Deciphering Functional Redundancy in the Human Microbiome

Liang Tian[1], Ang-Kun Wu[1], Jonathan Friedman[2], Matthew K. Waldor[3,4], Scott T. Weiss[1] & Yang-Yu Liu[1, 5*]

[1]Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA

[2]Physics of Living Systems, Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

[3]Division of Infectious Diseases, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA

[4]Howard Hughes Medical Institute, Boston, Massachusetts 02115, USA

[5]Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA

August 14, 2017

* Corresponding author: yyl@channing.harvard.edu

Abstract

The taxonomic composition of the human microbiome varies tremendously across individuals, while its gene composition or functional capacity is highly conserved across individuals[1-5]. Such a strikingly conserved functional capacity implies an ecological property known as *functional redundancy*[6-9]. Although this redundancy is thought to underlie the stability and resilience of the human microbiome[10, 11], its origin is elusive. Here, we decipher the functional redundancy in the human microbiome by depicting its genomic content network --- a bipartite graph that links microbes to the genes in their genomes. We find that this network exhibits several topological features, such as highly-nested structure and very heterogeneous gene degree distribution, which favor high functional redundancy. Furthermore, we find that the human gut microbiota displays a unique assembly pattern --- a functional core-periphery structure, which enables it to reach a higher functional redundancy than other body sites. This unique assembly pattern and the associated high functional redundancy of the gut microbiota are gradually established during the first year of life in healthy infants. Acute disturbances (such as in *Clostridium difficile* infection) abolish the assembly pattern and the functional redundancy weakens, but several chronic diseases, including inflammatory bowel diseases, asthma, obesity, autism, and Parkinson's disease, do not markedly alter the assembly pattern and the functional redundancy of the gut microbiota. This result suggests that the functional redundancy can serve as a resilience indicator of diseased microbiota against compositional interventions and hence predict the efficacy of microbiome-based therapies such as fecal microbiota transplantation. Our framework, based on the genomic content network of the human microbiome, paves the way for a mechanistic understanding of species-function relationships, which is critical for developing functional-based diagnostics and therapeutics[12].

As a classical concept in ecology, functional redundancy (FR) means that phylogenetically unrelated species perform similar functions in ecosystems so that they can be interchanged with little impact on overall ecosystem functioning[6-9]. Compelling evidence of strong FR in the human microbiome has been uncovered recently[11, 13]. For example, in the human gut microbiota strong FR was found in the processing of dietary carbohydrates[14], the modification of bile acids[15], and the promotion of interleukin secretion[16, 17]. Moreover, several metagenomic studies reported that the carriage of microbial taxa varies tremendously within healthy populations, while microbiome gene compositions or functional profiles remain remarkably similar across individuals[1-5]. In fact, if we simply plot the function richness vs. taxon richness for human microbiome samples, we see a positive but decelerating relationship (i.e., the relationship is steep at low taxon richness, and levels off quickly as taxon richness increases further) for each body site, a clear evidence of FR in the human microbiome (SI Sec.1, Fig.S1). These results suggest that changing the taxonomic composition will not drastically alter the genetic potential or functional capacity of the human microbiome. Hence, the strong FR establishes the basis for the stability and resilience of the human microbiome in response to disturbance[10, 11].

The origin of the remarkable FR in the human microbiome has not been well understood. From the host perspective, although strong FR does not necessarily imply that the host is regulating the diversity of microbiota to promote FR[18], natural selection could favor functionally stable communities with FR encoded in their genomes[19]. From the microbial perspective, however, selective pressures could favor niche specialization to avoid inter-species competition, thus minimizing functional overlaps between species and leading to low FR within the community[19].

Here we addressed whether there is any organizing principle or assembly rule of the human microbiome to achieve both FR and functional diversity (FD). In particular, we borrowed tools from network science[20] to construct the *genomic content network* (GCN) of the human microbiome, and then study its network properties that determine the FR and FD of microbiome samples.

Consider a pool of N taxa, hereafter referred to as a metacommunity[21], which contains a collection of M genes. The taxonomic profile $\boldsymbol{p}^{(\nu)} = \left[p_1^{(\nu)}, \cdots, p_N^{(\nu)}\right]$ of a local community ν (i.e., a microbiome sample from a particular body site of a human subject ν) can be directly related to

its gene composition or functional profile $\boldsymbol{f}^{(\nu)} = \left[f_1^{(\nu)}, \cdots, f_M^{(\nu)}\right]$ through the GCN of the metacommunity (Fig.1a-c). Here, we define the GCN as a weighted bipartite graph connecting those taxa to their genes. The GCN can be represented by an $N \times M$ incidence matrix $\boldsymbol{G} = (G_{ia})$ with a non-negative integer G_{ia} indicates the copy number of gene-a in the genome of taxon-i (Fig.1b). The functional profile is given by $\boldsymbol{f}^{(\nu)} = c\boldsymbol{p}^{(\nu)}\,\boldsymbol{G}$, where c is a normalization constant (see SI Sec.2 for details).

A key advantage of GCN is that it enables us to calculate the FR for each local community (i.e., the within-sample alpha functional redundancy, hereafter denoted as FR_α). In the literature of ecology, the FR of a local community is often interpreted as the part of its alpha taxonomic diversity (TD_α) that cannot be explained by its alpha functional diversity (FD_α)[22-24], i.e., $\mathrm{FR}_\alpha \equiv \mathrm{TD}_\alpha - \mathrm{FD}_\alpha$. In Ref.[22, 23], TD_α is chosen to be the Gini-Simpson index: $\mathrm{GSI} \equiv 1 - \sum_{i=1}^{N} p_i^2$, which equals the probability that two members chosen at random from the local community (with replacement) belong to different taxa; and FD_α is chosen to be Rao's quadratic entropy $Q \equiv \sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}p_ip_j$, a classical alpha diversity measure that characterizes the mean functional distance between any two randomly chosen members in the local community. Here $d_{ij} = d_{ji} \in [0, 1]$ denotes the functional distance between taxon-i and taxon-j, which can be calculated as the weighted Jaccard distance between the genomes of the two taxa (see SI Sec.3 for other definitions of d_{ij}). By definition $d_{ii} = 0$ for $i = 1, \cdots, N$. Note that with $\mathrm{TD}_\alpha = \mathrm{GSI}$ and $\mathrm{FD}_\alpha = Q$, we can derive $\mathrm{FR}_\alpha = \sum_{i=1}^{N}\sum_{j\neq i}^{N}(1 - d_{ij})p_ip_j$, representing the functional similarity (or overlap) of two randomly chosen members in the local community. Of course, one can also use other definitions for TD_α and FD_α. Then the expression of FR_α will be different (see SI Sec.4 for details).

The FR_α of each local community (or microbiome sample) is closely related to the system-level FR quantified over a collection of samples. Consider two extreme cases. (1) Each taxon is completely specialized and has its own unique genome (Fig.1b1), hence $d_{ij} = 1$ for any $i \neq j$. In this case, for each sample we have $\mathrm{FD}_\alpha = \mathrm{TD}_\alpha$ and $\mathrm{FR}_\alpha = 0$. The functional profiles vary drastically across samples (Fig. 1c1). Correspondingly, when plotting the function richness (N_f) vs. taxon richness (N_t) across all the samples, we roughly have a linear relationship: $N_\mathrm{f} \sim N_\mathrm{t}$, indicating the lack of system-level FR (Fig.1d1). (2) All taxa share exactly the same genome (Fig.1b3), rendering $d_{ij} = 0$ for all i and j. In this case, for each sample we have

$\mathrm{FD}_\alpha = 0$ and $\mathrm{FR}_\alpha = \mathrm{TD}_\alpha$. The function profiles are exactly the same for all samples (Fig.1c3) Correspondingly, we have constant N_f for any $N_\mathrm{t} > 0$, suggesting the strongest system-level FR (Fig.1d3).

The above two extreme scenarios are unrealistic. An intermediate scenario is as follows. The GCN has certain structural features such that different taxa share a few common functions, but certain taxa are specialized to perform some unique functions (Fig.1b2). In this case, the FD_α and FR_α of each sample can both be high. Moreover, the functional profiles can be highly conserved across samples (Fig.1c2), and the whole collection of samples will display strong FR in the sense that $N_\mathrm{f} \sim N_\mathrm{t}^a$ with $0 < a < 1$, i.e., a positive but decelerating relationship between N_f and N_t (Fig.1d2).

To test if this intermediate scenario is true for the human microbiome, we constructed a reference GCN using the Integrated Microbial Genomes (IMG) database[25]. The IMG database used here includes in total 2,833 strains and 6,727 KEGG Orthologs (KOs) (see SI Sec.5). Here each KO is a group of genes representing functional orthologs in the molecular networks[26]. In order to reduce the culturing and sequencing bias for certain species (e.g., *Escherichia coli*), we randomly chose a representative strain (genome) for each species, which results in a reference GCN of 1,386 species and 6,685 KOs. This GCN is depicted in Fig.2a, where for visualization purposes each taxon node represents a phylum and each function node represents a KEGG super-pathway. Several subgraphs of this GCN (associated with three phyla: Actinobacteria, Bacteroidetes, and Firmicutes; and three KEGG super-pathways: Carbohydrate Metabolism, Immune System Diseases, and Cardiovascular Diseases) are shown in SI Sec.1, Fig.S2.

In order to characterize the structure of this reference GCN, we systematically analyzed its network properties at the species-KO level. We first visualized its incidence matrix (Fig.2b), where the presence (or absence) of a link connecting a species and a KO is colored in yellow (or blue), respectively. The matrix is organized appropriately to emphasize its highly-nested structure[27-29], i.e., the KOs of those species in the lower rows (with smaller genome size) tend to be subsets of KOs for those species in the higher rows (with larger genome size). (See SI Sec.7 for the detailed quantification of nestedness.) We then calculated the functional distances among different species, finding a unimodal distribution with the peak centered around 0.7 (Fig.2c). We finally calculated the unweighted degree distributions of taxon nodes (species) and function nodes (KOs). Here, the unweighted degree of a species is just the number of distinct KOs in its

genome, and the unweighted degree of a KO is the number of species whose genomes contain this KO. We found that the unweighted degrees of species nodes follow a Poisson-like distribution, implying that in general species contain very similar numbers of distinct KOs (Fig.2e). By contrast, the unweighted degree distribution of KOs is highly heterogeneous and displays a fat tail, indicating that most KOs are specialized and only exist in the genomes of very few species, and a few housekeeping KOs appear in almost every species' genome to maintain basic cellular functions (Fig.2d). We also analyzed the reference GCN constructed from another database: MBGD (Microbial Genome Database for Comparative Analysis)[30], finding very similar network properties (SI Sec.1, Fig. S3a).

When analyzing human microbiome samples, many taxa do not appear in the IMG database. Their genomic contents can be inferred from closely related reference genomes[31] (see SI Sec.5 for details). This yields an extended reference GCN, which contains 203,452 Operational Taxonomic Units (OTUs) and 6,909 KOs, and displays very similar network properties as the original reference GCN (SI Sec.1, Fig.S3b).

Leveraging this extended reference GCN, we computed FR_α and FD_α for microbiome samples from two large-scale metagenomic studies: the Human Microbiome Project (HMP)[1, 32] and the Student Microbiome Project (SMP)[33]. The GCNs associated with different body sites display similar network properties as the original reference GCN (SI Sec.1, Fig.S4-S6). Interestingly, we found that in both HMP and SMP studies and for most body sites FR_α and FD_α are generally comparable for each sample (Fig.3a-d, black boxes).

To identify the structural features of the GCN that determine FR_α and FD_α, we randomized the GCN using four different randomization schemes, yielding four different null models (SI Sec.6). Then we recalculated FR_α and FD_α for each sample (Fig. 3a-d, colored boxes), finding that for all the body sites examined all the four different null models yield lower FR_α and higher FD_α than those calculated from the real GCN. Analyzing the network properties of those null models (SI Sec.1, Fig.S7), we found that those randomized GCNs all display lower nestedness and higher $\langle d_{ij} \rangle$ than those of the real GCN. Hence, the highly-nested structure and low $\langle d_{ij} \rangle$ of the real GCN contribute to the FR_α and FD_α values of those microbiome samples. Moreover, for the first two null models (Null-1 and Null-2, where both the highly-nested structure and high gene degree heterogeneity of the real GCN are destroyed), FR_α is much lower (and FD_α is much higher) than those of the other two null models (Null-3 and Null-4, where the

highly-nested structure is destroyed, but the high gene degree heterogeneity is kept). This suggests that the high gene degree heterogeneity also contribute to the FR_α and FD_α values of those microbiome samples. Hence the GCN exhibits at least three different structural features (highly-nested structure, low $\langle d_{ij} \rangle$, and heterogeneous gene degree distribution) that jointly determine the FR_α and FD_α values of microbiome samples. This result does not depend on the detailed definitions of d_{ij}, FR_α, and FD_α (see SI Sec.1, Fig S8-S10).

Note that for each microbiome sample if we preserve the taxonomic profile but randomly replace the OTUs by OTUs present in the extended reference GCN, this will typically yield lower FR_α and higher FD_α than those of the original sample (SI Sec.1, Fig.S11). This suggests that the OTUs present in each local community are not assembled at random, but follow certain functional assembly rules[34]. Notably, we found that for almost all body sites analyzed here the OTU functional distance distribution $P(d_{ij})$ displays a similar shape across samples (grey lines in Fig.3e-k). The consistency between the $P(d_{ij})$ of each sample and the overall $P(d_{ij})$ calculated from the union of OTUs in all samples of a given body site (red thick lines in Fig.3e-k) suggests a universal assembly rule --- for each local community the microbes are assembled based on a host-independent $P(d_{ij})$, although the collection of microbes that reside in this body site is highly host-specific (i.e., personalized). The universality of $P(d_{ij})$ is not as apparent for vaginal microbiome samples (Fig.3l), which is consistent with the fact that the functional profiles of the vaginal microbiome also display larger variation across individuals than other body sites[1].

Among all body sites examined, the gut microbiome displays the highest FR_α. This can be partly explained by the unique bimodality in $P(d_{ij})$ of gut microbes at the OTU level (Fig. 3e,h), which was not observed at any other body sites nor in the reference GCN. The left peak (centered around a small d_{ij} value) in the bimodal distribution suggests that certain OTUs in the human gut microbiota have relatively high functional overlap; these OTUs can be considered as a functional "core" of the gut microbiota, as shown in the clustering result of the functional distance matrix (SI Sec.1, Fig.S12a,d, the large blue block in the heatmap). The right peak centered around a large d_{ij} is due to large functional differences either among those "periphery" OTUs or between the "core" and the "periphery" OTUs (SI Sec.1, Fig. S12a,d, red part in the heatmap).

Many studies of the human gut microbiota during early development have shown that there is a systematic turnover of species until a stable adult state is reached[35-38]. In early life, the gut microbiota of healthy individuals is usually dominated by bifidobacteria[39, 40]; in contrast, in the adult gut microbiota there are 1,183~3,180 bacterial phylotypes[41]. It has been conjectured that FR in the infant gut may be much higher than that found in the adult gut[11]. To test this conjecture, we analyzed samples from a longitudinal study of infant gut microbiome in the first two years of life[42]. We found that FR_α of newborns displays large variation with median much lower than that of mothers; gradually, the median FR_α reaches a relatively stable value with small variations by the time infants reach 14 months old (Fig. 4a). Roughly at the same time, $P(d_{ij})$ displays strong bimodality (Fig. 4b). When the infants become older, their $P(\mathrm{FR}_\alpha)$, $P(\mathrm{FD}_\alpha)$, and $P(d_{ij})$ become more and more stable and similar to their mothers.

It is worth emphasizing that the primary succession of the infant gut microbiota has a universal feature: at first, functionally more different taxa (corresponding to the initial single peak in $P(d_{ij})$ centered around a high d_{ij} value) colonize the infant gut, ensuring high functional diversity of the microbiota. Subsequently, functionally more similar taxa (corresponding to the emerging additional peak in $P(d_{ij})$ centered around a low d_{ij} value) colonize, resulting in high functional redundancy of the gut microbiota. This is also consistent with the observation that the peak of $P(\mathrm{FD}_\alpha)$ almost doesn't change over time, while the peak of $P(\mathrm{FR}_\alpha)$ gradually shifts towards higher value by the time infants reach 14 months (Fig. 4a). Using samples from an independent study of the infant gut microbiome in the first five years of life[43], we obtained similar results (Fig.4c-d).

It has been postulated that microbiota-altering disease may be accompanied by a loss of FR over time[11]. To test this hypothesis, we analyzed the gut microbiome samples from subjects with *C. difficile* infection (CDI)[44] before and after treatment with fecal microbiota transplantation (FMT), subjects with inflammatory bowel disease (IBD)[45], asthma[43], obesity[2], autism[46], or Parkinson's disease (PD)[47], as well as their respective healthy controls. Stool samples of CDI patients before FMT treatment exhibited a much lower FR_α median value than that calculated for Post-FMT samples or samples from healthy donors (Fig.5a). In contrast, their FD_α values do not differ significantly from those determined for Post-FMT samples or samples from healthy donors (Fig.5a). The relatively low FR_α of CDI patients is expected to reduce the resilience of their gut microbiota against external perturbation, potentially facilitating the efficacy of FMT in restoring

a healthy gut microbiota. Notably, FMT treatment also restored the bimodality in $P(d_{ij})$ that was not observed in CDI patients (Fig.5b).

Unlike those of CDI patients, the stool samples from IBD, asthma, obesity, autism, and PD patients generally do not differ significantly from their respective healthy controls in their FR_{α} values and $P(d_{ij})$ (Fig.5c-l), although some reduction in FR_{α} values was seen in samples from mucosal tissue biopsies (rectum and terminal ileum) of IBD patients (Fig.5c). Thus, in contrast to CDI patients, the gut microbiomes of IBD, asthma, obesity, autism, or PD patients have been driven to relatively stable diseased states, which are resilient to compositional transitions and hence less likely to be susceptible to FMT that "engineers" the microbial community as a whole. FMT may have the most promise in countering IBD, which showed some reduction in FR_{α} values.

Targeting the human microbiota for therapeutic intervention holds great promise[12, 48, 49]. However, the design of effective microbiome-targeted therapies for complex diseases requires deep understanding of how the microbiota changes both structurally and functionally in response to various perturbations[50-52]. The framework established here, based on the analysis of GCN, sheds light on this direction. Indeed, from a translational perspective, our ability to quantify the resilience of the human microbiota against disturbance using the within-sample FR could offer significant potential clinical benefits[49]. The GCN framework can also be applied to general microbial communities beyond human microbiome[53-56]. We anticipate that it will inspire more studies on the biodiversity-ecosystem functioning relationship from a completely new angle.

Figure Captions

Fig.1: Structure of the genomic content network is crucial in determining the functional redundancy and functional diversity of microbial communities. Here we use hypothetical examples to demonstrate this point. **a**, The taxonomic profiles vary drastically across many local communities (microbiome samples). **b**, Genomic content networks are bipartite graphs that connect taxa to the genes in their genomes. Here, the left-hand side nodes (circles) represent different taxa and the right-hand side nodes (squares) represent different genes. The edge weight represents the gene copy number. **b1**, Each taxon has a unique genome. **b2**, Different taxa share a few common genes, some taxa are specialized to have some unique genes. **b3**, All taxa share exactly the same genome. **c**, The functional profiles can be calculated from the taxonomic profile in (a) and the genomic content network in (b). **c1**, The functional profiles vary drastically across samples. For each local community (sample), the functional diversity is maximized while the functional redundancy is minimized. **c2**, The functional profiles are highly conserved across samples. The within-sample functional diversity and redundancy are comparable. **c3**, The functional profiles are exactly the same across samples. For each sample, the functional diversity is minimized while the functional redundancy is maximized. **d**, The number of genes (function richness) vs. the number of taxa (taxon richness) in different local communities (samples). **d1**, The function richness increases linearly with the taxon richness, indicating zero functional redundancy at the system level. **d2**, There is a positive but decelerating relationship between function richness and taxon richness, indicating strong functional redundancy at the system level. **d3**, The function richness is constant for any nonzero taxon richness, implying complete functional redundancy at the system level.

Fig.2: The reference genomic content network constructed from the Integrated Microbial Genomes database has highly-nested structure and very heterogeneous gene degree distribution. a, The reference GCN is visualized at the phylum level for taxon nodes and the KEGG super-pathway level for function nodes. The bar height of each phylum corresponds to the average genome size of those species belonging to that phylum. The thickness of a link connecting a phylum and a KEGG super-pathway is proportional to the number of KOs in the phylum belonging to that pathway. For each pathway, we draw a dotted box to cover those links that start from different phyla and end at this pathway. In SI Sec.1, Fig.S2, we show several subgraphs of this GCN associated with a few phyla and pathways. An animated version of this GCN that clearly shows the links for each phylum (or pathway) can be found in the Supplementary File (RGCN.html). **b**, The incidence matrix of the reference GCN is shown at the species-KO level, where the presence (or absence) of a link between a species and a KO is colored in yellow (or blue), respectively. We organized this matrix using the Nestedness Temperature Calculator to emphasize its nested structure[27]. The nestedness value (~0.3655) of this network is calculated based on the classical NODF measure[28] (see SI Sec.7). **c**, The probability distribution of functional distances (d_{ij}) among different species. The bin size is 0.01. **d**, The unweighted KO degree distribution. Here, the unweighted degree of a KO is the number of species whose genomes contain this KO. **e**, The unweighted species degree distribution. Here, the unweighted degree of a species is the number of distinct KOs in its genome. See SI Sec.5 for the details of constructing this reference GCN.

Fig.3: The high functional redundancy in the human microbiome is due to several structural features of the genomic content network. Microbiome samples from the SMP study with 4 different body sites[33], as well as the HMP study with 18 different body sites[1, 32], were analyzed. See SI Sec.8 for detailed description of the two datasets. **a-b**, Functional redundancy. **c-d**, Functional diversity. The box plots of function redundancy and function diversity of those samples were calculated from the real GCN (black box), as well as the randomized GCNs (colored boxes) using four different randomization schemes: complete randomization (blue); OTU-degree preserving randomization (red); KO-degree preserving randomization (green); OTU- and KO-degree preserving randomization (yellow). Here the KO-degree is the sum of copy numbers of this KO in those genomes that contain it, and the OTU-degree is the sum of copy numbers of those KOs in this OTU's genome. See SI Sec.6 for details. **e-l**, The distribution of functional distances (d_{ij}) among different OTUs. Thin gray lines: $P(d_{ij})$ for different samples. Thick red lines: $P(d_{ij})$ calculated from the union of OTUs present in all the samples for each body site.

Fig.4: The infant gut microbiota gradually gains functional redundancy through a universal succession pattern. Microbiome samples are from two longitudinal studies of infant gut: (**a-b**) during the first two years of life[42]; (**c-d**) during the first five years of life[43]. See SI Sec.8 for detailed description of the two datasets. (**a,c**), The distributions of within-sample functional redundancy (blue), functional diversity (yellow), and taxonomic diversity (gray) of gut microbiome samples from infants at different ages and their mothers. (**b,d**), The OTU functional distance distribution of gut microbiome samples from infants at different ages and their mothers. Thin gray lines: $P(d_{ij})$ for different samples. Thick red lines: $P(d_{ij})$ calculated from the union of OTUs present in all the samples at each time point.

Fig.5: Disrupted gut microbiota associated with complex diseases can display similar functional redundancy as the healthy controls. The box plots of function redundancy (blue), function diversity (yellow); and OTU functional distance distribution (thin gray lines: $P(d_{ij})$ for different samples; thick red lines: $P(d_{ij})$ calculated from the union of OTUs present in all the samples) for microbiome samples of patients with the following diseases: (**a-b**), Recurrent *Clostridium difficile* infection[44]. (**c-d**), Inflammatory bowel disease[45]. (**e-f**), Asthma[43]. (**g-h**), Obesity[2]. (**i-j**), Autism[46]. (**k-l**), Parkinson's disease[47]. The results of their respective healthy controls are also shown. Significance levels: FDR<0.05 (*), <0.01(**), <0.001(***); >0.05 (NS, non-significant); Mann-Whitney *U* test. See SI Sec.8 for detailed description of those datasets.

References:

1. HMP Consortium Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207-214 (2012).
2. Turnbaugh, P.J. et al. A core gut microbiome in obese and lean twins. *Nature* **457**, 480-484 (2009).
3. Franzosa, E.A. et al. Identifying personal microbiomes using metagenomic codes. *Proc Natl Acad Sci USA* **112**, E2930-E2938 (2015).
4. Qin, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59-65 (2010).
5. Lozupone, C.A., Stombaugh, J.I., Gordon, J.I., Jansson, J.K. & Knight, R. Diversity, stability and resilience of the human gut microbiota. *Nature* **489**, 220-230 (2012).
6. Lawton, J.H. & Brown, V.K. in Biodiversity and Ecosystem Function. (eds. E.-D. Schulze & H.A. Mooney) 255-270 (Springer Berlin Heidelberg, Berlin, Heidelberg; 1994).
7. Loreau, M. Does functional redundancy exist? *Oikos* **104**, 606-611 (2004).
8. Hubbell, S.P. Neutral theory in community ecology and the hypothesis of functional equivalence. *Functional Ecology* **19**, 166-172 (2005).
9. Allison, S.D. & Martiny, J.B.H. Resistance, resilience, and redundancy in microbial communities. *Proc Natl Acad Sci USA* **105**, 11512-11519 (2008).
10. Sommer, F., Anderson, J.M., Bharti, R., Raes, J. & Rosenstiel, P. The resilience of the intestinal microbiota influences health and disease. *Nat Rev Micro* **advance online publication** (2017).
11. Moya, A. & Ferrer, M. Functional Redundancy-Induced Stability of Gut Microbiota Subjected to Disturbance. *Trends in Microbiology* **24**, 402-413 (2016).
12. Lemon, K.P., Armitage, G.C., Relman, D.a. & Fischbach, M.a. Microbiota-targeted therapies: an ecological perspective. *Science Translational Medicine* **4**, 137rv135 (2012).
13. Vieira-Silva, S. et al. Species–function relationships shape ecological properties of the human gut microbiome. *Nature Microbiology* **1**, 16088 (2016).

14. Ferrer, M. et al. Microbiota from the distal guts of lean and obese adolescents exhibit partial functional redundancy besides clear differences in community structure. *Environmental Microbiology* **15**, 211-226 (2013).
15. Labbé, A., Ganopolsky, J.G., Martoni, C.J., Prakash, S. & Jones, M.L. Bacterial Bile Metabolising Gene Abundance in Crohn's, Ulcerative Colitis and Type 2 Diabetes Metagenomes. *PLOS ONE* **9**, e115175 (2014).
16. Sokol, H. et al. Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc Natl Acad Sci USA* **105**, 16731-16736 (2008).
17. Morgan, X.C. et al. Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. *Genome Biology* **16**, 67 (2015).
18. Foster, K.R., Schluter, J., Coyte, K.Z. & Rakoff-Nahoum, S. The evolution of the host microbiome as an ecosystem on a leash. *Nature* **548**, 43-51 (2017).
19. Ley, R.E., Peterson, D.a. & Gordon, J.I. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* **124**, 837-848 (2006).
20. Albert, R. & Barabási, A.-L. Statistical mechanics of complex networks. *Reviews of Modern Physics* **74**, 47-97 (2002).
21. Costello, E.K., Stagaman, K., Dethlefsen, L., Bohannan, B.J.M. & Relman, D.a. The application of ecological theory toward an understanding of the human microbiome. *Science* **336**, 1255-1262 (2012).
22. de Bello, F. et al. Importance of species abundance for assessment of trait composition: an example based on pollinator communities. *Community Ecology* **8**, 163-170 (2007).
23. Pillar, V.D. et al. Functional redundancy and stability in plant communities. *Journal of Vegetation Science* **24**, 963-974 (2013).
24. Kang, S. et al. Functional redundancy instead of species redundancy determines community stability in a typical steppe of inner Mongolia. *PLoS ONE* **10**, 1-11 (2015).
25. Markowitz, V.M. et al. IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Research* **40**, D115-D122 (2012).
26. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Research* **32**, D277-D280 (2004).
27. Atmar, W. & Patterson, B.D. The measure of order and disorder in the distribution of species in fragmented habitat. *Oecologia* **96**, 373-382 (1993).
28. Almeida-Neto, M., Guimara˜es, P., Guimara˜es Jr, P.R., Loyota, R.D. & Ulrich, W. A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. *Oikos* **117**, 1227-1239 (2008).
29. Saavedra, S., Reed-tsochas, F. & Uzzi, B. A simple model of bipartite cooperation for ecological and organizational networks. *Nature* **457**, 436-466 (2009).
30. Uchiyama, I., Higuchi, T. & Kawai, M. MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity. *Nucleic Acids Research* **38**, D361-D365 (2010).
31. Langille, M.G.I. et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology* **31**, 814-821 (2013).

32. HMP Consortium A framework for human microbiome research. *Nature* **486**, 215-221 (2012).
33. Flores, G.E. et al. Temporal variability is a personalized feature of the human microbiome. *Genome Biology* **15**, 531 (2014).
34. Shafquat, A., Joice, R., Simmons, S.L. & Huttenhower, C. Functional and phylogenetic assembly of microbial communities in the human microbiome. *Trends in Microbiology* **22**, 261-266 (2014).
35. Palmer, C., Bik, E.M., DiGiulio, D.B., Relman, D.a. & Brown, P.O. Development of the human infant intestinal microbiota. *PLoS Biology* **5**, e177 (2007).
36. Koenig, J.E. et al. Succession of microbial consortia in the developing infant gut microbiome. *Proc Natl Acad Sci USA* **108**, 4578-4585 (2011).
37. La Rosa, P.S. et al. Patterned progression of bacterial populations in the premature infant gut. *Proceedings of the National Academy of Sciences*, 6-11 (2014).
38. Yassour, M. et al. Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Science Translational Medicine* **8** (2016).
39. Nylund, L., Satokari, R., Salminen, S. & de Vos, W.M. Intestinal microbiota during early life – impact on health and disease. *Proceedings of the Nutrition Society* **73**, 457-469 (2014).
40. Bäckhed, F. et al. Dynamics and stabilization of the human gut microbiome in the first year of life. *Cell Host & Microbe*, 690-703 (2015).
41. Claesson, M.J. et al. Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proc Natl Acad Sci USA* **108**, 4586-4591 (2011).
42. Bokulich, N.A. et al. Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Science Translational Medicine* **8** (2016).
43. Litonjua, A.A. et al. The Vitamin D Antenatal Asthma Reduction Trial (VDAART): Rationale, design, and methods of a randomized, controlled trial of vitamin D supplementation in pregnancy for the primary prevention of asthma and allergies in children. *Contemporary Clinical Trials* **38**, 37-50 (2014).
44. Youngster, I. et al. Fecal microbiota transplant for relapsing Clostridium difficile infection using a frozen inoculum from unrelated donors: a randomized, open-label, controlled pilot study. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **58**, 1515-1522 (2014).
45. Gevers, D. et al. The Treatment-Naive Microbiome in New-Onset Crohn's Disease. *Cell Host & Microbe* **15**, 382-392 (2014).
46. Kang, D.W. et al. Reduced Incidence of Prevotella and Other Fermenters in Intestinal Microflora of Autistic Children. *PLoS ONE* **8** (2013).
47. Hill-Burns, E.M. et al. Parkinson's disease and Parkinson's disease medications have distinct signatures of the gut microbiome. *Movement Disorders* **32**, 739-749 (2017).
48. Dethlefsen, L., McFall-Ngai, M. & Relman, D.A. An ecological and evolutionary perspective on human–microbe mutualism and disease. *Nature* **449**, 811-818 (2007).
49. Relman, D.A. The human microbiome: ecosystem resilience and health. *Nutrition Reviews* **70 Suppl 1**, S2-9 (2012).

50. Manor, O. & Borenstein, E. Revised computational metagenomic processing uncovers hidden and biologically meaningful functional variation in the human microbiome. *Microbiome* **5**, 19 (2017).
51. Manor, O. & Borenstein, E. Systematic Characterization and Analysis of the Taxonomic Drivers of Functional Shifts in the Human Resource Systematic Characterization and Analysis of the Taxonomic Drivers of Functional Shifts in the Human Microbiome. *Cell Host & Microbe* **21**, 1-14 (2017).
52. Bradley, P.H. & Pollard, K.S. Proteobacteria explain significant functional variability in the human gut microbiome. *Microbiome* **5**, 36 (2017).
53. Jurburg, S.D. & Salles, J.F. in Biodiversity in Ecosystems - Linking Structure and Function. (eds. Y.-H. Lo, J.A. Blanco & S. Roy) Ch. 02 (InTech, Rijeka; 2015).
54. Sunagawa, S. et al. Structure and function of the global ocean microbiome. *Science* **348** (2015).
55. Louca, S. et al. High taxonomic variability despite stable functional structure across microbial communities. *Nature Ecology & Evolution* **1**, 0015 (2016).
56. Louca, S., Parfrey, L.W. & Doebeli, M. Decoupling function and taxonomy in the global ocean microbiome. *Science* **353**, 1272-1277 (2016).

Acknowledgements This work is supported in part by the John Templeton Foundation (Award number 51977). We thank Curtis Huttenhower, Gabriel Billings, Amir Bashan, Yandong Xiao, Weiliang Qiu, Hao Chang, and Yu Chen for valuable discussions. Special thanks to Brigid Davis and Edwin K. Silverman for a careful reading of the manuscript.

Contributions Y.-Y.L conceived and designed the project. L.T., A.-K.W., and Y.-Y.L. did the analytical and numerical calculations. L.T. analyzed all the empirical data. All authors interpreted the results. Y.-Y.L. wrote the manuscript. L.T., M.K.W and S.T.W. edited the manuscript.

Author Information The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to Y.-Y.L. (yyl@channing.harvard.edu).

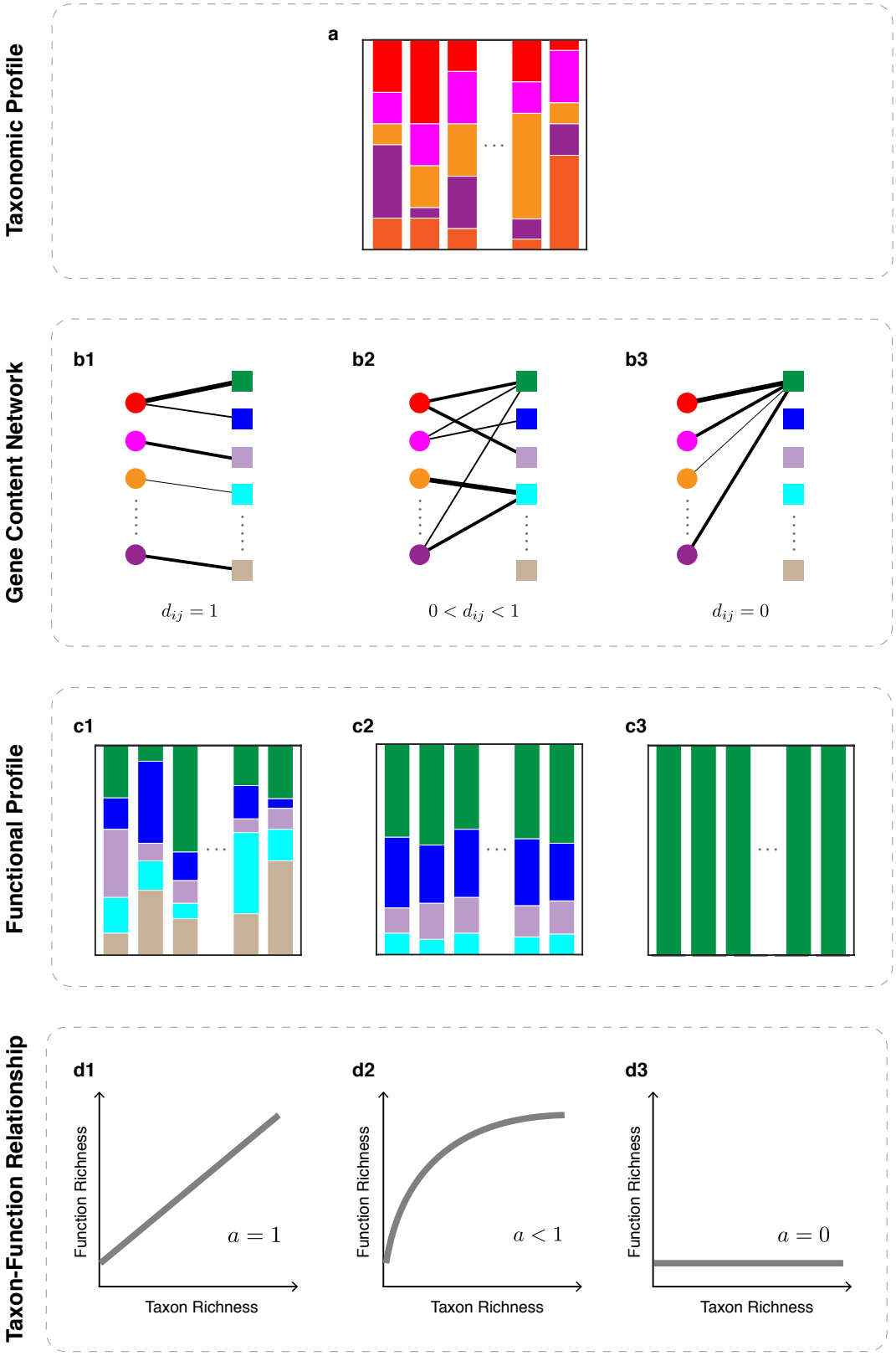
Taxonomic Profile
a
Gene Content Network
b1
b2
b3
$d_{ij} = 1$
$0 < d_{ij} < 1$
$d_{ij} = 0$
Functional Profile
c1
c2
c3
Taxon-Function Relationship
d1
d2
d3
Function Richness
Taxon Richness
$a = 1$
$a < 1$
$a = 0$

FIG. 1:

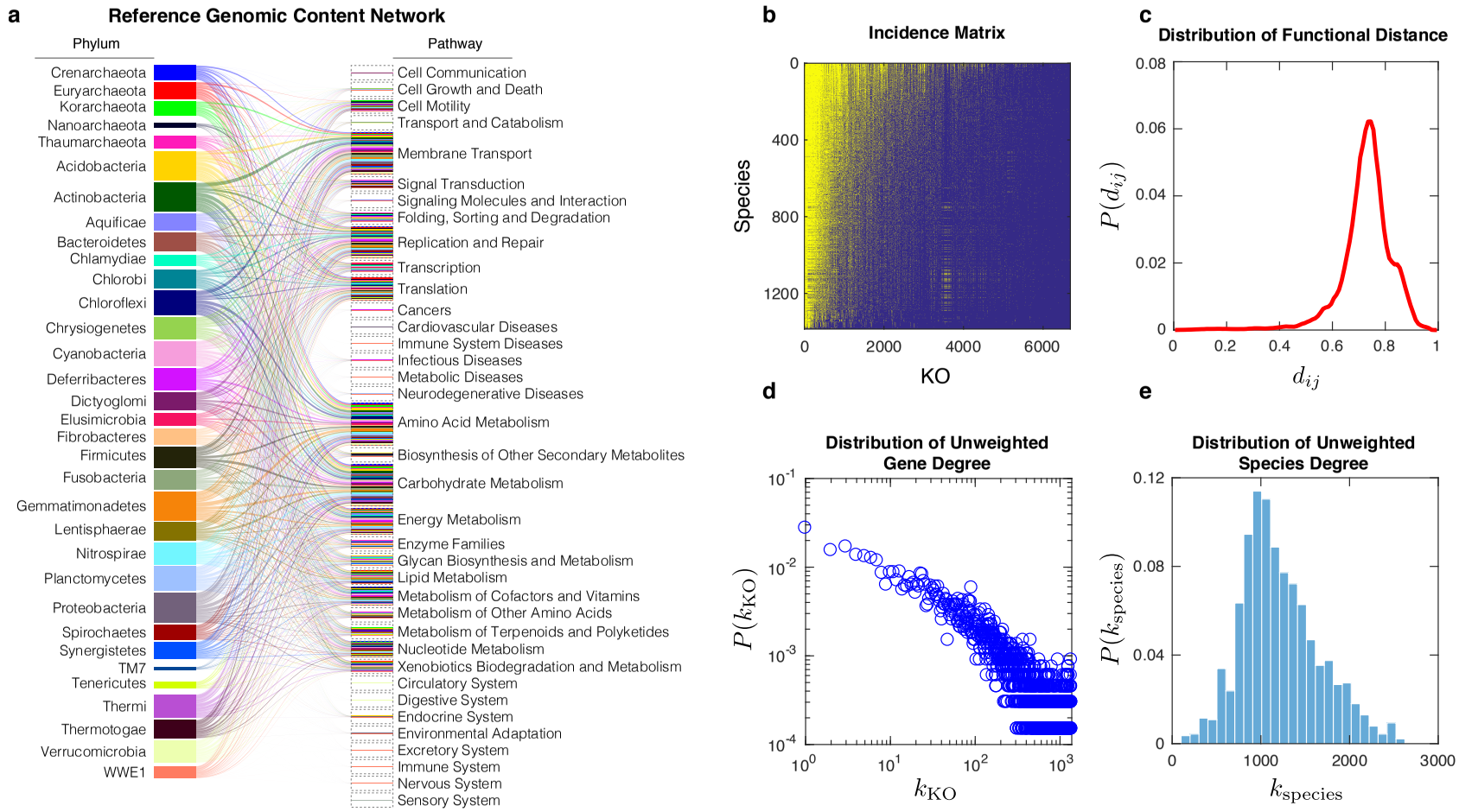
a
Reference Genomic Content Network
Phylum
Crenarchaeota
Euryarchaeota
Korarchaeota
Nanoarchaeota
Thaumarchaeota
Acidobacteria
Actinobacteria
Aquificae
Bacteroidetes
Chlamydiae
Chlorobi
Chloroflexi
Chrysiogenetes
Cyanobacteria
Deferribacteres
Dictyoglomi
Elusimicrobia
Fibrobacteres
Firmicutes
Fusobacteria
Gemmatimonadetes
Lentisphaerae
Nitrospirae
Planctomycetes
Proteobacteria
Spirochaetes
Synergistetes
TM7
Tenericutes
Thermi
Thermotogae
Verrucomicrobia
WWE1
Pathway
Cell Communication
Cell Growth and Death
Cell Motility
Transport and Catabolism
Membrane Transport
Signal Transduction
Signaling Molecules and Interaction
Folding, Sorting and Degradation
Replication and Repair
Transcription
Translation
Cancers
Cardiovascular Diseases
Immune System Diseases
Infectious Diseases
Metabolic Diseases
Neurodegenerative Diseases
Amino Acid Metabolism
Biosynthesis of Other Secondary Metabolites
Carbohydrate Metabolism
Energy Metabolism
Enzyme Families
Glycan Biosynthesis and Metabolism
Lipid Metabolism
Metabolism of Cofactors and Vitamins
Metabolism of Other Amino Acids
Metabolism of Terpenoids and Polyketides
Nucleotide Metabolism
Xenobiotics Biodegradation and Metabolism
Circulatory System
Digestive System
Endocrine System
Environmental Adaptation
Excretory System
Immune System
Nervous System
Sensory System
b
Incidence Matrix
Species
KO
c
Distribution of Functional Distance
$P(d_{ij})$
d_{ij}
d
Distribution of Unweighted Gene Degree
$P(k_{\mathrm{KO}})$
k_{KO}
e
Distribution of Unweighted Species Degree
$P(k_{\mathrm{species}})$
k_{species}

FIG. 2:

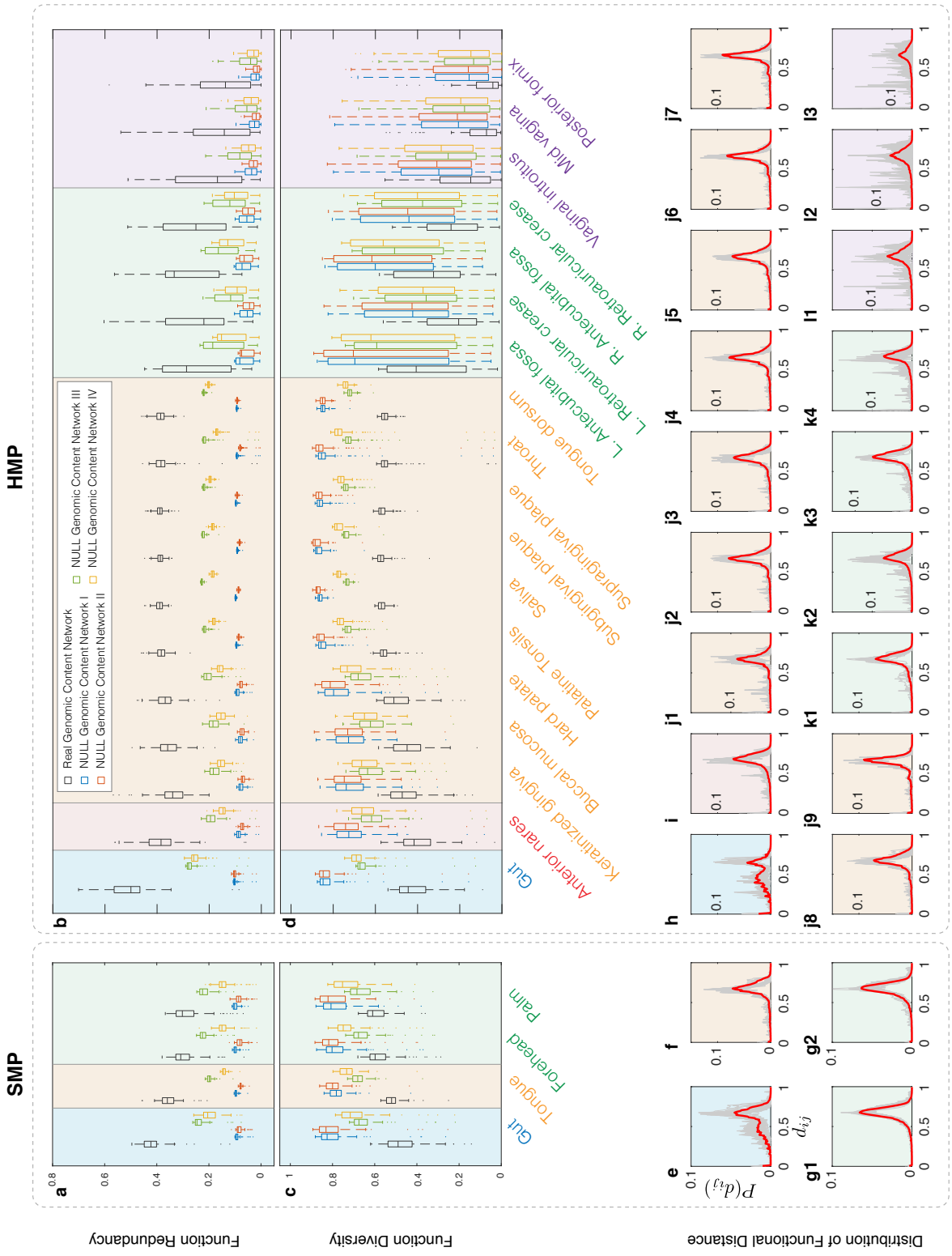
SMP
HMP
Real Genomic Content Network
NULL Genomic Content Network I
NULL Genomic Content Network II
NULL Genomic Content Network III
NULL Genomic Content Network IV
Gut
Tongue
Forehead
Palm
Gut
Anterior nares
Keratinized gingiva
Buccal mucosa
Hard palate
Palatine Tonsils
Saliva
Subgingival plaque
Supragingival plaque
Throat
Tongue dorsum
L. Anteocubital fossa
L. Retroauricular crease
R. Anteocubital fossa
R. Retroauricular crease
Vaginal introitus
Mid vagina
Posterior fornix
Function Redundancy
Function Diversity
Distribution of Functional Distance
$P(d_{ij})$
d_{ij}

FIG. 3:

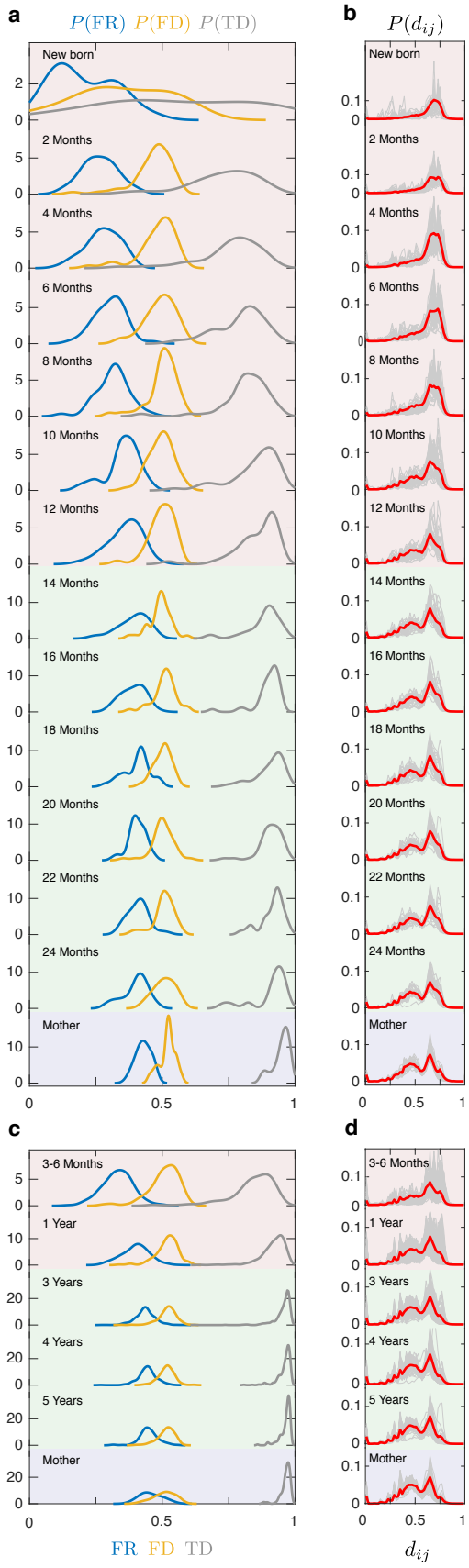
a
P(FR) P(FD) P(TD)
b
P(d_ij)
New born
2 Months
4 Months
6 Months
8 Months
10 Months
12 Months
14 Months
16 Months
18 Months
20 Months
22 Months
24 Months
Mother
c
d
3-6 Months
1 Year
3 Years
4 Years
5 Years
Mother
FR FD TD
d_ij

FIG. 4:

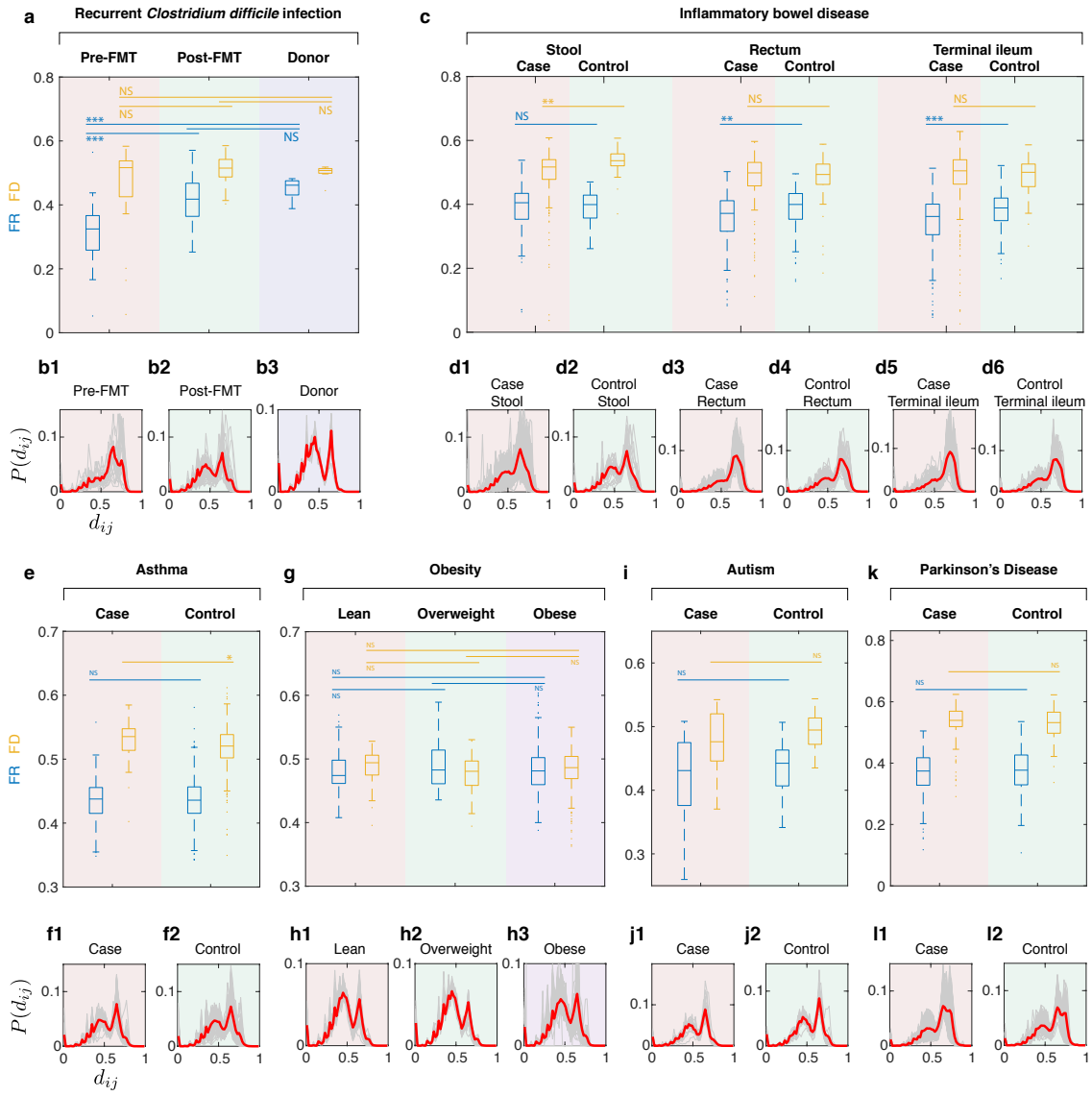
a
Recurrent Clostridium difficile infection
Pre-FMT
Post-FMT
Donor
FR FD
c
Inflammatory bowel disease
Stool
Rectum
Terminal ileum
Case
Control
b1
b2
b3
d1
d2
d3
d4
d5
d6
Case Stool
Control Stool
Case Rectum
Control Rectum
Case Terminal ileum
Control Terminal ileum
$P(d_{ij})$
d_{ij}
e
Asthma
g
Obesity
Lean
Overweight
Obese
i
Autism
k
Parkinson's Disease
f1
f2
h1
h2
h3
j1
j2
l1
l2

FIG. 5:

Deciphering Functional Redundancy in the Human Microbiome: *Supplementary Information*

Liang Tian, Ang-Kun Wu, Jonathan Friedman,
Matthew K. Waldor, Scott T. Weiss, Yang-Yu Liu*

August 14, 2017

Contents

1 Supplementary Figures **2**

2 Calculation of functional profiles **26**

3 Functional distances among different taxa **26**

4 Taxonomic and functional diversity measures based on Hill numbers **27**
4.1 Taxonomic diversity 27
4.2 Functional diversity 28
4.3 Functional redundancy 30

5 Constructing the GCN **32**

6 Randomizing the GCN **33**

7 Nestedness of the GCN **34**

8 Human microbiome datasets **35**

*Corresponding author: yyl@channing.harvard.edu

1 Supplementary Figures

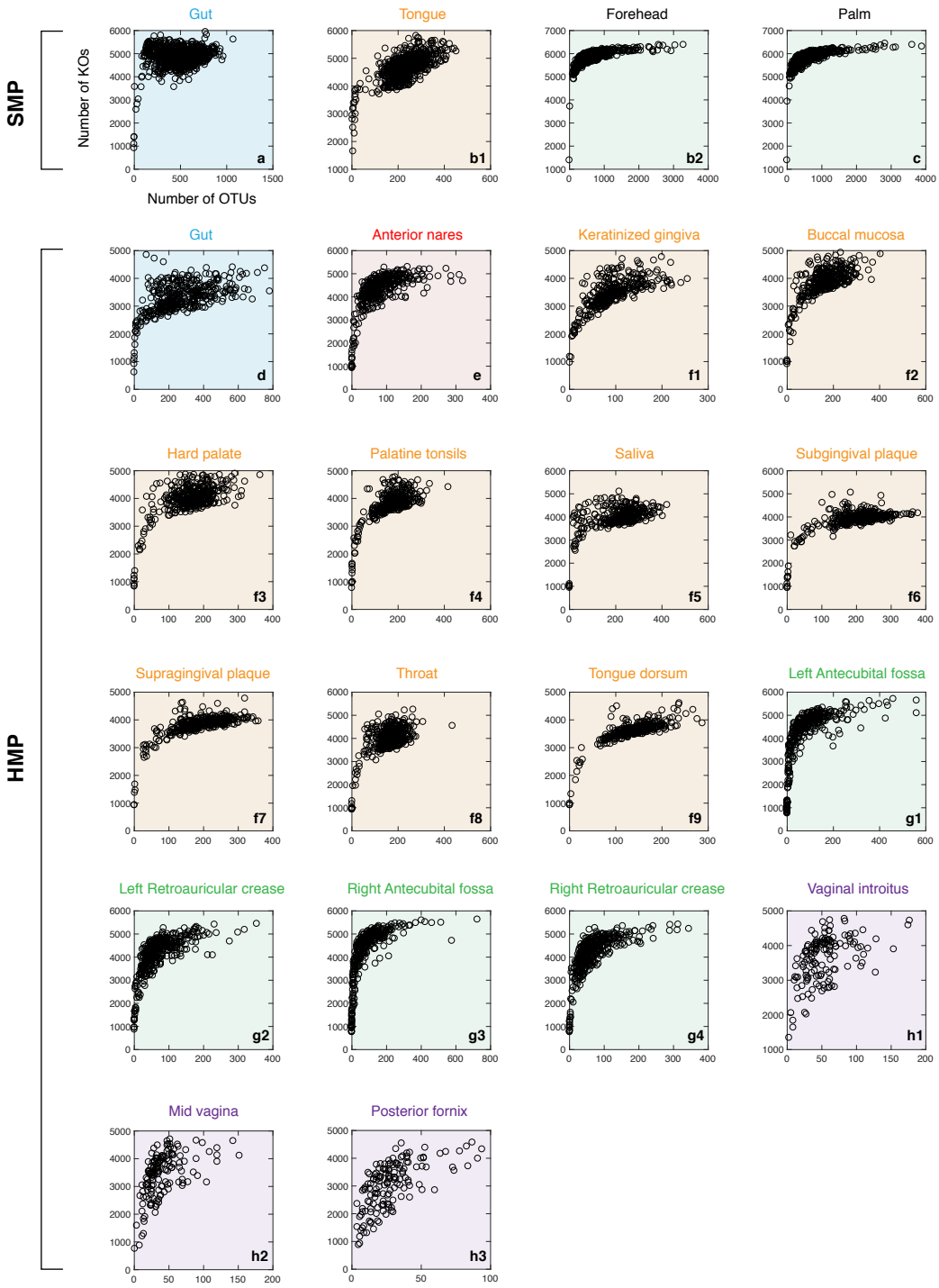

Figure S1

Figure S1: **Human microbiome displays high functional redundancy.** Functional richness (the number of KOs) vs. taxon richness (the number of OTUs) for microbiome samples from the Student Microbiome Project (SMP) [1] and the Human Microbiome Project (HMP) [2, 3]. For all the body sites, the function richness gradually saturates as taxon richness increases, implying strong functional redundancy.

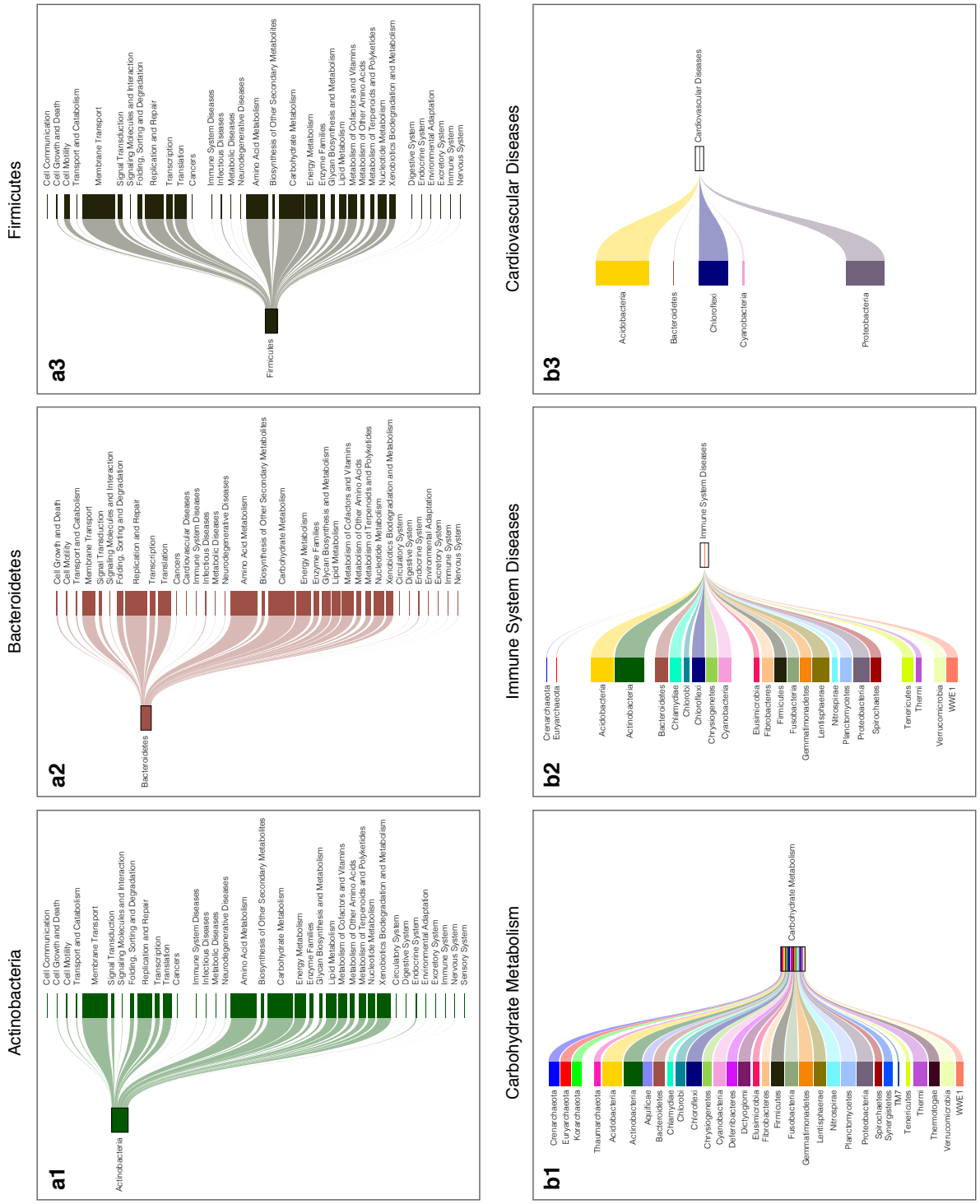
Actinobacteria
Bacteroidetes
Firmicutes
Carbohydrate Metabolism
Immune System Diseases
Cardiovascular Diseases
a1
a2
a3
b1
b2
b3
Actinobacteria
Bacteroidetes
Firmicutes
Carbohydrate Metabolism
Immune System Diseases
Cardiovascular Diseases
Cell Communication
Cell Growth and Death
Cell Motility
Transport and Catabolism
Membrane Transport
Signal Transduction
Signaling Molecules and Interaction
Folding, Sorting and Degradation
Replication and Repair
Transcription
Translation
Cancers
Immune System Diseases
Infectious Diseases
Metabolic Diseases
Neurodegenerative Diseases
Amino Acid Metabolism
Biosynthesis of Other Secondary Metabolites
Carbohydrate Metabolism
Energy Metabolism
Enzyme Families
Glycan Biosynthesis and Metabolism
Lipid Metabolism
Metabolism of Cofactors and Vitamins
Metabolism of Other Amino Acids
Metabolism of Terpenoids and Polyketides
Nucleotide Metabolism
Xenobiotics Biodegradation and Metabolism
Circulatory System
Digestive System
Endocrine System
Environmental Adaptation
Excretory System
Immune System
Nervous System
Sensory System
Crenarchaeota
Euryarchaeota
Korarchaeota
Thaumarchaeota
Acidobacteria
Actinobacteria
Aquificae
Bacteroidetes
Chlamydiae
Chlorobi
Chloroflexi
Chrysiogenetes
Cyanobacteria
Deferribacteres
Dictyoglomi
Elusimicrobia
Fibrobacteres
Firmicutes
Fusobacteria
Gemmatimonadetes
Lentisphaerae
Nitrospirae
Planctomycetes
Proteobacteria
Spirochaetes
Synergistetes
TM7
Tenericutes
Thermi
Thermotogae
Verrucomicrobia
WWE1

Figure S2

Figure S2: **Subgraphs of the reference genomic content network (GCN) constructed from the Integrated Microbial Genomes (IMG) database** [4]**.** **a**, Subgraphs of the reference GCN for three phyla: Actinobacteria, Bacteroidetes, and Firmicutes. The height of the bar on the left-hand side is proportional to the average genome size of the representative species in that phylum. The bar height on the right-hand side is proportional to the copy number of KOs in the genome of this phylum that belong to that KEGG super-pathway. **b**, Subgraphs of the reference GCN for three KEGG super-pathways: Carbohydrate Metabolism, Immune System Diseases, and Cardiovascular Diseases. The height of the bar on the left-hand side is proportional to the copy number of KOs in each phylum belonging to that pathway, and the colors of bars in the black square on the right-hand side represent the phyla whose genomes include KOs belonging to the pathway.

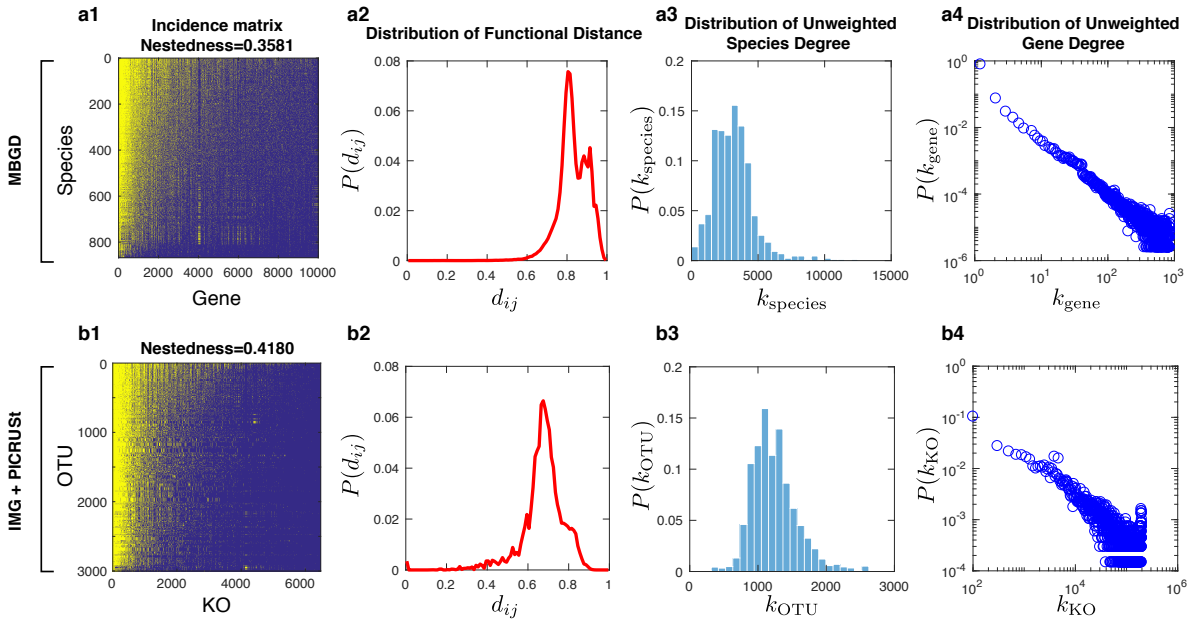
a1
Incidence matrix
Nestedness=0.3581
MBGD
Species
Gene
a2
Distribution of Functional Distance
$P(d_{ij})$
d_{ij}
a3
Distribution of Unweighted Species Degree
$P(k_{\mathrm{species}})$
k_{species}
a4
Distribution of Unweighted Gene Degree
$P(k_{\mathrm{gene}})$
k_{gene}
b1
Nestedness=0.4180
IMG + PICRUSt
OTU
KO
b2
$P(d_{ij})$
d_{ij}
b3
$P(k_{\mathrm{OTU}})$
k_{OTU}
b4
$P(k_{\mathrm{KO}})$
k_{KO}

Figure S3

Figure S3: **Structural properties of the genomic content networks (GCNs). a**, The reference GCN constructed from the MicroBial Genome Database (MBGD) [5]. **b**, The extended reference GCN constructed by using a reference gene content table derived from the Integrated Microbial Genomes (IMG) database [4], and a reference OTU phylogenetic tree derived from the Greengenes Database [6]. Note that this gene content inference is actually the first step in the workflow of PICRUSt (phylogenetic investigation of communities by reconstruction of unobserved states) [7]. (**column-1**) The incidence matrix of the GCN, where the presence (or absence) of a link is colored in yellow (or blue), respectively. We organized this matrix using the Nestedness Temperature Calculator (NTC) to emphasize its nested structure [8]. The nestedness of this network is calculated based on the classical NODF measure [9]. Note that NTC is computationally intensive. Here, for visualization purpose, in (a1) we only include the top 10,000 genes with the highest degree; and in (b1) we only include 3,000 randomly selected OTUs. (**column-2**), The distribution of functional distances (d_{ij}) between two different species or OTUs. (**column-3**), Distribution of the unweighted degrees of species or OTUs. (**column-4**), Distribution of the unweighted degrees of genes or KOs.

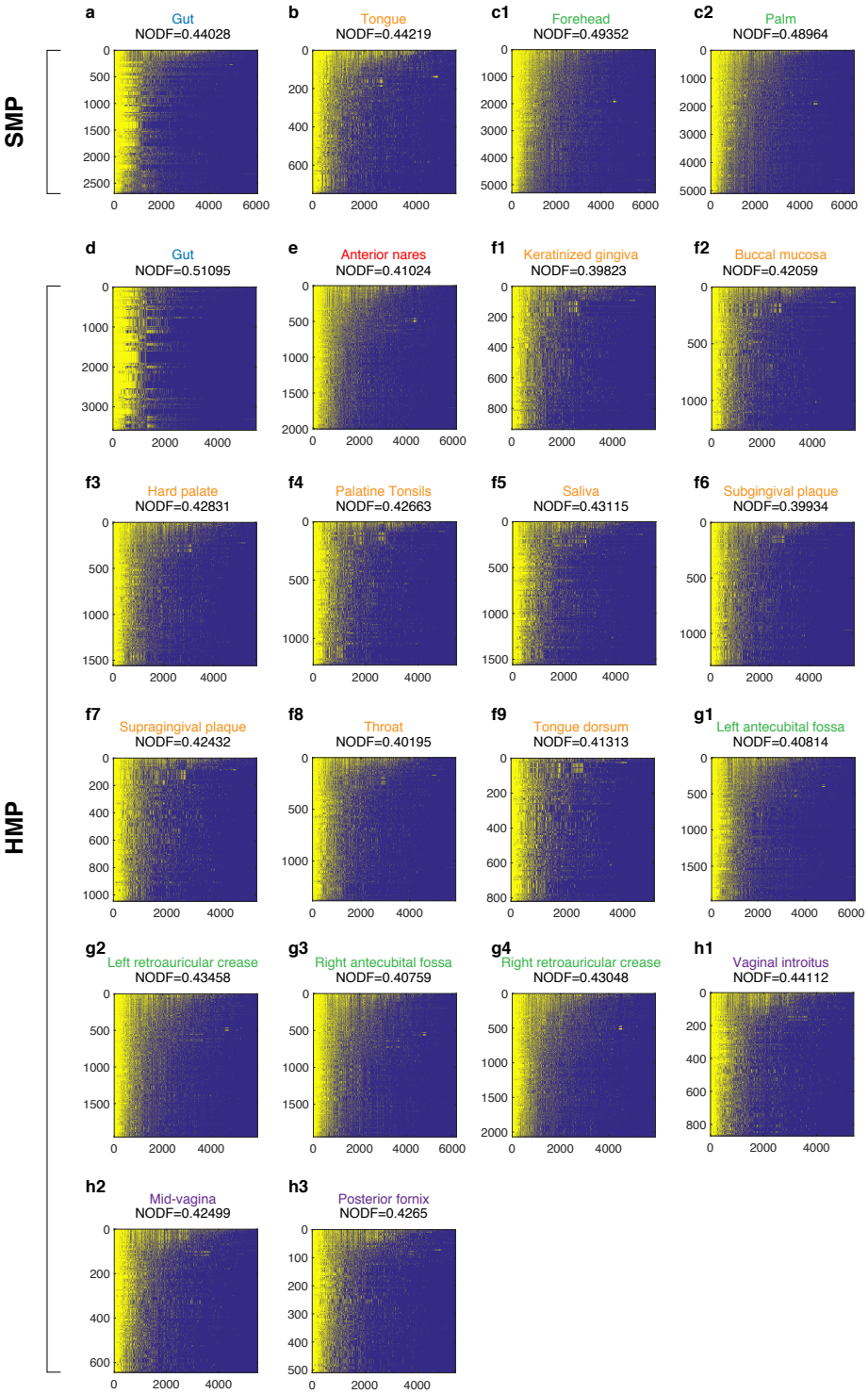
SMP
HMP
a
Gut
NODF=0.44028
b
Tongue
NODF=0.44219
c1
Forehead
NODF=0.49352
c2
Palm
NODF=0.48964
d
Gut
NODF=0.51095
e
Anterior nares
NODF=0.41024
f1
Keratinized gingiva
NODF=0.39823
f2
Buccal mucosa
NODF=0.42059
f3
Hard palate
NODF=0.42831
f4
Palatine Tonsils
NODF=0.42663
f5
Saliva
NODF=0.43115
f6
Subgingival plaque
NODF=0.39934
f7
Supragingival plaque
NODF=0.42432
f8
Throat
NODF=0.40195
f9
Tongue dorsum
NODF=0.41313
g1
Left antecubital fossa
NODF=0.40814
g2
Left retroauricular crease
NODF=0.43458
g3
Right antecubital fossa
NODF=0.40759
g4
Right retroauricular crease
NODF=0.43048
h1
Vaginal introitus
NODF=0.44112
h2
Mid-vagina
NODF=0.42499
h3
Posterior fornix
NODF=0.4265

Figure S4

Figure S4: **Incidence matrix of the bodysite-specific genomic content network (GCN) for two independent datasets: the Student Microbiome Project (SMP) [1] and the Human Microbiome Project (HMP) [2, 3].** For each body site, we consider the union of OTUs present in all the samples, from which we construct the bodysite-specific GCN based on the extended reference GCN. In each incidence matrix, the presence (or absence) of a link is colored in yellow (or blue), respectively. We organized all the matrices using the Nestedness Temperature Calculator to emphasize their nested structure [8]. The nestedness of each network is calculated based on the classical NODF measure [9].

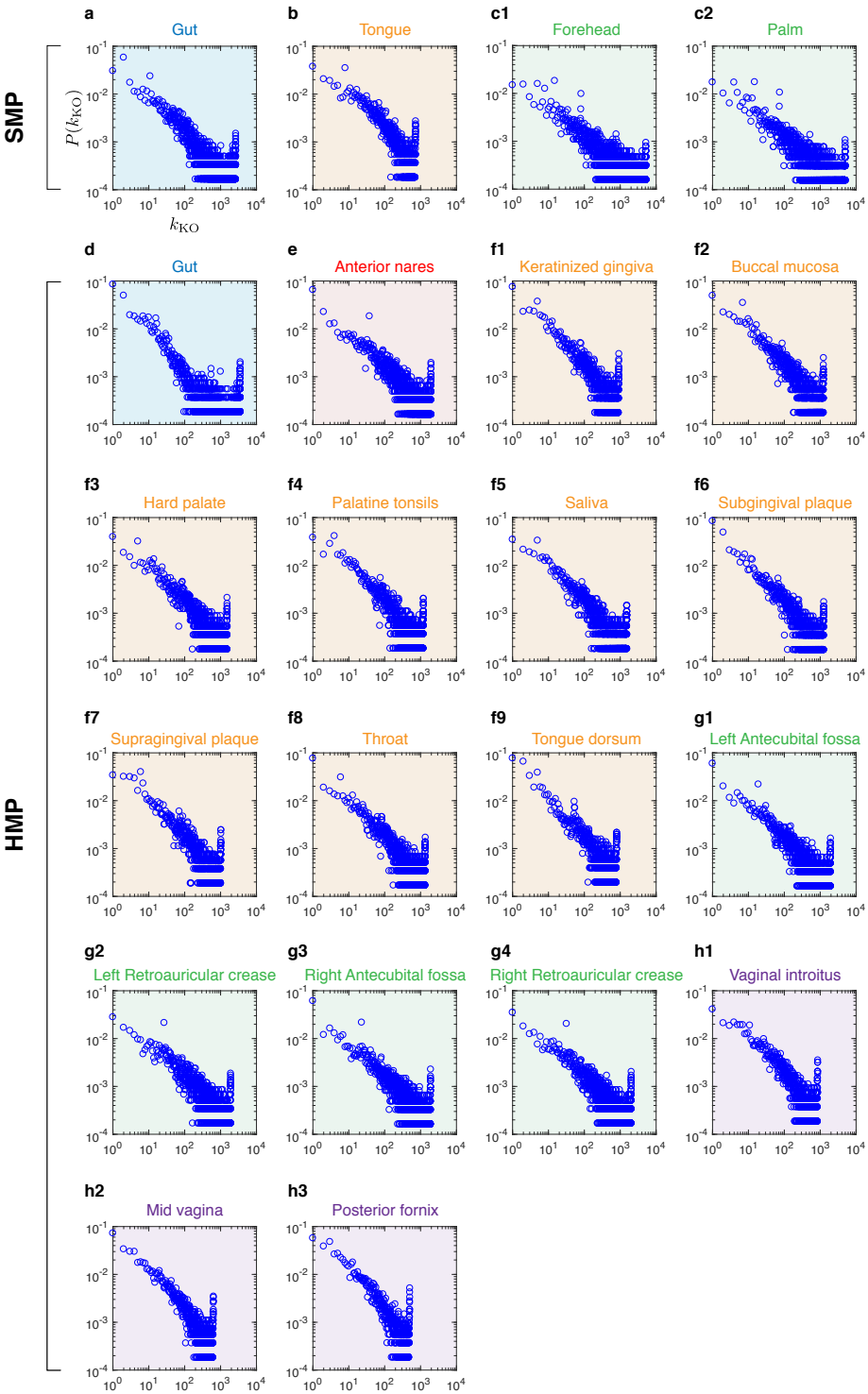
SMP
HMP
a
Gut
b
Tongue
c1
Forehead
c2
Palm
d
Gut
e
Anterior nares
f1
Keratinized gingiva
f2
Buccal mucosa
f3
Hard palate
f4
Palatine tonsils
f5
Saliva
f6
Subgingival plaque
f7
Supragingival plaque
f8
Throat
f9
Tongue dorsum
g1
Left Antecubital fossa
g2
Left Retroauricular crease
g3
Right Antecubital fossa
g4
Right Retroauricular crease
h1
Vaginal introitus
h2
Mid vagina
h3
Posterior fornix
$P(k_{\mathrm{KO}})$
k_{KO}

Figure S5

Figure S5: **The unweighted KO degree distribution of bodysite-specific GCN for two independent datasets: the Student Microbiome Project (SMP) [1] and the Human Microbiome Project (HMP) [2, 3].** For each body site, we consider the union of OTUs present in all the samples, from which we construct the bodysite-specific GCN based on the extended reference GCN. Here, the unweighted degree of a KO, denoted as k_{KO}, is just the number of OTUs whose genomes contain this KO. Note that a few housekeeping KOs exist in the genomes of almost all OTUs, resulting in a peak in the tail of the $P(k_{\mathrm{KO}})$.

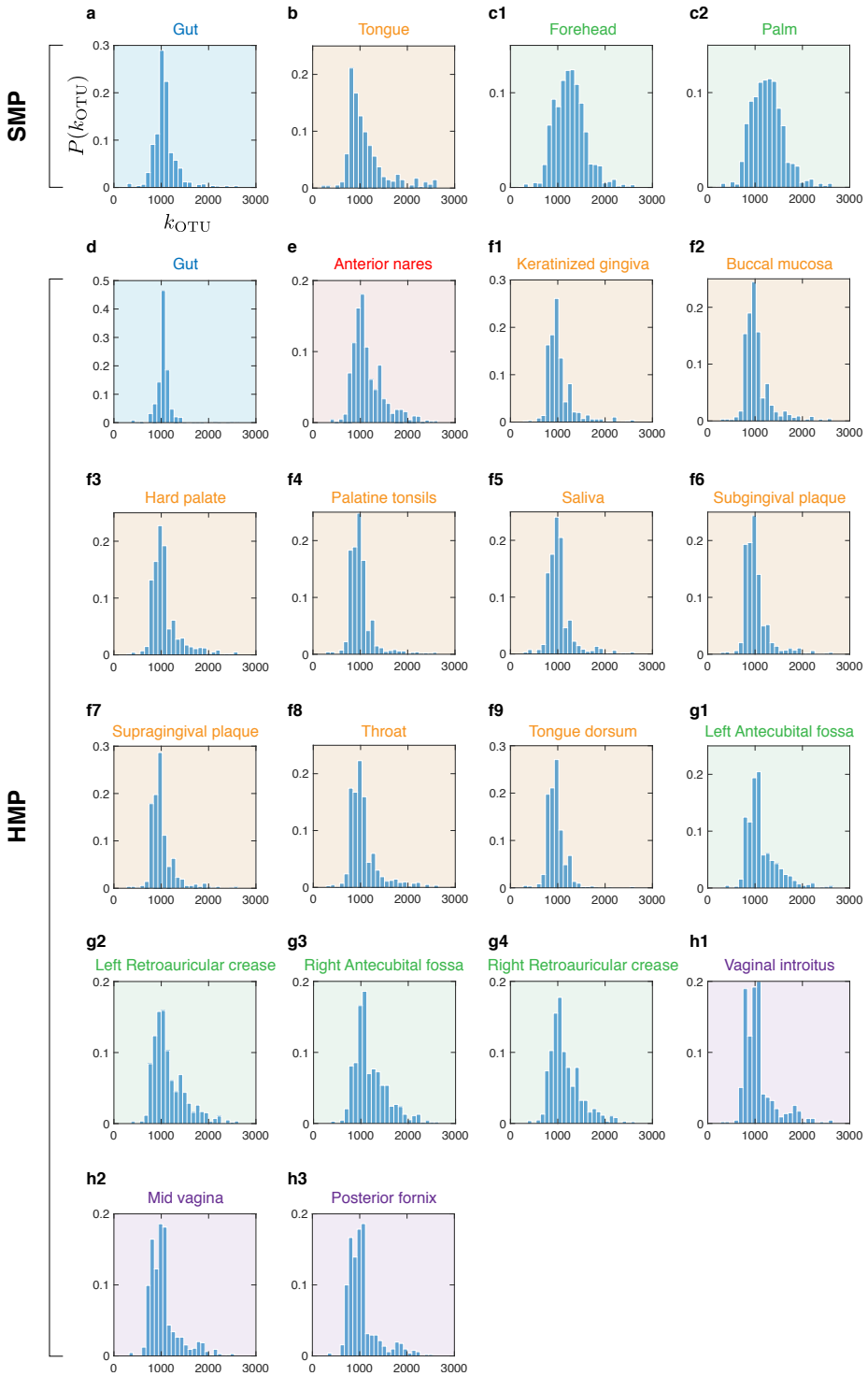
SMP
HMP
a
Gut
b
Tongue
c1
Forehead
c2
Palm
$P(k_{\mathrm{OTU}})$
k_{OTU}
d
Gut
e
Anterior nares
f1
Keratinized gingiva
f2
Buccal mucosa
f3
Hard palate
f4
Palatine tonsils
f5
Saliva
f6
Subgingival plaque
f7
Supragingival plaque
f8
Throat
f9
Tongue dorsum
g1
Left Antecubital fossa
g2
Left Retroauricular crease
g3
Right Antecubital fossa
g4
Right Retroauricular crease
h1
Vaginal introitus
h2
Mid vagina
h3
Posterior fornix

Figure S6

Figure S6: **The unweighted OTU degree distribution of bodysite-specific GCN for two independent datasets: the Student Microbiome Project (SMP) [1] and the Human Microbiome Project (HMP) [2, 3].** For each body site, we consider the union of OTUs present in all the samples, from which we construct the bodysite-specific GCN based on the extended reference GCN. The unweighted degree of an OTU is just the number of distinct KOs in its genome.

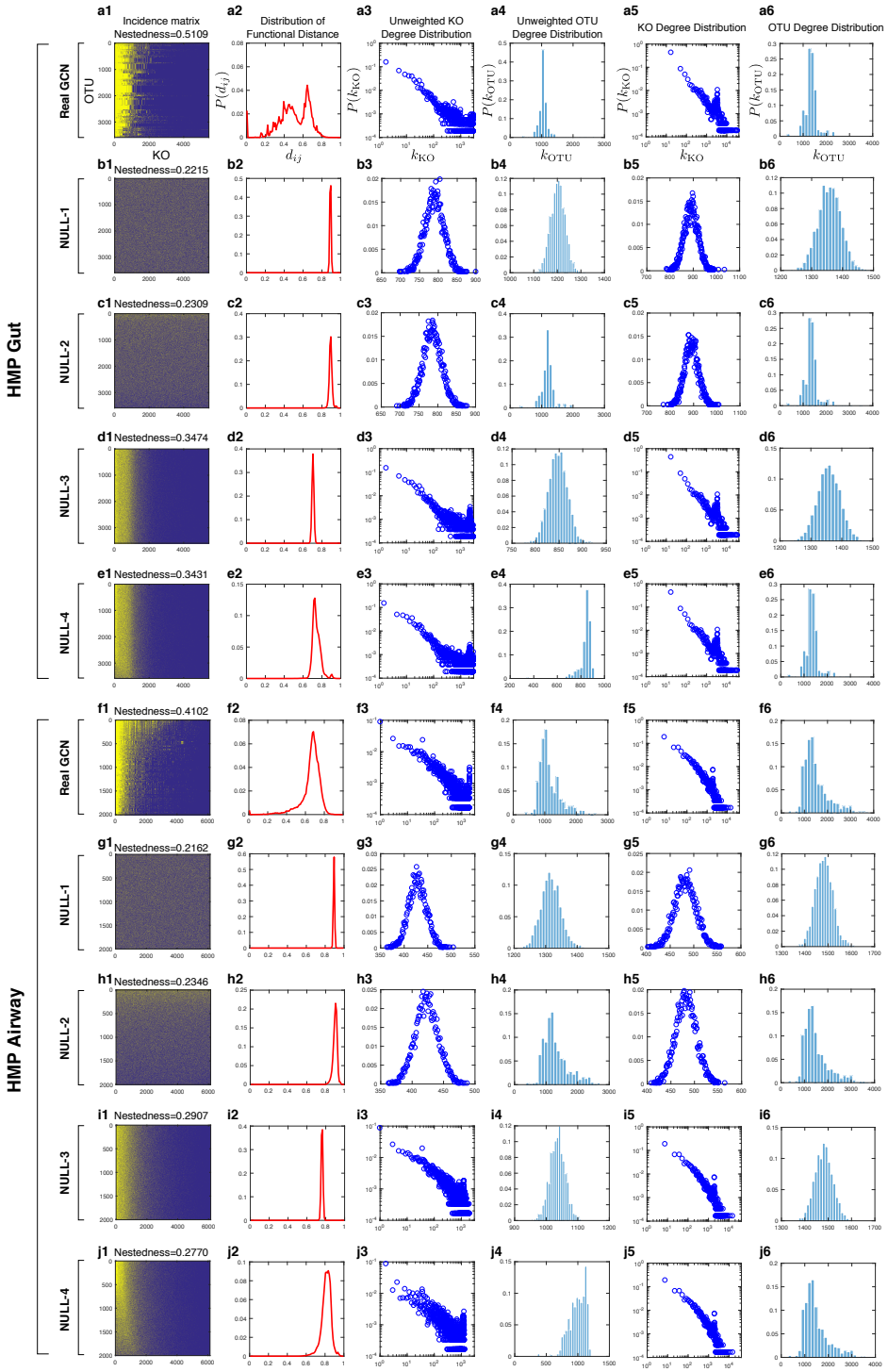
HMP Gut
HMP Airway
Real GCN
NULL-1
NULL-2
NULL-3
NULL-4
a1 Incidence matrix Nestedness=0.5109
a2 Distribution of Functional Distance
a3 Unweighted KO Degree Distribution
a4 Unweighted OTU Degree Distribution
a5 KO Degree Distribution
a6 OTU Degree Distribution
b1 Nestedness=0.2215
c1 Nestedness=0.2309
d1 Nestedness=0.3474
e1 Nestedness=0.3431
f1 Nestedness=0.4102
g1 Nestedness=0.2162
h1 Nestedness=0.2346
i1 Nestedness=0.2907
j1 Nestedness=0.2770
OTU
KO
$P(d_{ij})$
d_{ij}
$P(k_{\mathrm{KO}})$
k_{KO}
$P(k_{\mathrm{OTU}})$
k_{OTU}

Figure S7

Figure S7: **Structural properties of the gut-GCN and the airway-GCN of the Human Microbiome Project (HMP) [2, 3].** (**a-e**), Gut. (**f-j**), Airway. (**colomn-1**), The incidence matrix of the GCN shown at the OTU-KO level, where the presence (or absence) of a link between an OTU and a KO is colored in yellow (or blue), respectively. We organized this matrix using the Nestedness Temperature Calculator to emphasize its nested structure [8]. The nestedness of this network is calculated based on the NODF measure [9]. (**colomn-2**), The distribution of functional distances (d_{ij}) between different OTUs. (**column-3**), The unweighted KO degree distribution. Here, the unweighted degree of a KO is just the number of OTUs whose genomes contain this KO. (**column-4**), The unweighted OTU degree distribution. Here, the unweighted degree of an OTU is just the number of distinct KOs in its genome. (**column-5**), The KO degree distribution, where the degree of a KO is the sum of copy numbers of this KO in those OTUs whose genomes contain this KO. (**column-6**), The OTU degree distribution, where the degree of an OTU is the sum of copy numbers of those KOs in this OTU's genome. (**a,f**), The structural properties of the real GCN. (**b,g**), Null-1: Completely randomized GCN. (**c,h**) Null-2: Randomized GCN with OTU-degree preserved. (**d,i**) Null-3: Randomized GCN with KO-degree preserved. (**e,j**) Null-4: Randomized GCN with both OTU-degree and KO-degree preserved. Note that a few housekeeping KOs exist in the genome of almost every OTU, resulting in a peak in the tail of $P(k_{\mathrm{KO}})$. See SI Sec. 6 for the details of the four different null models.

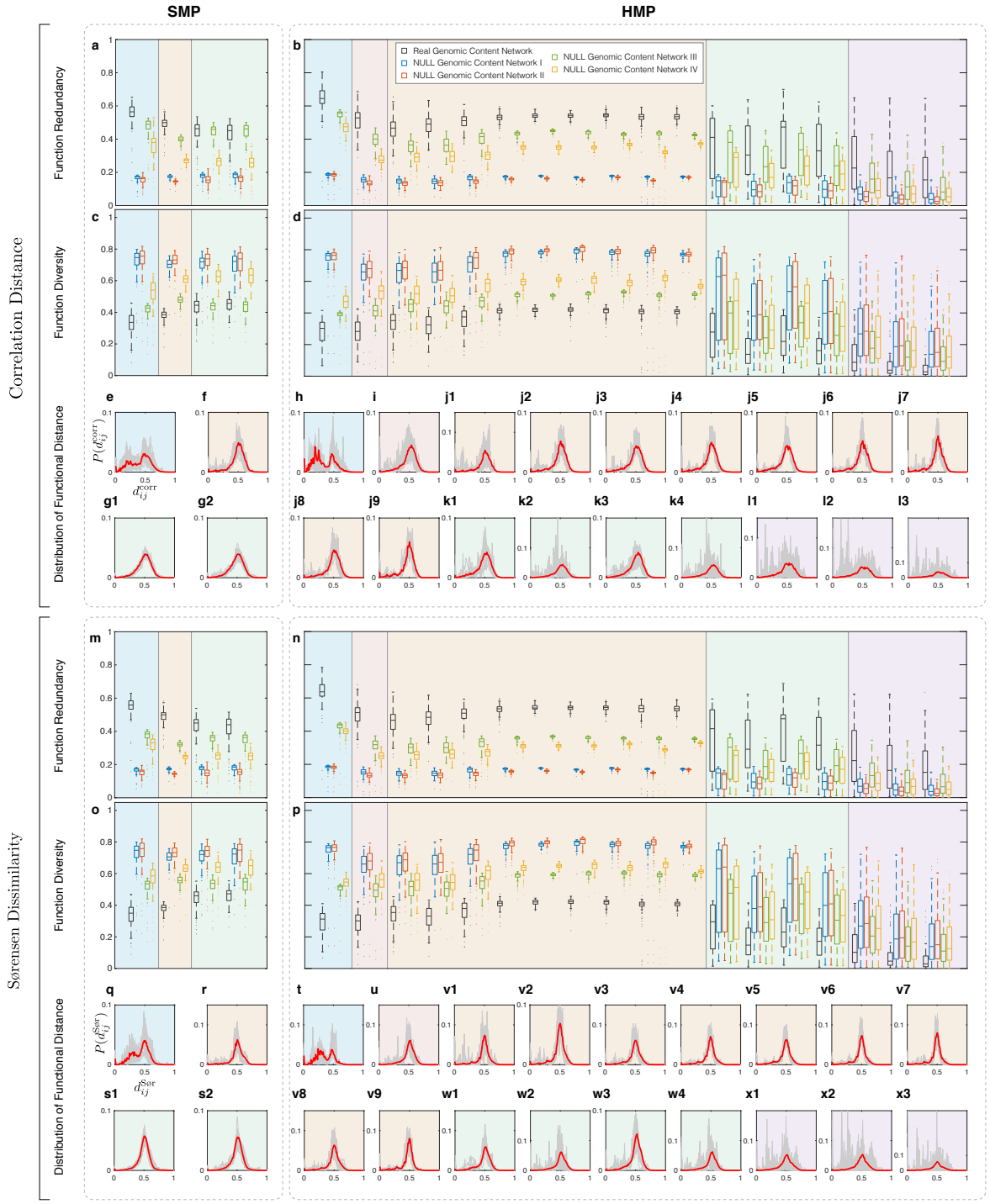
SMP
HMP
Correlation Distance
Sørensen Dissimilarity
Function Redundancy
Function Diversity
Distribution of Functional Distance
Distribution of Functional Distance
Real Genomic Content Network
NULL Genomic Content Network I
NULL Genomic Content Network II
NULL Genomic Content Network III
NULL Genomic Content Network IV
a
b
c
d
e
f
g1
g2
h
i
j1
j2
j3
j4
j5
j6
j7
j8
j9
k1
k2
k3
k4
l1
l2
l3
m
n
o
p
q
r
s1
s2
t
u
v1
v2
v3
v4
v5
v6
v7
v8
v9
w1
w2
w3
w4
x1
x2
x3

Figure S8

Figure S8: **Functional redundancy, functional diversity, and the distribution of OTU functional distances calculated using two different distance measures.** (**a-l**), Correlation distance [10]; (**m-x**), Sørensen dissimilarity [11]. Microbiome samples from the Student Microbiome Project (SMP) [1] with 4 different body sites, as well as the Human Microbiome Project (HMP) [2, 3] with 18 different body sites, were analyzed. (**a-b, m-n**), Functional redundancy. (**c-d, o-p**), Functional diversity. The box plots of function redundancy and function diversity of those samples were calculated from the real GCN (black box), as well as the randomized GCNs (colored boxes) using four different randomization schemes: complete randomization (blue); OTU-degree preserving randomization (red); KO-degree preserving randomization (green); OTU- and KO-degree preserving randomization (yellow). (**e-l**), The distribution of functional distances $P(d_{ij}^{\mathrm{corr}})$ calculated with correlation distance [10]. (**q-x**), The distribution of functional distances $P(d_{ij}^{\mathrm{Sør}})$ calculated with Sørensen dissimilarity [11]. In (**e-l, q-x**), thin gray lines: functional distance distribution for different samples for a given body site; thick red lines: functional distance distribution calculated from the union of OTUs present in all the samples for each body site.

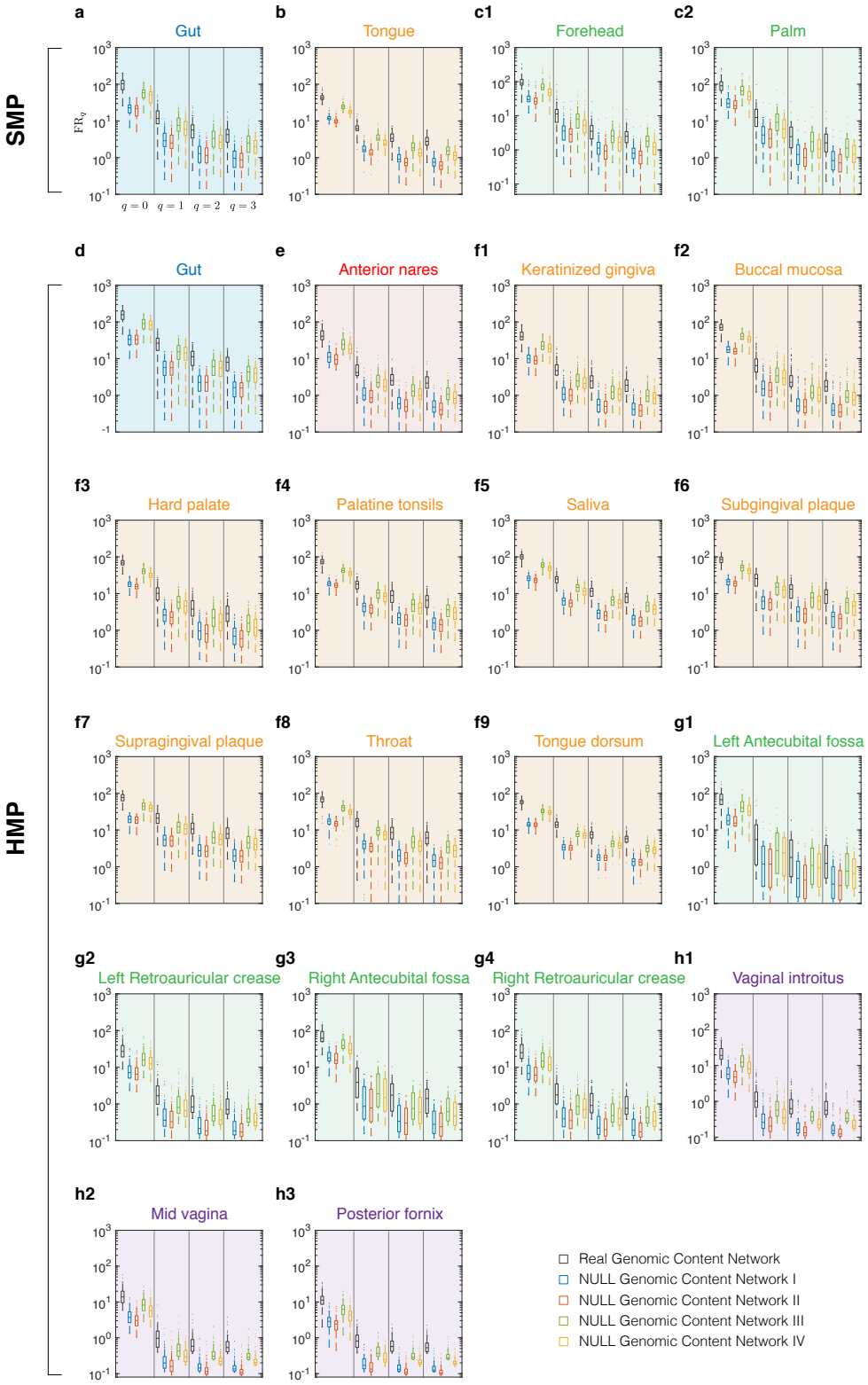
SMP
HMP
a
Gut
b
Tongue
c1
Forehead
c2
Palm
d
Gut
e
Anterior nares
f1
Keratinized gingiva
f2
Buccal mucosa
f3
Hard palate
f4
Palatine tonsils
f5
Saliva
f6
Subgingival plaque
f7
Supragingival plaque
f8
Throat
f9
Tongue dorsum
g1
Left Antecubital fossa
g2
Left Retroauricular crease
g3
Right Antecubital fossa
g4
Right Retroauricular crease
h1
Vaginal introitus
h2
Mid vagina
h3
Posterior fornix
FR_q
q = 0
q = 1
q = 2
q = 3
Real Genomic Content Network
NULL Genomic Content Network I
NULL Genomic Content Network II
NULL Genomic Content Network III
NULL Genomic Content Network IV

Figure S9

Figure S9: **Hill number based functional redundancy of microbiome samples in different body sites for two independent datasets: the Student Microbiome Project (SMP) [1] and the Human Microbiome Project (HMP) [2, 3].** The box plots of FR_q values of those samples were calculated from the real GCN (black box), as well as the randomized GCNs (colored boxes) using four different randomization schemes: complete randomization (blue); OTU-degree preserving randomization (red); KO-degree preserving randomization (green); OTU- and KO-degree preserving randomization (yellow). See Sec. 4 for the definition of Hill number based functional redundancy.

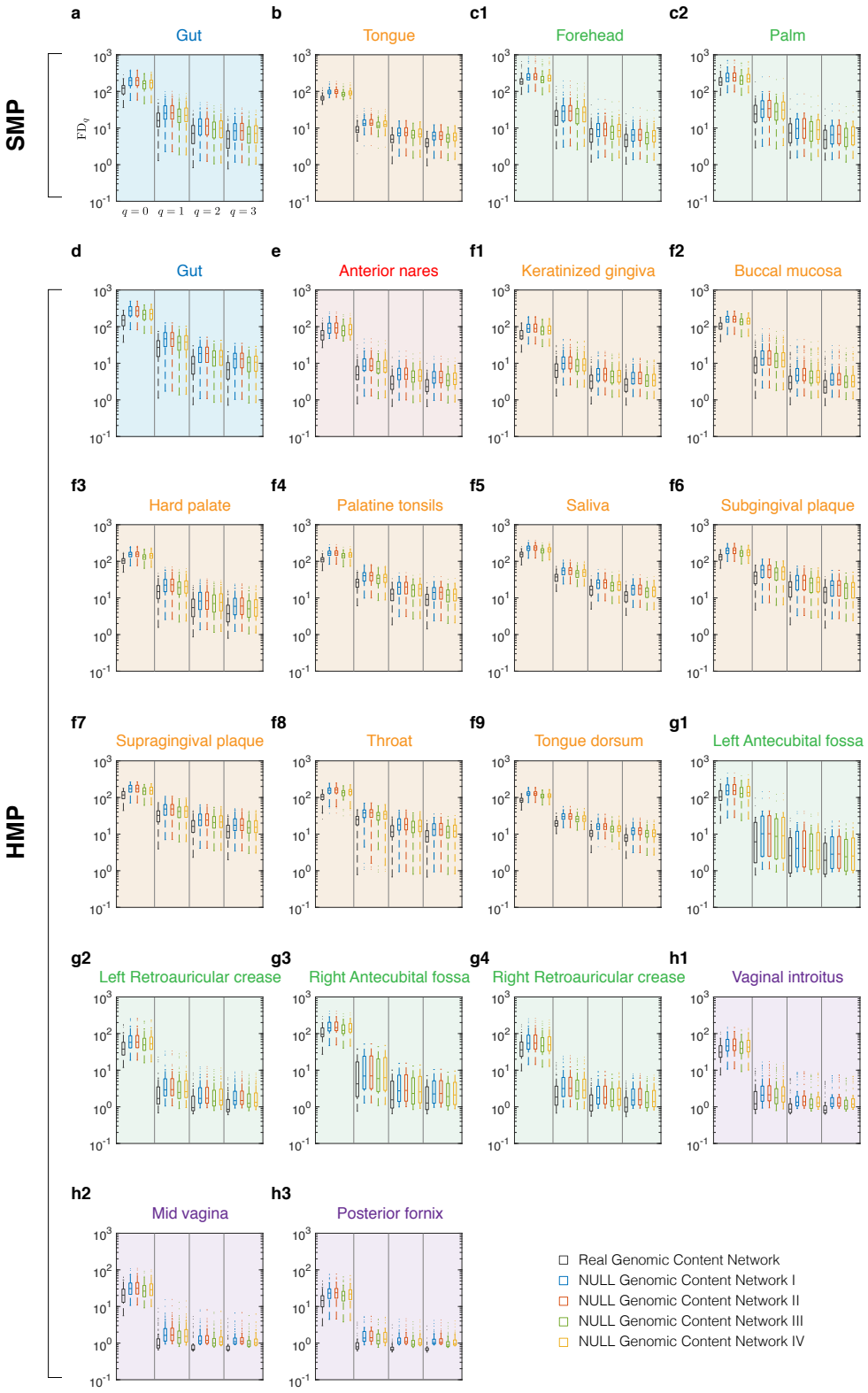
SMP
HMP
a
Gut
b
Tongue
c1
Forehead
c2
Palm
d
Gut
e
Anterior nares
f1
Keratinized gingiva
f2
Buccal mucosa
f3
Hard palate
f4
Palatine tonsils
f5
Saliva
f6
Subgingival plaque
f7
Supragingival plaque
f8
Throat
f9
Tongue dorsum
g1
Left Antecubital fossa
g2
Left Retroauricular crease
g3
Right Antecubital fossa
g4
Right Retroauricular crease
h1
Vaginal introitus
h2
Mid vagina
h3
Posterior fornix
FD_q
$q = 0$ $q = 1$ $q = 2$ $q = 3$
Real Genomic Content Network
NULL Genomic Content Network I
NULL Genomic Content Network II
NULL Genomic Content Network III
NULL Genomic Content Network IV

Figure S10

Figure S10: **Hill number based functional diversity for microbiome samples in different body sites for two independent datasets: the Student Microbiome Project (SMP) [1] and the Human Microbiome Project (HMP) [2, 3].** The box plots of FD_q values of those samples were calculated from the real GCN (black box), as well as the randomized GCNs (colored boxes) using four different randomization schemes: complete randomization (blue); OTU-degree preserving randomization (red); KO-degree preserving randomization (green); OTU- and KO-degree preserving randomization (yellow). See Sec. 4 for the definition of Hill number based functional diversity.

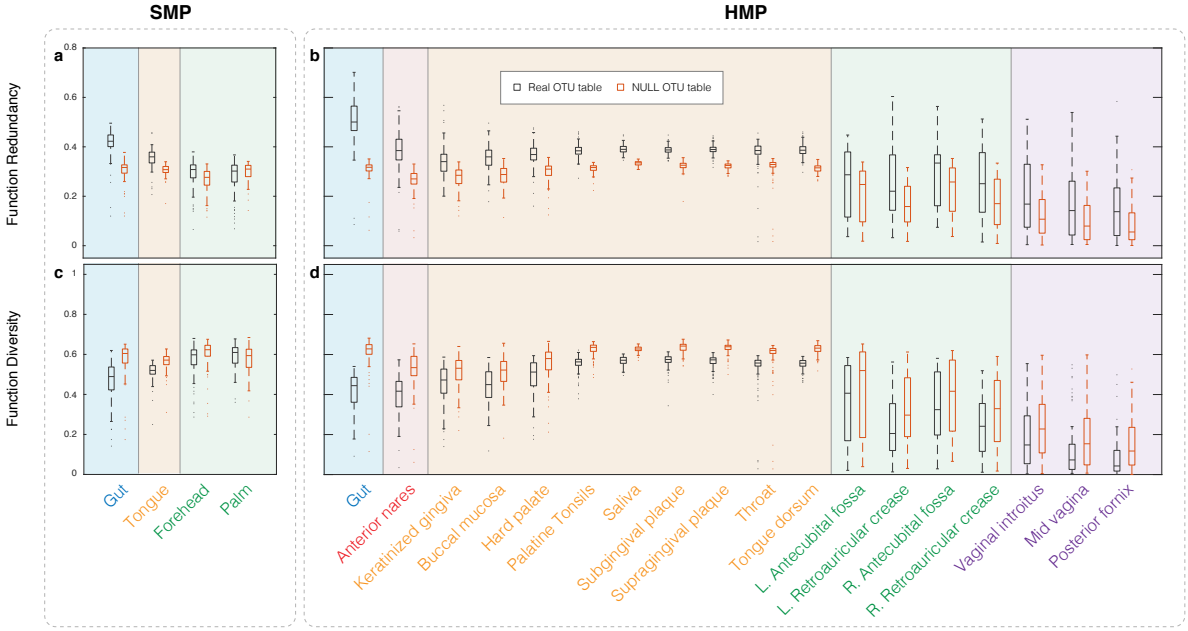
SMP
HMP
a
b
c
d
Function Redundancy
Function Diversity
Real OTU table
NULL OTU table
Gut
Tongue
Forehead
Palm
Gut
Anterior nares
Keratinized gingiva
Buccal mucosa
Hard palate
Palatine Tonsils
Saliva
Subgingival plaque
Supragingival plaque
Throat
Tongue dorsum
L. Antecubital fossa
L. Retroauricular crease
R. Antecubital fossa
R. Retroauricular crease
Vaginal introitus
Mid vagina
Posterior fornix

Figure S11

Figure S11: **Assemblages of OTUs in human microbiome samples lead to higher functional redundancy than randomly chosen OTUs.** Microbiome samples from the Student Microbiome Project (SMP) [1] with 4 different body sites, as well as the the Human Microbiome Project (HMP) [2, 3] with 18 different body sites, were analyzed. **a-b**, Functional redundancy. **c-d**, Functional diversity. The function redundancy and function diversity of each sample are calculated for the real OTU assemblage (black box), as well as random OTU assemblage (red box). For each sample, the random OTU assemblage is generated by randomly choosing the same number of OTUs from the extended reference GCN and keeping the OTU abundance profile unchanged.

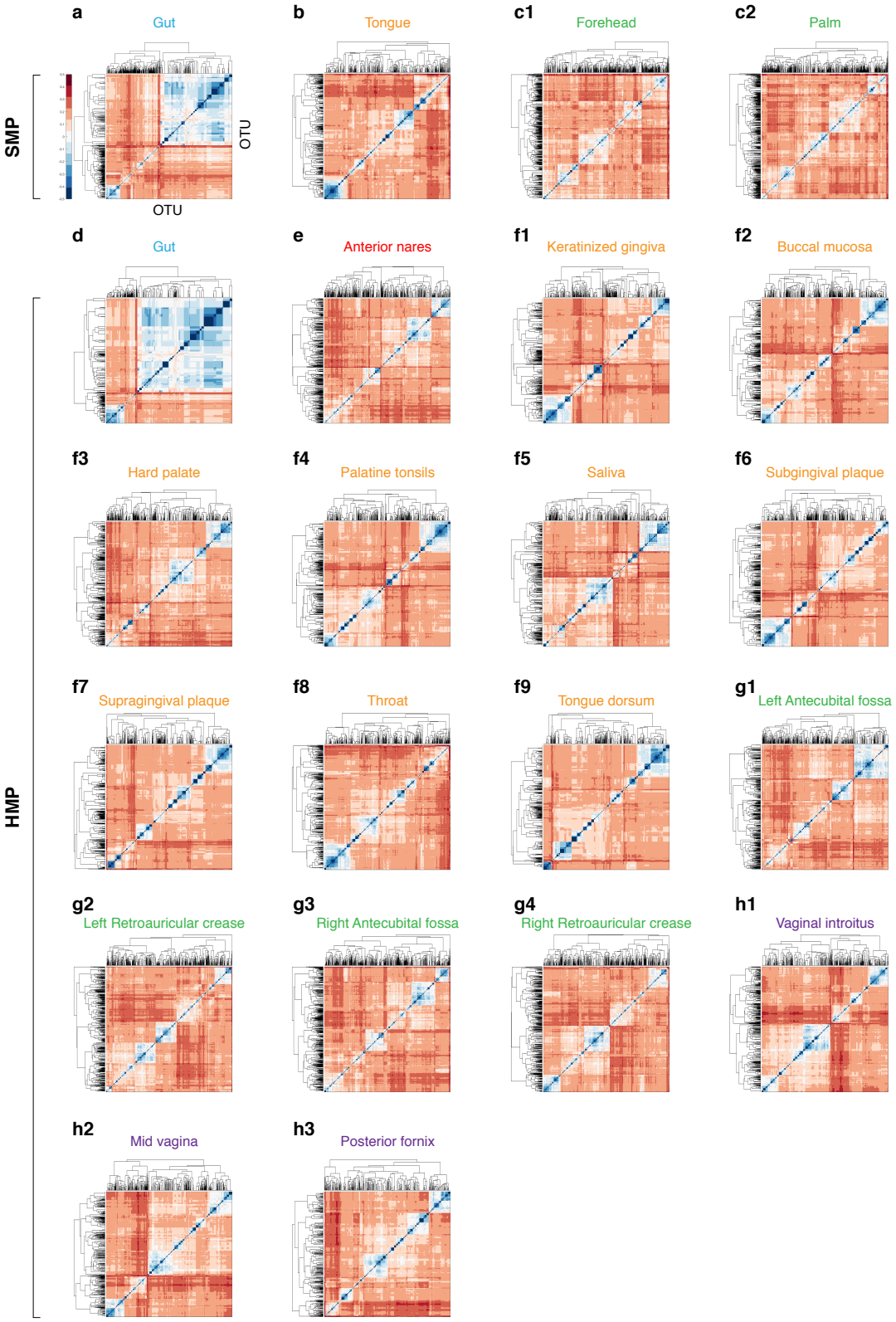
a
Gut
b
Tongue
c1
Forehead
c2
Palm
SMP
OTU
OTU
d
Gut
e
Anterior nares
f1
Keratinized gingiva
f2
Buccal mucosa
f3
Hard palate
f4
Palatine tonsils
f5
Saliva
f6
Subgingival plaque
f7
Supragingival plaque
f8
Throat
f9
Tongue dorsum
g1
Left Antecubital fossa
HMP
g2
Left Retroauricular crease
g3
Right Antecubital fossa
g4
Right Retroauricular crease
h1
Vaginal introitus
h2
Mid vagina
h3
Posterior fornix

Figure S12

Figure S12: **Functional distance matrix of OTUs in different body sites for two independent datasets: the Student Microbiome Project (SMP) [1] and the Human Microbiome Project (HMP) [2, 3].** The clustering is performed by the average-linkage hierarchical clustering [12].

2 Calculation of functional profiles

Consider a metacommunity of N taxa and M genes in total. Denote the taxonomic profile of a local community (e.g., the microbiome sample from a particular body site of subject ν) as $\mathbf{p}^{(\nu)} = \left[p_1^{(\nu)}, \cdots, p_N^{(\nu)}\right]$, where $p_i^{(\nu)}$ is the relative abundance of the i-th taxon and $\sum_{i=1}^{N} p_i^{(\nu)} = 1$. Denote the gene composition (or functional profile) of this local community as $\mathbf{f}^{(\nu)} = \left[f_1^{(\nu)}, \cdots, p_M^{(\nu)}\right]$, where $f_a^{(\nu)}$ is the relative abundance of the a-th gene and $\sum_{a=1}^{M} f_a^{(\nu)} = 1$. The genomic content network (GCN) can be represented by an $N \times M$ incidence matrix $\mathbf{G} = [G_{ia}]$, where $G_{ia} \geqslant 0$ is the copy number of gene-a in the genome of taxon-i. The GCN naturally connects the taxonomic profile and the functional profile as follows:

$$\mathbf{f}^{(\nu)} = c\mathbf{p}^{(\nu)} \cdot \mathbf{G} \tag{S1}$$

or equivalently

$$f_a^{(\nu)} = c \sum_{i=1}^{N} p_i^{(\nu)} G_{ia} \tag{S2}$$

for $a = 1, \cdots, M$. Here $c = \left[\sum_{a=1}^{M} \sum_{i=1}^{N} p_i^{(\nu)} G_{ia}\right]^{-1}$ is the normalization constant.

3 Functional distances among different taxa

In the main text the functional distance d_{ij} between taxon-i and j is calculated as the weighted Jaccard distance between the genomes of the two taxa:

$$d_{ij} = 1 - \frac{\sum_a \min(G_{ia}, G_{ja})}{\sum_a \max(G_{ia}, G_{ja})}. \tag{S3}$$

The reasons why we used the weighted Jaccard distance are twofold: (1) The genome of each taxon is represented by a vector of gene copy numbers, which are non-negative integers. Weighted Jaccard distance can naturally measure the distance between two vectors of non-negative integers. (2) The weighted Jaccard distance is normalized: $d_{ij} = 0$ indicates that taxon-i and taxon-j share exactly the same genome; $d_{ij} = 1$ means that they have totally different genomes.

Other distance or dissimilarity measures that satisfy the above conditions can also be used, such as the correlation distance [10],

$$d_{ij}^{\text{corr}} = 1 - \frac{\sum_a G_{ia} G_{ja}}{\sqrt{\left(\sum_a G_{ia}^2\right)\left(\sum_a G_{ja}^2\right)}} \tag{S4}$$

or the Sørensen dissimilarity [11],

$$d_{ij}^{\text{Sørensen}} = 1 - \frac{2\sum_a \min(G_{ia}, G_{ja})}{\sum_a G_{ia} + \sum_a G_{ja}}. \tag{S5}$$

We have checked that our results do not change quantitatively by using different distance (or dissimilarity) measures (see Fig.S8).

4 Taxonomic and functional diversity measures based on Hill numbers

4.1 Taxonomic diversity

Consider a sample of N species with the relative abundance profile given by a vector $\mathbf{p} = [p_1, \cdots, p_N]$. Here the term "species" is used in the general context of ecology, which doesn't necessarily represent the lowest major taxonomic rank. Hill introduced the concept of *species equivalents* or the *effective numbers of species* [13]. The basic idea is that the taxonomic diversity (of order q) of a given sample with relative abundance profile $\mathbf{p} = [p_1, \cdots, p_N]$ is the same as that of an idealized sample with D equally abundant species, i.e., $\tilde{\mathbf{p}} = [1/D, \cdots, 1/D]$. In other words,

$$\sum_{i=1}^{N} p_i^q = \sum_{i=1}^{D} \left(\frac{1}{D}\right)^q = D^{1-q}. \tag{S6}$$

This offers a parametric class of taxonomic diversity measures defined as follows [13]:

$$\text{TD}_q := \left(\sum_{i=1}^{N} p_i^q\right)^{\frac{1}{(1-q)}} \quad \text{for} \quad q \neq 1. \tag{S7}$$

and

$$\text{TD}_1 := \lim_{q\to 1} \text{TD}_q = \exp\left(-\sum_{i=1}^{N} p_i \log p_i\right). \tag{S8}$$

A few special cases:

- $q = 0$, $\text{TD}_0 = N$ is nothing but the *species richness*;
- $q = 1$, TD_1 is the exponential of the *Shannon entropy* (a.k.a. *Shannon index*);
- $q = 2$, $\text{TD}_2 = 1/\sum_{i=1}^{N} p_i^2$ is the *inverse Simpson index* (a.k.a. *Simpson diversity*).

Note that the *Gini-Simpson index* (GSI) used in the main text is related to TD_2 as follows:

$$\mathrm{GSI} := 1 - \sum_{i=1}^{N} p_i^2 = 1 - \frac{1}{\mathrm{TD}_2}. \tag{S9}$$

4.2 Functional diversity

Consider a sample of N species with the relative abundance profile $\mathbf{p} = [p_1, \cdots, p_N]$, and pair-wise functional distance matrix $\Delta = (d_{ij}) \in \mathbb{R}^{N\times N}$ with $d_{ii} = 0$ for all $i = 1, \cdots, N$, and $d_{ij} = d_{ji} \geq 0$ for all $i \neq j$. Recently, Chiu *et al.* extended the notion of Hill numbers to incorporate species pairwise functional distances and introduced the so-called functional Hill number [14]. In this framework, we can derive the functional Hill number as follows:

$$\sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} d_{ij}(p_i p_j)^q = \sum_{i=1}^{D} \sum_{j=1, j\neq i}^{D} Q^*(\frac{1}{D}\frac{1}{D})^q \tag{S10}$$

where

$$Q^* \equiv \frac{\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} d_{ij} p_i p_j}{\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} p_i p_j} \tag{S11}$$

denotes the average pair-wise distance weighted by their abundances. Note that when different species are equally distinct with a constant pairwise distance, Q^* must be equal to this constant. The functional Hill number D can be interpreted as "the effective number of equally abundant and (functionally) equally distinct species" with a constant distance Q^* for all species pairs (i, j) with $i \neq j$.

We notice that there are two drawbacks in this framework: (i) The functional Hill number D can not be explicitly solved from Eq. (S10); (2) When all different species in the assemblage are equally distinct (i.e., $d_{ij} = Q^*$ for all $i \neq j$), the functional Hill number D should reduce to the traditional or taxonomic Hill number. But according to (S10) and (S11), we have

$$\sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} (p_i p_j)^q = (D-1)D^{1-2q}. \tag{S12}$$

Apparently, D dose not reduce to TD_q in (S7) or (S8).

In order to overcome those two drawbacks, we introduce a new pair-wise functional distance matrix

$$\Delta' = (d'_{ij}) = \begin{pmatrix} \lambda_1 & d_{12} & d_{13} & \cdots & d_{1N} \\ d_{21} & \lambda_2 & d_{23} & \cdots & d_{2N} \\ d_{31} & d_{32} & \lambda_3 & \cdots & d_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{N1} & d_{N2} & d_{N3} & \cdots & \lambda_N \end{pmatrix}, \tag{S13}$$

where d_{ij} denotes the original functional distance between species-i and j, with $d_{ij} = d_{ji} \geq 0$ $(i \neq j)$, and

$$\lambda_i := \frac{\sum_{j \neq i}^{N} d_{ij}}{N-1}, \tag{S14}$$

is the average functional distance between species-i and all the other species. Note that, when different species are equally distinct with a constant pairwise distance, λ is equal to this constant.

Equipped with the modified distance matrix Δ', we can now derive the functional Hill number as follows:

$$\sum_{i=1}^{N}\sum_{j=1}^{N} d'_{ij}(p_i p_j)^q = \sum_{i=1}^{D}\sum_{j=1}^{D} Q'(\frac{1}{D}\frac{1}{D})^q \tag{S15}$$

where

$$Q' := \sum_{i=1}^{N}\sum_{j=1}^{N} d'_{ij} p_i p_j. \tag{S16}$$

Note that Q' shares almost the exact form as *Rao's quadratic entropy* $Q := \sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij} p_i p_j$, a classic functional diversity measure that characterizes the mean distance between any two randomly chosen taxa in the sample [15]. From (S15), we can derive a parametric class of functional Hill numbers:

$$D_q(Q') := \left(\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{d'_{ij}}{Q'}(p_i p_j)^q\right)^{\frac{1}{2(1-q)}} \quad \text{for} \quad q \neq 1, \tag{S17}$$

and

$$D_1(Q') := \lim_{q \to 1} D_q(Q') = \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{d'_{ij}}{Q'} p_i p_j \log(p_i p_j)\right]. \tag{S18}$$

The functional Hill number of order q, i.e., $D_q(Q')$, can be interpreted as "the effective number of equally abundant and (functionally) equally distinct species" with a constant distance Q' for all species pairs. Thus if $D_q(Q') = v$, it just means that the functional diversity (of order q) of a given real sample with relative abundance profile $\mathbf{p} = [p_1, \cdots, p_N]$ and pair-wise functional distance matrix Δ' is the same as that of an idealized sample containing v equally abundant species (i.e., $\tilde{\mathbf{p}} = [1/v, \cdots, 1/v]$), and a $v \times v$ distance matrix with constant pair-wise functional distance Q', i.e.,

$$\begin{pmatrix} Q' & Q' & Q' & \cdots & Q' \\ Q' & Q' & Q' & \cdots & Q' \\ Q' & Q' & Q' & \cdots & Q' \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Q' & Q' & Q' & \cdots & Q' \end{pmatrix}_{v \times v}. \tag{S19}$$

Hence, the functional diversity of a real sample can be defined as the column (or row) sum of the above idealized $v \times v$ constant matrix, yielding

$$\mathrm{FD}_q(Q') := D_q(Q') \cdot Q' = \left(\sum_{i=1}^{N} \sum_{j=1}^{N} \frac{d'_{ij}}{Q'} (p_i p_j)^q \right)^{\frac{1}{2(1-q)}} \cdot Q' \quad \text{for} \quad q \neq 1, \tag{S20}$$

and

$$\mathrm{FD}_1(Q') := D_1(Q') \cdot Q' = \exp\left[-\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{d'_{ij}}{Q'} p_i p_j \log(p_i p_j) \right] \cdot Q'. \tag{S21}$$

A few special cases:

- If all the different species in the sample are equally distinct, i.e., $d'_{ij} = Q'$ for all species pairs (i, j), $i \neq j$; and $d'_{ii} = \lambda_i = Q'$, then the functional Hill number $D_q(Q')$ reduces to the ordinary Hill number TD_q, and $\mathrm{FD}_q(Q') = \mathrm{TD}_q \cdot Q'$.

- If all the species in the sample are equally aundant, i.e., $p_i = 1/N$ for $i = 1, \cdots, N$, then for any distance matrix Δ', we have $Q' = N^{-2} \sum_{i=1}^{N} \sum_{j=1}^{N} d'_{ij}$, and $D_q(Q') = N$ for all orders of q, and $\mathrm{FD}_q(Q') = D_q(Q') \cdot Q' = N^{-1} \sum_{i=1}^{N} \sum_{j=1}^{N} d'_{ij}$.

4.3 Functional redundancy

In the literature of ecology, the *functional redundancy* (FR) of a sample is often considered to be the part of the taxonomic diversity (TD) that can not be explained by the functional diversity (FD) [16, 17, 18]. Hence FR is typically defined to be the difference between TD and FD:

$$\mathrm{FR} := \mathrm{TD} - \mathrm{FD}. \tag{S22}$$

In Ref. [16, 17], TD was chosen to be the Gini-Simpson index: $\mathrm{GSI} := 1 - \sum_{i=1}^{N} p_i^2$, and FD be Rao's quadratic entropy $Q := \sum_{i=1}^{N} \sum_{j=1}^{N} d_{ij} p_i p_j$. Hence,

$$\mathrm{FR} = \mathrm{GSI} - Q = 1 - \sum_{i=1}^{N} p_i^2 - \sum_{i=1}^{N} \sum_{j=1}^{N} d_{ij} p_i p_j = \sum_{i=1}^{N} \sum_{j=1}^{N} (1 - d_{ij}) p_i p_j. \tag{S23}$$

Consider $d_{ij} \in [0, 1]$. We can verify that FR defined in Eq.(S23) takes the range $[0, 1]$.

- When the species are completely different in their functions, i.e., for all pairs $i \neq j$, $d_{ij} = 1$; and $d_{ii} = 0$, then $Q = \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} p_i p_j = \sum_{i=1}^{N} p_i (1 - p_i) = \mathrm{GSI}$, hence $\mathrm{FR} = 0$. This makes perfect sense, because all the different species are functionally different, hence the functional redundancy of this sample (community) is zero.

- When all the species are functionally identical: $d_{ij} = 0$ for $i, j = 1, \cdots, N$, then $Q = 0$, and FR = GSI. In this case, the functional redundancy of this sample (community) is maximized and equal to the Gini-Simpson index of taxa diversity.

- When all the species are functionally identical: $d_{ij} = 0$ for $i, j = 1, \cdots, N$, and in addition the number of species N is very large and the species are equally abundant $p_i = 1/N \to 0$, then FR = GSI $\to 1$.

Now we consider the Hill number based taxonomic diversity TD_q and functional diversity $\mathrm{FD}_q(Q')$, and define a parametric class of functional redundancy:

$$\mathrm{FR}_q(Q') := \mathrm{TD}_q - \mathrm{FD}_q(Q'). \tag{S24}$$

For $q \neq 1$,

$$\mathrm{FR}_q(Q') := \left[\sum_{i=1}^{N} p_i^q\right]^{\frac{1}{(1-q)}} - \left[\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{d'_{ij}}{Q'}(p_i p_j)^q\right]^{\frac{1}{2(1-q)}} \cdot Q'. \tag{S25}$$

For $q = 1$,

$$\mathrm{FR}_1(Q') := \exp\left[-\sum_{i=1}^{N} p_i \log p_i\right] - \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{d'_{ij}}{Q'} p_i p_j \log(p_i p_j)\right] \cdot Q'. \tag{S26}$$

Consider two extreme cases:

(1) Zero redundancy: $d_{ij} = 1$ for all species pair $i \neq j$. Then, we have $d'_{ij} = 1$ for all i and j, and $Q' = 1$. For $q \neq 1$

$$\mathrm{FD}_q(Q) = \left[\sum_{i=1}^{N}\sum_{j=1}^{N}(p_i p_j)^q\right]^{\frac{1}{2(1-q)}} = \left[\left(\sum_{i=1}^{N} p_i^q\right) \cdot \left(\sum_{j=1}^{N} p_j^q\right)\right]^{\frac{1}{2(1-q)}} = \mathrm{TD}_q.$$

For $q = 1$,

$$\begin{aligned}
\mathrm{FD}_1(Q) &= \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} p_i p_j \log(p_i p_j)\right] \\
&= \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} p_i p_j \log(p_i) - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} p_i p_j \log(p_j)\right] \\
&= \exp\left[-\frac{1}{2}\sum_{i=1}^{N} p_i \log(p_i) - \frac{1}{2}\sum_{j=1}^{N} p_j \log(p_j)\right] \\
&= \mathrm{TD}_1.
\end{aligned} \tag{S27}$$

Therefore, for $d'_{ij} = 1$ we have $\mathrm{FR}_q(Q') = \mathrm{TD}_q - \mathrm{FD}_q(Q') = 0$.

(2) Complete redundancy: $d_{ij} = 0$ for all species pair $i \neq j$, then $d'_{ij} = 0$, $Q' = 0$, $\mathrm{FD}_q(Q') = D_q(Q') \cdot Q' = 0$, and $\mathrm{FR}_q(Q') = \mathrm{TD}_q$ is maximized.

5 Constructing the GCN

We downloaded the bacterial and archaeal genomes from the Integrated Microbial Genomes (IMG) database (v3.5) [4]. In this version of IMG, there are 2,833 strains (genomes) annotated by 6,727 KEGG Ortholog (KO) gene families. In constructing the reference genomic content network (GCN), we randomly chose a representative strain (genome) for each species to reduce the culturing and sequencing bias for certain species (e.g., *Escherichia coli*). There are in total 1,386 species and 6,685 KOs in this network. In main text Fig. 2a, this reference GCN is visualized at the phylum level for taxa nodes and the KEGG super-pathway level for functional nodes. The genome of each phylum is obtained by averaging the genomes of all the species belonging to this phylum. The bar height of each phylum corresponds to genome size of that phylum. The thickness of a link connecting a phylum and a pathway is proportional to the number of KOs belonging to the pathway. The network properties of this reference GCN are shown in main text Fig. 2b-e, where each taxon node represents a species, and each functional node represents a KO.

A large portion of human microbiome consists of uncultivated organisms, whose genomes and functional traits are unknown. Those microbes play important roles in the functional capacity of the whole community. To predict their gene contents, we leveraged the existing pipeline of PICRUSt (phylogenetic investigation of communities by reconstruction of unobserved states) [7], a computational approach to predict the functional composition of a metagenome using marker gene (e.g., 16S rRNA) data and a database of reference genomes. The first step in the PICRUSt workflow is gene content inference, which predicts the gene content (including the marker gene copy number) for each OTU with unknown gene content, given a reference OTU tree (derived from Greengenes [6]) and a reference gene content table (i.e., counts of genes for reference OTUs with known gene content, derived from IMG [4]). The basic idea is that despite various important forms of microbial genome plasticity (e.g., gene loss, duplication, or gene transfer), the genes present in microbial genomes are much more similar among related bacteria or archaea than distant relatives. Therefore, when sufficient genome sequences are available, it is possible to predict which gene families are present in a given microbial OTU from phylogeny alone. To do this, PICRUSt uses existing gene content and 16S copy numbers from reference bacterial and archaeal genomes in the IMG database, and performs an extended ancestral state reconstruction to make predictions of gene content for all organisms represented in the Greengenes phylogenetic tree of 16S sequences. This enables us to construct the extended reference GCN, consisting of 203,452 KO-annotated genomes (OTUs) and 6,909 KOs. The network properties of this extended

reference GCN are shown in SI Sec.1, Fig. S3b.

6 Randomizing the GCN

To identify the structural features of the GCN that detemine the functional redundancy and functional diversity of microbial communities, we randomized the GCN using four different randomization schemes, yielding four different null models:

- (Null-1) Complete randomization. We keep the number of taxa (N) and number of genes (M) unchanged, but otherwise completely rewire the links between taxa and genes. The randomized GCN displays Poisson-like degree distributions for both taxon- and gene-nodes (Fig. S7b,g).

- (Null-2) Taxon-degree preserving randomization. We keep N, M, and the degree of each taxon node unchanged, but selects randomly the gene nodes that link to each taxon. The randomized GCN will have exactly the same taxon degree distribution as the original GCN, but a Poisson-like gene degree distribution (Fig. S7c,h).

- (Null-3) Gene-degree preserving randomization. Here we keep N, M, and the degree of each gene unchanged, but selects randomly the taxa that link to each gene. The randomized GCN will have exactly the same gene degree distribution as the original GCN, and a Poisson-like taxon degree distribution (Fig. S7d,i).

- (Null-4) Gene-degree and taxon-degree preserving randomization. Here we keep N, M, and the degree of each gene and the degree of each taxon unchanged, but randomly rewire the links between taxon nodes and gene nodes. This is achieved as follows. In each step, we randomly pick two links (i, a) and (j, b), where the taxon nodes $i \neq j$ and the gene nodes $a \neq b$. Then we rewire those two links by switching their taxon nodes (or gene nodes), yielding two new links (i, b) and (j, a). In case gene-a (gene-b) already exists in the genome of taxon-j (taxon-i), we simply increase its copy number by one. We repeat this process until every link has been rewired for at least once. The gene degree distribution and the taxon degree distribution of the randomized GCN will be exactly the same as those of the original GCN. (Fig. S7e,j)

Note that the weight of each link in the original GCN is the gene copy number, which is a non-negative integer. In all the randomization processes, we treat the weighted link as multiple links of unit weight.

7 Nestedness of the GCN

As a classical concept in ecology, *nestedness* characterizes the nested structure of ecological systems, such as the species-site network (describing the distribution of species across geographic locations), and the species-species interaction networks (e.g., host-parasite, plant-pollinator interactions) [9, 19, 20, 21, 22, 23, 24]. Roughly speaking, an ecological system is said to be *nested* if the items belonging to "smaller" elements (e.g., a small island containing few species, or a specialist species with few interactions) tend to be a subset of the items belonging to "larger" elements (e.g., a large island containing many species, or a generalist species with many interacting partners). Mathematically, those ecological systems can be represented as bipartite graphs with two types of nodes, e.g., sites and species, hosts and parasites, plants and pollinators, etc. In this work, we focus on the genomic content network of microbial communities, which is also a bipartite graph with two types of nodes: species and genes.

Consider a general bipartite graph with N type-1 nodes and M type-2 nodes. The structure of this bipartite graph can be represented by its $N \times M$ binary incidence matrix $\mathbf{B} = (B_{ia})$, where $B_{ia} = 1$ if there is a link connecting the i-th type-1 node and the a-th type-2 node, and 0 otherwise. Mathematically, nestedness can be defined as a property of the incidence matrix $\mathbf{B}$. If there exists a permutation of rows and columns such that the set of links in row-i contains the links in row-$(i+1)$, and the set of links in column-a contains those in column-$(a+1)$, then $\mathbf{B}$ is a perfectly nested binary matrix. For example, consider the mainland and a series of islands sorted according to their distances to the mainland. The mainland contains all the species, the first island has a subset of species in the mainland, the second island has a subset of species in the first island, etc.

To emphasize and visualize the nested structure of the incidence matrix, we can use the Nestedness Temperature Calculator (NTC) [8] based on the BINMATNESS algorithm [25], which also provides a nestedness measure. But NTC is generally time consuming for large incidence matrix. In this work we adopt the classical NODF measure (Nestedness metric based on Overlap and Decreasing Fill) to characterize the nested structure of a general bipartite graph [9]. Comparing with alternative nestedness measures, NODF reduces potential bias due to network size and shape.

For a given bipartite graph (say, the species-gene network) with binary incidence matrix $\mathbf{B}$. The (unweighted) degree of the i-th species node is $k_i = \sum_{a=1}^{M} B_{ia}$, and the (unweighted) degree of the a-th gene node is $k_a = \sum_{i=1}^{N} B_{ia}$. The number of common genes shared by the genomes of the i-th and the j-th species is given by $P_{ij} = \sum_{a=1}^{M} B_{ia}B_{ja}$. Similarly, the number of common species that both the a-th and the b-th genes appear in their genomes is given by $Q_{ab} = \sum_{i=1}^{N} B_{ia}B_{ib}$. Define $\tilde{P}_{ij} = 0$ if $k_i = k_j$, and $\tilde{P}_{ij} = P_{ij}/\min(k_i, k_j)$ otherwise. Similarly, define $\tilde{Q}_{ab} = 0$ if $k_a = k_b$, and $\tilde{Q}_{ab} = Q_{ab}/\min(k_a, k_b)$ otherwise. The NODF

measure is defined as follows:

$$\text{NODF} = \frac{\sum_{i<j}^{N} \tilde{P}_{ij} + \sum_{a<b}^{M} \tilde{Q}_{ab}}{\frac{N(N-1)}{2} + \frac{M(M-1)}{2}}. \tag{S28}$$

8 Human microbiome datasets

The microbiome data analyzed in this work are all from published studies. The original experiments and corresponding power analysis have been reported in previous publications.

To quantitatively compare the functional redundancy and diversity of microbiome samples collected from different body sites, we leveraged two large-scale metagenomics studies:

- Human Microbiome Project (HMP) [2, 3]. We analyzed the 16S rRNA gene-based dataset (variable regions V3 to V5) of the human microbiome samples from HMP. The dataset is available at http://hmpdacc.org/ and https://qiita.ucsd.edu/. This dataset covers 18 body sites in five areas: Gut (one site: stool (388 samples)); Airway (one site: anterior nares (387 samples)); Oral (nine sites: attached keratinized gingiva (359 samples), buccal mucosa (362 samples), hard palate (366), palatine tonsils (373 samples), saliva (346 samples), subgingival plaque (373 samples), supragingival plaque (374 samples), throat (371 samples), tongue dorsum (365 samples)); Skin (four sites: left antecubital fossa (340 samples), left retroauricular crease (396 samples), right antecubital fossa (344 samples), right retroauricular crease (414 samples)); Vagina (three sites: introitus (141 samples), mid-vagina (152 samples), and posterior fornix (169 samples)). For each body site, we only choose one sample from each subject in our analysis.

- Student Microbiome Project (SMP) [1]. We analyzed the 16S rRNA gene-based dataset (variable region V4) from 85 college-aged adults in the SMP study. The dataset is available at https://qiita.ucsd.edu/. This dataset covers four body sites: gut (917 samples), tongue (947 samples), forehead (939 samples) and palm (930 samples). In case there are multiple samples measured for one subject, we used the sample from the first visit.

For both datasets (HMP and SMP), sequence depth for all samples are rarefied to 5,000 reads/sample. Those samples whose original sequence depth is lower than this level are excluded in our analysis. We anticipate that using other sequence depth will not qualitatively change our results.

To analyze the development of the functional redundancy, we analyzed gut microbiome samples of infants in the beginning of their lives, leveraging two existing longitudinal studies:

- Microbial development profile during the first 2 years of life in a cohort of 43 U.S. infants [26]. Stool samples from the mothers were also collected for comparative analysis. DNA was extracted from stool samples. The V4 region of the bacterial 16S rRNA gene was sequenced using the Illumina MiSeq platform in five different sequencing runs. All the samples with antibiotic exposure are excluded in our analysis.

- Vitamin D Antenatal Asthma Reduction Trial (VDAART), a clinical trial of prenatal vitamin D supplementation and outcomes in pregnant women and their offspring [27]. The stool samples from 668 infants were collected longitudinally at 3-6 months, 1 year, 3 years, 4 years, and 5 years of age. The stool samples from 120 mothers at delivery were also collected. DNA extractions were performed on stool samples, and analyzed by Roche 454 Titanium pyrosequencing of the V3-V5 hypervariable regions of the bacterial rRNA 16S gene segments. All the samples with antibiotic exposure are excluded in our analysis.

To compare the functional redundancy of gut microbiome samples from healthy subjects and subjects with various diseases, we analyzed six different datasets:

- Recurrent *Clostridium difficile* infection (rCDI) [28]. Stool samples of patients with rCDI and their healthy donors were collected from a fecal microbiota transplantation (FMT) clinical trial, which was approved by the Partners Human Research Committee as well as by the US Food and Drug Administration (FDA) (Investigational New Drug application number 15199) and registered at Clinical Trials.gov (NCT01704937). Informed consent was obtained from all participants. Microbial samples from 17 patients with rCDI were analyzed in the groups of pre-FMT and post-FMT. Stool samples from 10 healthy donors were also analyzed for comparisons. The V4 region of the 16S rRNA gene was sequenced using an Illumina MiSeq.

- Inflammatory bowel disease (IBD) [29]. We analyzed the mucosal- and lumen-associated microbiota in a large, well-characterized inception cohort for children with Crohn's disease (CD), a particular type of IBD. This cohort included subjects from 3 to 17 years of age with a well-established diagnosis of CD ($n = 447$), and control subjects with noninflammatory conditions ($n = 221$). Mucosal tissue biopsies (terminal ileum and rectum) were collected as part of the diagnostic colonoscopic examination prior to the initiation of treatment. A subset of the enrolled patients ($n = 233$) also provided a fecal sample prior to treatment start. The diagnosis and disease categorization were confirmed after a minimum of 6 months' follow-up, based on a combination of endoscopic, histological, and radiological investigations. A total of 1,321 samples, including 630 ileal and 387 rectal tissue biopsies and 304 stool samples, were submitted for microbiome profiling using 16S rRNA gene sequencing (V4 variable region) on the Illumina MiSeq platform. The dataset is available at https://qiita.ucsd.edu/.

- Asthma [27]. We analyzed the stool samples of infants at age 3 from the VDAART study. The subjects were divided into groups of asthma (79 subjects) and control (362 subjects) according to the clinical outcomes.

- Obesity [30]. Stool samples of 31 monozygotic twin pairs, 23 dizygotic twin pairs and, where available, their mothers (46 subjects) were analyzed. Subjects are divided into three groups: obesity (BMI>30), overweight (25<BMI<30), and leanness (BMI= 18.5 $\sim$ 24.9). The bacterial lineages present in the stool samples were analyzed through 16S rRNA sequencing (V2/V6 regions). The dataset is available at https://qiita.ucsd.edu/.

- Autism [31]. Stool samples from 20 neurotypical and 20 autistic children accompanied by a survey of both autistic severity and gastrointestinal symptoms were collected. Frozen fecal samples were shipped overnight to Arizona State University with a cold pack, and stored in -80°C until DNA extraction. By pyrosequencing the V2/V3 regions in bacterial 16S rDNA from fecal DNA samples, the microbial composition were characterized. The sequences data can be downloaded from the Sequence Research Archive (SRA) at NCBI with the accession number SRA053656.

- Parkinson's disease (PD) [32]. A total of 197 PD cases and 130 controls were included. Stool samples were collected at home using DNA/ RNA-free sterile swabs and shipped immediately via standard U.S. postal service at ambient temperature. Microbial composition was determined by 16S rRNA gene sequencing (V4 region) of DNA extracted from stool, using an Illumina MiSeq platform. The dataset is available at https://qiita.ucsd.edu/.

References

[S1] Flores, G. E. *et al.* Temporal variability is a personalized feature of the human microbiome. *Genome Biology* **15**, 531 (2014).

[S2] The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207 (2012).

[S3] Methé, B. A. *et al.* A framework for human microbiome research. *Nature* **486**, 215 (2012).

[S4] Markowitz, V. M. *et al.* Img: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Research* **40**, D115–D122 (2011).

[S5] Uchiyama, I., Higuchi, T. & Kawai, M. Mbgd update 2010: toward a comprehensive resource for exploring microbial genome diversity. *Nucleic Acids Research* **38**, D361–D365 (2009).

[S6] DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. *Applied and Environmental Microbiology* **72**, 5069–5072 (2006).

[S7] Langille, M. G. *et al.* Predictive functional profiling of microbial communities using 16s rrna marker gene sequences. *Nature Biotechnology* **31**, 814–821 (2013).

[S8] Atmar, W. & Patterson, B. D. The measure of order and disorder in the distribution of species in fragmented habitat. *Oecologia* **96**, 373–382 (1993).

[S9] Almeida-Neto, M., Guimarães, P., Guimarães, 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).

[S10] McCune, B., Grace, J. B. & Urban, D. L. *Analysis of ecological communities*, vol. 28 (MjM software design Gleneden Beach, OR, 2002).

[S11] Sørensen, T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biol. Skr.* **5**, 1–34 (1948).

[S12] Sokal, R. R. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin* **38**, 1409–1438 (1958).

[S13] Hill, M. O. Diversity and evenness: A unifying notation and its consequences. *Ecology* **54**, 427–432 (1973).

[S14] Chiu, C.-H. & Chao, A. Distance-based functional diversity measures and their decomposition: A framework based on hill numbers. *PLOS ONE* **9**, e100014 (2014).

[S15] Rao, C. R. Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology* **21**, 24–43 (1982).

[S16] de Bello, F. *et al.* Importance of species abundance for assessment of trait composition: an example based on pollinator communities. *Community Ecology* **8**, 163–170 (2007).

[S17] Pillar, V. D. *et al.* Functional redundancy and stability in plant communities. *Journal of Vegetation Science* **24**, 963–974 (2013).

[S18] Kang, S. *et al.* Functional redundancy instead of species redundancy determines community stability in a typical steppe of inner mongolia. *PLoS ONE* **10**, 1–11 (2015).

[S19] Patterson, B. D. & Atmar, W. Nested subsets and the structure of insular mammalian faunas and archipelagos. *Biological Journal of the Linnean Society* **28**, 65–82 (1986).

[S20] Patterson, B. D. The principle of nested subsets and its implications for biological conservation. *Conservation Biology* **1**, 323–334 (1987).

[S21] Brualdi, R. A. & Sanderson, J. G. Nested species subsets, gaps, and discrepancy. *Oecologia* **119**, 256–264 (1999).

[S22] Bascompte, J., Jordano, P., Melián, C. J. & Olesen, J. M. The nested assembly of plant–animal mutualistic networks. *Proceedings of the National Academy of Sciences of U.S.A.* **100**, 9383–9387 (2003).

[S23] Rezende, E. L., Lavabre, J. E., Guimarães, P. R., Jordano, P. & Bascompte, J. Non-random coextinctions in phylogenetically structured mutualistic networks. *Nature* **448**, 925 (2007).

[S24] Saavedra, S., Stouffer, D. B., Uzzi, B. & Bascompte, J. Strong contributors to network persistence are the most vulnerable to extinction. *Nature* **478**, 233 (2011).

[S25] Rodríguez-Gironés, M. A. & Santamaría, L. A new algorithm to calculate the nestedness temperature of presence–absence matrices. *Journal of Biogeography* **33**, 924–935 (2006).

[S26] Bokulich, N. A. *et al.* Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Science Translational Medicine* **8**, 343ra82–343ra82 (2016).

[S27] Litonjua, A. A. *et al.* The Vitamin D Antenatal Asthma Reduction Trial (VDAART): rationale, design, and methods of a randomized, controlled trial of vitamin d supplementation in pregnancy for the primary prevention of asthma and allergies in children. *Contemporary Clinical Trials* **38**, 37–50 (2014).

[S28] Youngster, I. *et al.* Fecal microbiota transplant for relapsing clostridium difficile infection using a frozen inoculum from unrelated donors: a randomized, open-label, controlled pilot study. *Clinical Infectious Diseases* **58**, 1515–1522 (2014).

[S29] Gevers, D. *et al.* The treatment-naive microbiome in new-onset crohn's disease. *Cell host & microbe* **15**, 382–392 (2014).

[S30] Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480 (2009).

[S31] Kang, D.-W. *et al.* Reduced incidence of prevotella and other fermenters in intestinal microflora of autistic children. *PLoS ONE* **8**, e68322 (2013).

[S32] Hill-Burns, E. M. *et al.* Parkinson's disease and parkinson's disease medications have distinct signatures of the gut microbiome. *Movement Disorders* **32**, 739–749 (2017).