and its fraction in the control group (without added sugars).
596

597 **Generation of GCN and PCN.** For the ultra-deep metaproteomic dataset, the genus-COG
598 version of GCN and PCN tables were directly obtained from the previous work[14]. In addition,
599 here we generated a genus-KEGG version of GCN and PCN for each individual microbiome
600 using a similar method. Briefly, for the genus-KEGG GCN, by searching raw metagenomic
601 reads against an integrated gene catalog (IGC) database of the human gut microbiome[38], we
602 obtained a list of proteins quantified by read counts. FASTA sequences of these proteins were
603 searched against the KEGG database using GhostKOALA[91]. Taxonomic origination of the
604 proteins was obtained by searching against an in-house database generated with the NCBI
605 non-redundant (nr) database (downloaded 2/3/2016). To generate genus-KEGG PCN, the
606 taxonomic table of the metaproteomics dataset was directly obtained from MetaLab, and
607 KEGG annotation was also performed by querying protein FASTA sequences with
608 GhostKOALA. Protein group intensity was used as the quantification information in PCNs. For
609 the proteomic dataset of single strains, the whole proteomic FASTA database was submitted
610 to EggNOG mapper (http://eggnog-mapper.embl.de/, submitted Oct-30-2021, ran emapper.py
611 2.1.6) to obtain functional annotations. To generate GCN, protein coding sequence (CDS) files
612 were downloaded from NCBI, and the count of each protein id in the CDS files was considered
613 as the copy number of each gene in the GCN. For PCN generation, intensities of identified
614 proteins matched to each strain were used. Note that protein ids in the CDS file were 100%
615 matched with those in the proteomic FASTA database in each strain. For the metaproteomics
616 dataset of the cultured microbiome samples, functional information for the generation of PCN
617 was obtained from the resulting functional table automatically generated by the MetaLab
618 software. Taxonomic assignment was performed using the 'protein-peptide bridge' method as
619 described previously[14]. The PCNs for this dataset were then generated based on intensities of
620 COG-genus pairs.
621
622 **Normalized gene-level functional redundancy ($\mathrm{nFR_g}$) and normalized protein-level**
623 **functional redundancy ($\mathrm{nFR_p}$).** Across multiple samples, it is pointless to compare the $\mathrm{FR_g}$
624 or $\mathrm{FR_p}$ directly because of the difference in microbial taxonomic diversities. In fact, it has been
625 shown in the past that the normalized functional redundancy, which is the functional
626 redundancy divided by the taxonomic diversity, can be compared across samples[12]. In our
627 study, the definition for $\mathrm{nFR_g}$ is

628 $$\mathrm{nFR_g} = \frac{\sum_{i=1}^{N}\sum_{j\neq i}^{N}\left(1-d_{ij}^{\mathrm{GCN}}\right)p_i p_j}{\sum_{i=1}^{N}\sum_{j\neq i}^{N}p_i p_j}, \quad (1)$$

629 and the definition for $\mathrm{nFR_p}$ is

630 $$\mathrm{nFR_p} = \frac{\sum_{i=1}^{N}\sum_{j\neq i}^{N}\left(1-d_{ij}^{\mathrm{PCN}}\right)p_i p_j}{\sum_{i=1}^{N}\sum_{j\neq i}^{N}p_i p_j}. \quad (2)$$

631

632 **The community assembly model.**

633 *Step 1: Assignment of species' genomic capacity.* Three types of protein functions are modeled:
634 niche function, specialist function, and essential function. Both specialist function and niche
635 function are considered as the capacity to consume a unique and externally supplied resource.
636 The probability of a random consumer being assigned the ability to have a niche function is
637 0.7. To make fewer species own specialist functions in their genomes, the probability of a
638 random consumer being assigned the ability to have a specialist function is 0.2, much lower
639 than the probability of owning a niche function. The maximal consumption rate of a resource
640 by one species represents the consumption rate that the species would have if it allocates the
641 entire proteome (100%) to the consumption of the resource. If many resources are consumed,
642 the total proteome has to be divided into several parts and the consumption rates would be a
643 fraction of the corresponding maximal consumption rates. The essential function is not
644 modeled as the consumption of alternative resources due to its metabolic essentiality. Instead,
645 the essential function is modeled as multiplying the growth rate by a factor of 0.95 for each
646 missing essential function.

647 *Step 2: Assignment of species' protein functions based on their genomic capacity.* Each
648 species sub-samples its genomic potential functions with a sub-sampling probability *p* (which
649 is a random number uniformly distributed between 0 and 1) to obtain its protein functions (i.e.
650 which resource it can truly consume). As a result, all protein functions of species form the basis
651 for PCN. The true consumption rate of one species on a resource is its maximal consumption
652 rate on the resource divided by the number of resources that can be utilized by the species.
653 This process can be thought of as the proteome allocation to consume several resources
654 simultaneously[32,33]. This assumption imposes a trade-off between a generalist and a specialist
655 species: a generalist species utilizes more resources but has lower consumption rates for all
656 resources, while a specialist species consumes fewer resources but has higher consumption
657 rates for consumed resources.

658 *Step 3: Community assembly.* We assumed a chemostat environment, similar to the setting
659 considered by many Consumer-Resource models[75,77]. The dilution rate *D* is considered as 0.1
660 per hour. A fixed number of resources is considered and the pool concentrations (or supply
661 rates) for all resources are assumed to be the same for simplicity. For each species, the growth

rate is treated as the sum of consumption rates for different resources divided by the yield. For simplicity, all yields are assumed to be equal (Y = 1). Overall, the dynamics for the concentrations of resource i (denoted as C_i) and the abundance of the species α (written as B_α):

$$\frac{dC_i}{dt} = h_i - DC_i - \frac{\sum_\beta a_{\beta i}\gamma^{N_m} B_\beta C_i}{Y}, \quad (3)$$

$$\frac{dB_\alpha}{dt} = -DB_\alpha + \sum_j a_{\alpha j}\,\gamma^{N_m} B_\alpha C_j, \quad (4)$$

where $a_{\alpha i}$ is the consumption rate of species α on resource i, h_i is the supply rate of resource i, Y is the same yield assumed for all resources, $\gamma(= 0.95)$ is the diminishing rate for the overall consumption rate that is multiplied for each missing essential function, and N_m is the number of missing essential functions. The consumption rate of one species of a resource is randomly drawn from the uniform distribution between 0 and 1. Eventually, for each species, its true consumption rates are its randomly drawn consumption rates divided by the number of resources the species can use to constrain the total proteome budget[32,33]. The incidence matrix of the consumption abilities establishes part of PCN for niche functions and specialist functions of the species. The entire PCN is completed by including the presence/absence information of all essential functions.

Step 4: Generate GCN and PCN for survived species. When we simulated the above community assembly process to reach a steady-state in the chemostat environment, survived species can be found as species existing with non-negative abundances at the end of the simulation. For survived species, we can reconstruct the GCN and PCN for them. Within equipped GCN and PCN, we would be able to compute $\mathrm{FR_g}$, $\mathrm{FR_p}$, and network degrees (k_{GCN} and k_{PCN}).

Calculation of nestedness. To reveal the nested structure of an incidence matrix, we first need to use the Nestedness Temperature Calculator (NTC)[92] to organize the matrix. Then we adopted the NODF (Nestedness based on Overlap and Decreasing Fill) measure previously defined[39]. The measure can only be computed for binary incidence matrices. As with any perfectly nested matrix, two properties must be present: (1) decreasing fill, which means that the columns below and to the right should have fewer entries than the columns above and to the left; and (2) paired overlap, which implies that when an entry appears in the columns below and to the right, it should also appear in the columns above and to the left. The NODF measure is calculated by averaging these two properties across all pairs of an upper and lower row and a left and right column. For the comparison of each pair, if decreasing fill is not

satisfied, the pair will contribute 0 to the total nestedness. Otherwise, the pair's contribution is the percentage overlap in non-zero entries between the two rows or two columns.

Statistics. To calculate correlation throughout the study, we used Pearson's correlation coefficient. Wherever we used *P* values, we explained in the Methods how we calculated them, since for all such measurements in the study, we calculated the associated null distributions from scratch. All statistical tests were performed using standard numerical and scientific computing libraries in the Python programming language (version 3.7.1) and Jupyter Notebook (version 6.1).

Data and code availability. All code for simulations used in this manuscript can be found at XXX.

Acknowledgements. We thank Janice Mayne for the help with providing biobanking samples and thank Zhibin Ning for the help with running mass spectrometry. 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. acknowledges grants from the National Institutes of Health (R01AI141529, R01HD093761, RF1AG067744, UH3OD023268, U19AI095219, and U01HL089856).

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.

721 **References**

722 1. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of
723 microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
724 2. Flint, H. J., Scott, K. P., Louis, P. & Duncan, S. H. The role of the gut microbiota in
725 nutrition and health. *Nature Reviews Gastroenterology & Hepatology* **9**, 577–589 (2012).
726 3. Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory
727 bowel diseases. *Nature* **569**, 655–662 (2019).
728 4. Paerl, null & Pinckney, null. A Mini-review of Microbial Consortia: Their Roles in
729 Aquatic Production and Biogeochemical Cycling. *Microb Ecol* **31**, 225–247 (1996).
730 5. Falkowski, P. G., Fenchel, T. & Delong, E. F. The Microbial Engines That Drive
731 Earth's Biogeochemical Cycles. *Science* **320**, 1034–1039 (2008).
732 6. Louca, S., Parfrey, L. W. & Doebeli, M. Decoupling function and taxonomy in the
733 global ocean microbiome. *Science* **353**, 1272–1277 (2016).
734 7. Levy, R. & Borenstein, E. Metabolic modeling of species interaction in the human
735 microbiome elucidates community-level assembly rules. *PNAS* **110**, 12804–12809 (2013).
736 8. Pacheco, A. R., Moel, M. & Segrè, D. Costless metabolic secretions as drivers of
737 interspecies interactions in microbial ecosystems. *Nature Communications* **10**, 103 (2019).
738 9. Fahimipour, A. K. & Gross, T. Mapping the bacterial metabolic niche space. *Nature*
739 *Communications* **11**, 4887 (2020).
740 10. Louca, S. *et al.* High taxonomic variability despite stable functional structure across
741 microbial communities. *Nat Ecol Evol* **1**, 15 (2016).
742 11. Louca, S. *et al.* Function and functional redundancy in microbial systems. *Nature*
743 *Ecology & Evolution* **2**, 936 (2018).
744 12. Tian, L. *et al.* Deciphering functional redundancy in the human microbiome. *Nature*
745 *Communications* **11**, 6217 (2020).
746 13. Franzosa, E. A. *et al.* Relating the metatranscriptome and metagenome of the human
747 gut. *Proc Natl Acad Sci U S A* **111**, E2329-2338 (2014).
748 14. Li, L. *et al.* Revealing Protein-Level Functional Redundancy in the Human Gut
749 Microbiome using Ultra-deep Metaproteomics. *bioRxiv* 2021.07.15.452564 (2021)
750 doi:10.1101/2021.07.15.452564.
751 15. Ibba, M. & Soll, D. Aminoacyl-tRNA synthesis. *Annu Rev Biochem* **69**, 617–650
752 (2000).
753 16. Parker, D. J. *et al.* Growth-Optimized Aminoacyl-tRNA Synthetase Levels Prevent
754 Maximal tRNA Charging. *Cell Syst* **11**, 121-130.e6 (2020).
755 17. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J*
756 *Mol Biol* **3**, 318–356 (1961).
757 18. Okano, H., Hermsen, R., Kochanowski, K. & Hwa, T. Regulation underlying
758 hierarchical and simultaneous utilization of carbon substrates by flux sensors in Escherichia
759 coli. *Nat Microbiol* **5**, 206–215 (2020).
760 19. Kumari, S. *et al.* Regulation of Acetyl Coenzyme A Synthetase inEscherichia coli.
761 *Journal of Bacteriology* **182**, 4173–4179 (2000).
762 20. Starai, V. J. & Escalante-Semerena, J. C. Acetyl-coenzyme A synthetase (AMP
763 forming). *Cell Mol Life Sci* **61**, 2020–2030 (2004).
764 21. Rosenzweig, R. F., Sharp, R. R., Treves, D. S. & Adams, J. Microbial evolution in a
765 simple unstructured environment: genetic differentiation in Escherichia coli. *Genetics* **137**,
766 903–917 (1994).

22. Treves, D. S., Manning, S. & Adams, J. Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of Escherichia coli. *Mol Biol Evol* **15**, 789–797 (1998).
23. Lin, H., Castro, N. M., Bennett, G. N. & San, K.-Y. Acetyl-CoA synthetase overexpression in <Emphasis Type="Italic">Escherichia coli</Emphasis> demonstrates more efficient acetate assimilation and lower acetate accumulation: a potential tool in metabolic engineering. *Appl Microbiol Biotechnol* **71**, 870–874 (2006).
24. Kinnersley, M. A., Holben, W. E. & Rosenzweig, F. E Unibus Plurum: genomic analysis of an experimentally evolved polymorphism in Escherichia coli. *PLoS Genet* **5**, e1000713 (2009).
25. Penzlin, A. *et al.* Pipasic: similarity and expression correction for strain-level identification and quantification in metaproteomics. *Bioinformatics* **30**, i149-156 (2014).
26. Ram, R. J. *et al.* Community proteomics of a natural microbial biofilm. *Science* **308**, 1915–1920 (2005).
27. VerBerkmoes, N. C., Denef, V. J., Hettich, R. L. & Banfield, J. F. Systems biology: Functional analysis of natural microbial consortia using community proteomics. *Nat Rev Microbiol* **7**, 196–205 (2009).
28. Zhang, X. *et al.* MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. *Microbiome* **4**, 31 (2016).
29. Abrams, P. A. Character displacement and niche shift analyzed using consumer-resource models of competition. *Theor Popul Biol* **29**, 107–160 (1986).
30. Ispolatov, I. & Doebeli, M. A note on the complexity of evolutionary dynamics in a classic consumer-resource model. *Theor Ecol* **13**, 79–84 (2020).
31. Arthur, R. M. Species Packing, and What Competition Minimizes. *PNAS* **64**, 1369–1371 (1969).
32. You, C. *et al.* Coordination of bacterial proteome with metabolism by cyclic AMP signalling. *Nature* **500**, 301–306 (2013).
33. Basan, M. *et al.* Overflow metabolism in Escherichia coli results from efficient proteome allocation. *Nature* **528**, 99–104 (2015).
34. Hardin, G. The competitive exclusion principle. *Science* **131**, 1292–1297 (1960).
35. 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).
36. Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
37. Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* **9**, 811–814 (2012).
38. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat Biotechnol* **32**, 834–841 (2014).
39. 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).
40. 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).

41. Galperin, M. Y., Wolf, Y. I., Garushyants, S. K., Vera Alvarez, R. & Koonin, E. V. Nonessential Ribosomal Proteins in Bacteria and Archaea Identified Using Clusters of Orthologous Genes. *Journal of Bacteriology* **203**, e00058-21 (2021).
42. Shoji, S., Dambacher, C. M., Shajani, Z., Williamson, J. R. & Schultz, P. G. Systematic chromosomal deletion of bacterial ribosomal protein genes. *J Mol Biol* **413**, 751–761 (2011).
43. Dabbs, E. R. Mutants lacking individual ribosomal proteins as a tool to investigate ribosomal properties. *Biochimie* **73**, 639–645 (1991).
44. Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* **2**, 2006.0008 (2006).
45. Steinsiek, S. & Bettenbrock, K. Glucose transport in Escherichia coli mutant strains with defects in sugar transport systems. *J Bacteriol* **194**, 5897–5908 (2012).
46. Fath, M. J. & Kolter, R. ABC transporters: bacterial exporters. *Microbiol Rev* **57**, 995–1017 (1993).
47. Nikaido, H. Maltose transport system of Escherichia coli: an ABC-type transporter. *FEBS Lett* **346**, 55–58 (1994).
48. Yilmaz, B. & Li, H. Gut Microbiota and Iron: The Crucial Actors in Health and Disease. *Pharmaceuticals* **11**, 98 (2018).
49. Seyoum, Y., Baye, K. & Humblot, C. Iron homeostasis in host and gut bacteria – a complex interrelationship. *Gut Microbes* **13**, 1874855 (2021).
50. Cech, T. R. The Ribosome Is a Ribozyme. *Science* **289**, 878–879 (2000).
51. Xue, S. & Barna, M. Specialized ribosomes: a new frontier in gene regulation and organismal biology. *Nat Rev Mol Cell Biol* **13**, 355–369 (2012).
52. Bubunenko, M., Baker, T. & Court, D. L. Essentiality of Ribosomal and Transcription Antitermination Proteins Analyzed by Systematic Gene Replacement in Escherichia coli. *Journal of Bacteriology* **189**, 2844–2853 (2007).
53. Kanehisa, M. A database for post-genome analysis. *Trends Genet* **13**, 375–376 (1997).
54. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27–30 (2000).
55. Mao, X., Cai, T., Olyarchuk, J. G. & Wei, L. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics* **21**, 3787–3793 (2005).
56. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44**, D457-462 (2016).
57. Kotrba, P., Inui, M. & Yukawa, H. Bacterial phosphotransferase system (PTS) in carbohydrate uptake and control of carbon metabolism. *Journal of Bioscience and Bioengineering* **92**, 502–517 (2001).
58. Romano, A. H. & Conway, T. Evolution of carbohydrate metabolic pathways. *Research in Microbiology* **147**, 448–455 (1996).
59. Yan, Y. *Engineering Microbial Metabolism For Chemical Synthesis: Reviews And Perspectives*. (World Scientific, 2017).
60. E, H. G. The multivariate niche. *Cold Spring Harbor Symposia on Quantitative Biology* **22**, 415–421 (1957).
61. Leibold, M. A. The Niche Concept Revisited: Mechanistic Models and Community Context. *Ecology* **76**, 1371–1382 (1995).
62. Holt, R. D. Bringing the Hutchinsonian niche into the 21st century: Ecological and evolutionary perspectives. *PNAS* **106**, 19659–19665 (2009).

63. Li, L. *et al.* RapidAIM: a culture- and metaproteomics-based Rapid Assay of Individual Microbiome responses to drugs. *Microbiome* **8**, 33 (2020).
64. Li, L. *et al.* An in vitro model maintaining taxon-specific functional activities of the gut microbiome. *Nat Commun* **10**, 4146 (2019).
65. Creskey, M. *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).
66. Tilman, D. Resource Competition between Plankton Algae: An Experimental and Theoretical Approach. *Ecology* **58**, 338–348 (1977).
67. Wandersman, C. & Delepelaire, P. Bacterial iron sources: from siderophores to hemophores. *Annu Rev Microbiol* **58**, 611–647 (2004).
68. Hibbing, M. E., Fuqua, C., Parsek, M. R. & Peterson, S. B. Bacterial competition: surviving and thriving in the microbial jungle. *Nature Reviews Microbiology* **8**, 15–25 (2010).
69. Smith, R. L. & Smith, T. M. *Elements of ecology*. (Benjamin Cummings, 2003).
70. Hutchinson, G. E. The Paradox of the Plankton. *The American Naturalist* **95**, 137–145 (1961).
71. Marsland, R., Cui, W. & Mehta, P. The Minimum Environmental Perturbation Principle: A New Perspective on Niche Theory. *The American Naturalist* **196**, 291–305 (2020).
72. Rozen, D. E. & Lenski, R. E. Long-Term Experimental Evolution in Escherichia coli. VIII. Dynamics of a Balanced Polymorphism. *Am. Nat.* **155**, 24–35 (2000).
73. Good, B. H., McDonald, M. J., Barrick, J. E., Lenski, R. E. & Desai, M. M. The dynamics of molecular evolution over 60,000 generations. *Nature* **551**, 45–50 (2017).
74. MacArthur, R. Species packing and competitive equilibrium for many species. *Theoretical Population Biology* **1**, 1–11 (1970).
75. Goyal, A. & Maslov, S. Diversity, Stability, and Reproducibility in Stochastically Assembled Microbial Ecosystems. *Phys. Rev. Lett.* **120**, 158102 (2018).
76. Goldford, J. E. *et al.* Emergent simplicity in microbial community assembly. *Science* **361**, 469–474 (2018).
77. Marsland, R. *et al.* Available energy fluxes drive a transition in the diversity, stability, and functional structure of microbial communities. *PLoS Comput Biol* **15**, e1006793 (2019).
78. Li, Z. *et al.* Modeling microbial metabolic trade-offs in a chemostat. *PLoS Comput Biol* **16**, e1008156 (2020).
79. 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).
80. Cappelletti, V. *et al.* Dynamic 3D proteomes reveal protein functional alterations at high resolution in situ. *Cell* **184**, 545-559.e22 (2021).
81. Schreiber, F. *et al.* Phenotypic heterogeneity driven by nutrient limitation promotes growth in fluctuating environments. *Nat Microbiol* **1**, 1–7 (2016).
82. Gutierrez-Ríos, R. M. *et al.* Identification of regulatory network topological units coordinating the genome-wide transcriptional response to glucose in Escherichia coli. *BMC Microbiology* **7**, 53 (2007).
83. 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).

84. Zhang, X. *et al.* Evaluating live microbiota biobanking using an ex vivo microbiome assay and metaproteomics. *Gut Microbes* **14**, 2035658 (2022).
85. Zhang, X. *et al.* Assessing the impact of protein extraction methods for human gut metaproteomics. *Journal of Proteomics* **180**, 120–127 (2018).
86. Perez-Riverol, Y. *et al.* The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Research* **50**, D543–D552 (2022).
87. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc* **11**, 2301–2319 (2016).
88. Cheng, K. *et al.* MetaLab 2.0 Enables Accurate Post-Translational Modifications Profiling in Metaproteomics. *J. Am. Soc. Mass Spectrom.* **31**, 1473–1482 (2020).
89. Huang, T. *et al.* MSstatsTMT: Statistical Detection of Differentially Abundant Proteins in Experiments with Isobaric Labeling and Multiple Mixtures. *Mol Cell Proteomics* **19**, 1706–1723 (2020).
90. Ma, W. *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).
91. 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).
92. Atmar, W. & Patterson, B. D. The measure of order and disorder in the distribution of species in fragmented habitat. *Oecologia* **96**, 373–382 (1993).

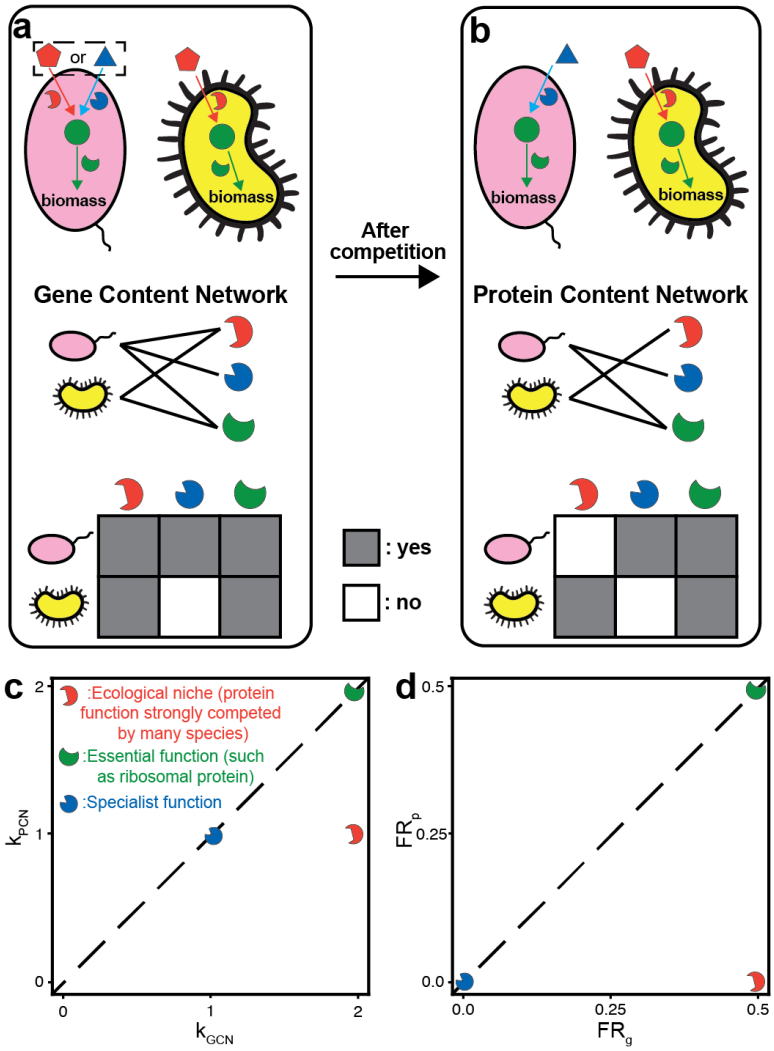

Figure 1: **Protein functions involved in determining ecological niches are postulated to have larger discrepancies between the gene-level functional redundancy FR_g and protein-level functional redundancy FR_p.** Here we use a hypothetical example with three representative proteins (3 broken circles with complementary shapes to their substrates) to demonstrate this point. **a,** Schematic of 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 clusters: specialist function, essential function, and niche function. In the calculation of FR_g and FR_p, we we assume equal abundances of the two species, i.e., $p_1 = p_2 = 0.5$.

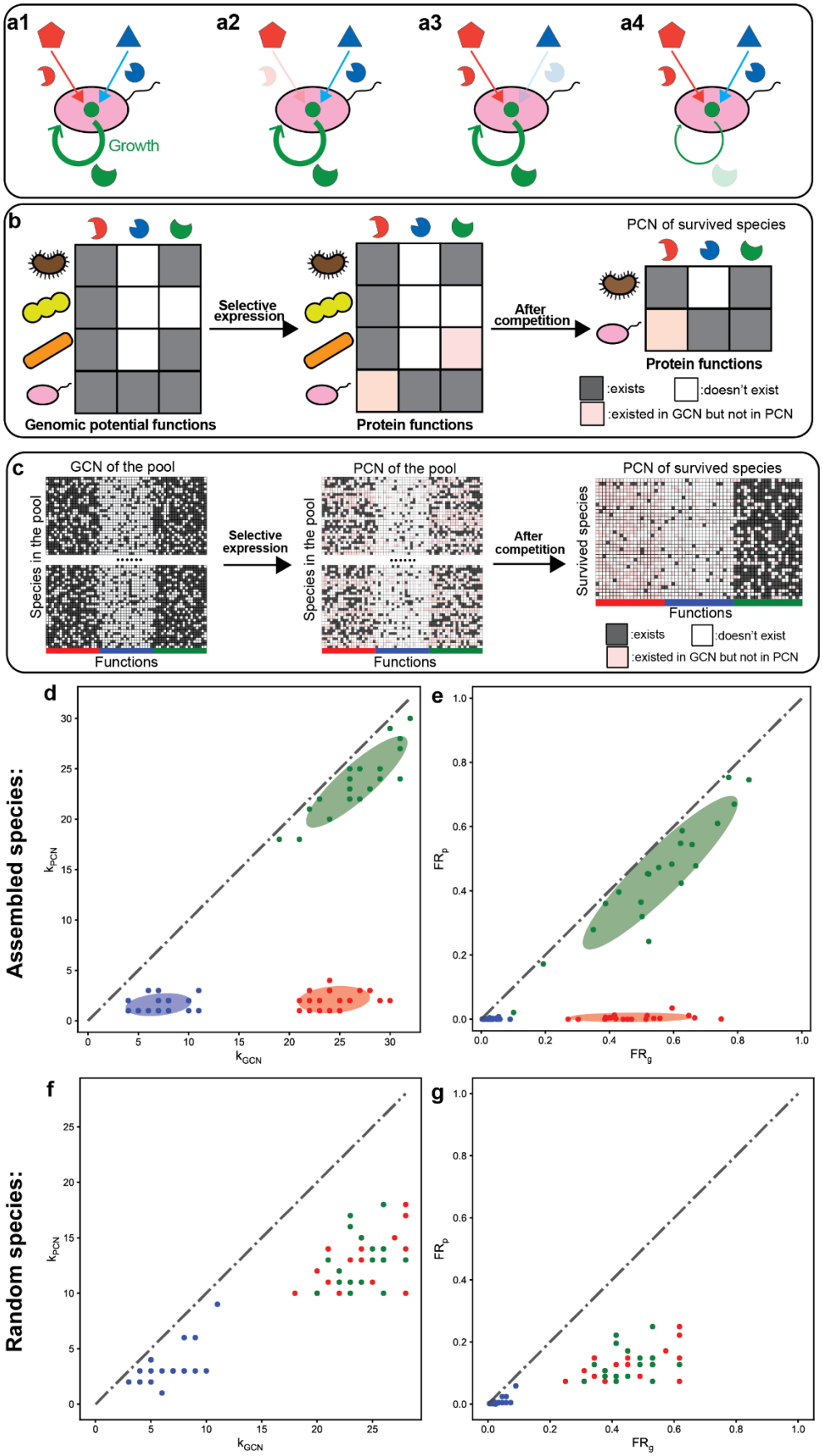
a1
a2
a3
a4
Growth
b
Selective expression
After competition
PCN of survived species
Genomic potential functions
Protein functions
Protein functions
:exists
:doesn't exist
:existed in GCN but not in PCN
c
GCN of the pool
PCN of the pool
PCN of survived species
Species in the pool
Survived species
Functions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

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 protein 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.

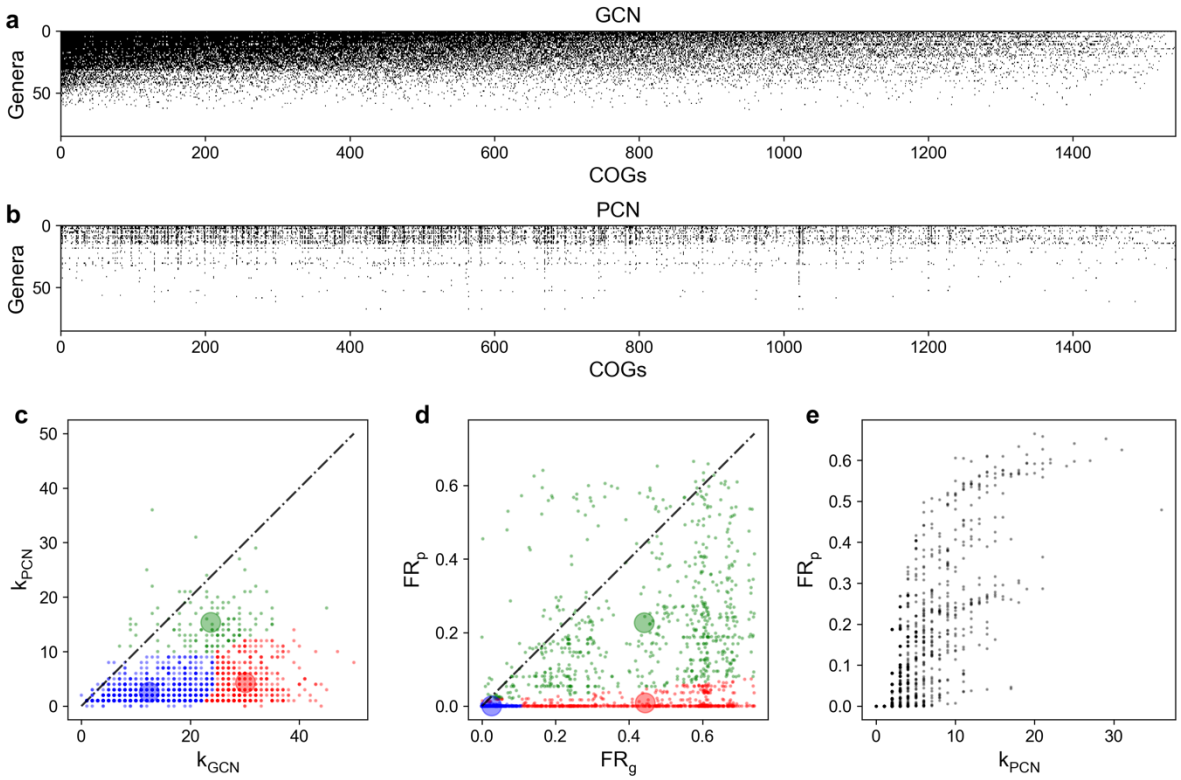

Figure 3: R**eal data of the human gut microbiome showing three clusters on the plot that compares FR_g with FR_p.** Metagenome and metaproteome of subject HM454 mucosal-luminal interface samples[28] were used to construct GCN and PCN, respectively. **a,** The GCN shows if a genus owns (or doesn't own) 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.

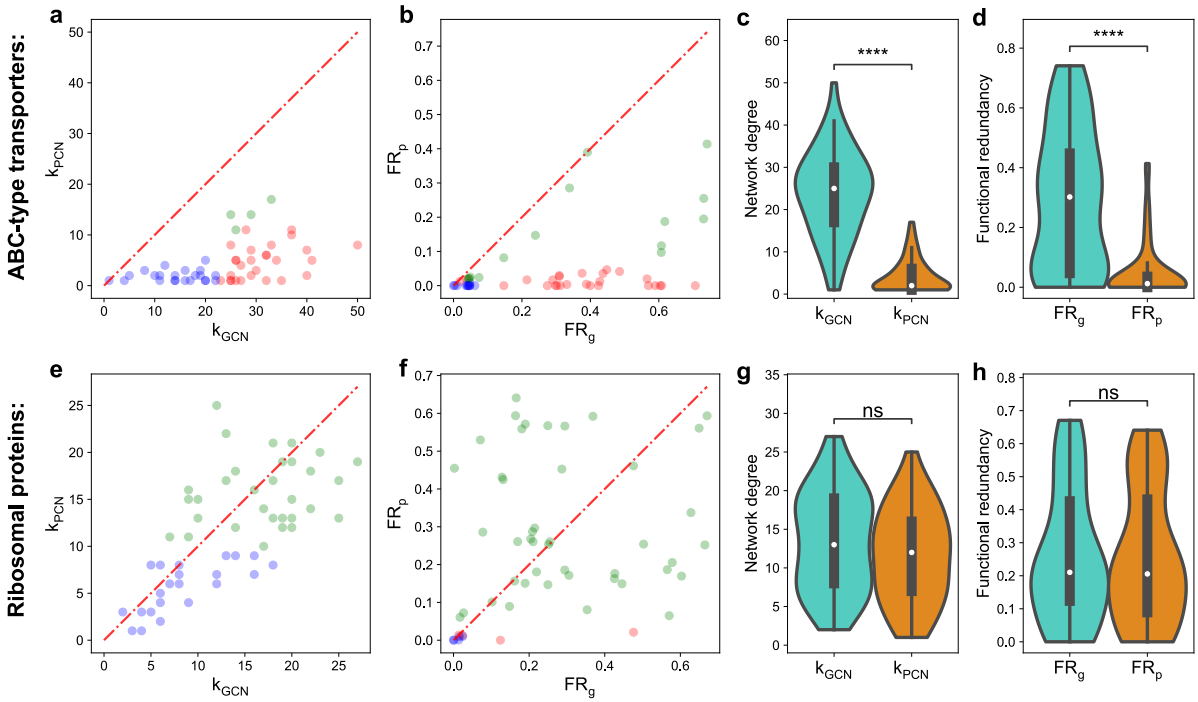

Figure 4: **Comparison of network degree and functional redundancy between the gene and protein level for ABC-type transporters and ribosomal proteins. a,** Network degrees in GCN are larger than network degrees in PCN for most ABC-type transporter COGs. k_{GCN} (or k_{PCN}) is the network degree of 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 (violin plots and boxplots) for ABC-type transporter COGs show a significantly huge reduction from k_{GCN} to k_{PCN} or from FR_g to FR_p. **e,** Network degrees in GCN are comparable with that in PCN for most ribosomal protein COGs. **f,** FR_g is comparable with FR_p for most ribosomal protein COGs. Points in scatter plots are colored by the same colors used in Fig. 3d. **g-h,** The distribution of network degrees and functional redundancies (violin plots and boxplots) for ribosomal protein COGs show no significant reduction from k_{GCN} to k_{PCN} or from FR_g to FR_p. In all boxplots, the middle white dot is the median, the lower and upper hinges correspond to the first and third quartiles, and the black line ranges from the 1.5 × IQR (where IQR is the interquartile range) below the lower hinge to 1.5 × IQR above the upper hinge. All violin plots are smoothed by a kernel density estimator and 0 is set as the lower bound. All statistical analyses were performed using the two-sided Mann-Whitney-Wilcoxon U Test with Bonferroni correction between genomic capacity (GCN) and protein functions (PCN). P values obtained from the test is divided into 5 groups: (1) $p > 0.05$ (ns), (2) $0.01 < p \leq 0.05$ (*), (3) $10^{-3} < p \leq 0.01$ (**), (4) $10^{-4} < p \leq 10^{-3}$ (***), and (5) $p \leq 10^{-4}$ (****). Network degree comparison of ABC transporters: $p = 7.11 \times 10^{-16}$. Network degree comparison of ribosomal proteins: proteins: $p = 0.10$. Redundancy comparison of ABC transporters: $p = 2.19 \times 10^{-11}$. Redundancy comparison of ribosomal proteins: $p = 1.00$.

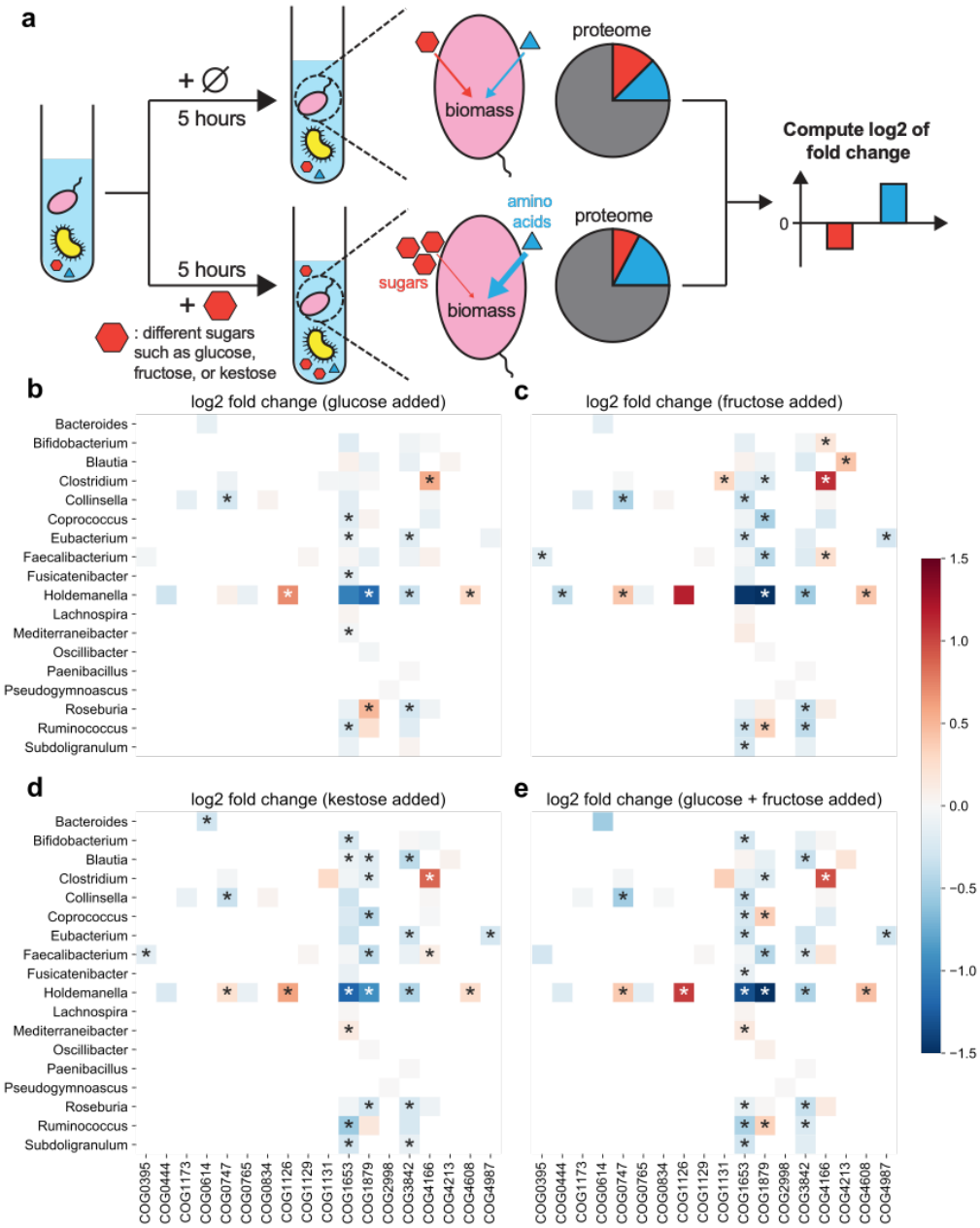

Figure 5: **Microbes modify their expression for ABC-type transporters to adapt to added sugars.** All heatmaps share the same color bar on the right. **a,** Schematic of in-vitro cultures of a collected human gut microbiome. In the treatment group, one sugar is added to the community. Metaproteomic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