Two Competing Guilds as a Core Microbiome Signature for Chronic Diseases

Guojun Wu[2,3]*, Ting Xu[1]*, Naisi Zhao[6]*, Yan Y. Lam[7]*, Xiaoying Ding[5]*, Dongqin Wei[4], Jian Fan[4], Yajuan Shi[4], Xiaofeng Li[4], Mi Li[4], Shenjie Ji[4], Xuejiao Wang[5], Huaqing Fu[1], Feng Zhang[8], Yongde Peng[5]†, Yu Shi[4]†, Chenhong Zhang[1]†, Liping Zhao[1,2,3]†

[1]State Key Laboratory of Microbial Metabolism and Ministry of Education Key Laboratory of Systems Biomedicine, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China

[2]Department of Biochemistry and Microbiology, School of Environmental and Biological Sciences and Center for Microbiome, Nutrition, and Health, New Jersey Institute for Food, Nutrition, and Health, Rutgers, The State University of New Jersey, New Brunswick, NJ 08901, USA

[3]Rutgers-Jiaotong Joint Laboratory for Microbiome and Human Health, New Brunswick, NJ, USA

[4]Department of Endocrinology and Metabolism, Qidong People's Hospital, Jiangsu 226200, China

[5]Department of Endocrinology and Metabolism, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200080, China

[6]Department of Public Health and Community Medicine, School of Medicine, Tufts University, Boston, MA, 02111, USA

[7]Gut Microbiota and Metabolism Group, Centre for Chinese Herbal Medicine Drug Development, School of Chinese Medicine, Hong Kong Baptist University, Units 201-207, Building 15W, 15 Science Park West Avenue, Pak Shek Kok, N.T., Hong Kong, China

[8]Nutrition Department (Clinical Study Center of Functional Food), The Affiliated Hospital of Jiangnan University Wuxi, Jiangsu 214122, China

*co-first author

†co-corresponding author

Summary Paragraph

Gut microbiota may work as an essential organ and its members interact closely with each other and form a higher-level organization called guilds. How such guild-level structure supports the gut microbiota to stably provide essential health-relevant functions to the host remains elusive. With high quality metagenome-assembled genomes as network nodes, here we identified a core microbiome signature made up of two robust competing guilds that together correlate with a wide range of host health conditions. Genomes in these two guilds kept their ecological relationship unchanged despite experiencing profound abundance changes during a 3-month high fiber intervention and 1-year follow-up in patients with type 2 diabetes. The genomes of one guild harbored more genes for plant polysaccharide degradation and butyrate production, while the other guild had more genes for virulence or antibiotic resistance. A Random Forest regression model showed that the abundance distributions of these genomes were associated with 41 out of 43 bio-clinical parameters in the study cohort. With these genomes as reference, Random Forest modeling successfully classified case and control of 8 chronic diseases in 12 independent metagenomic datasets from 1,816 participants across ethnicity and geography. This core microbiome signature may facilitate ecological management of chronic diseases.

Introduction

Over eons of co-evolution, humans have developed a robust symbiotic relationship with their gut microbiome[1,2]. The gut microbiome supports the host's homeostasis in metabolism, immunity, development, and behavior, etc.[3] It has been regarded as an essential organ because the attenuation or loss of such health-relevant functions of a dysbiotic gut microbiome has been linked with the initiation and/or progression of many chronic diseases, including type 2 diabetes

(T2DM)[4-6]. However, the underlying gut microbiome structural signatures that support the stable provision of health-relevant functions to the host remain to be identified.

The gut microbiota is a complex adaptive system[7], in which the minimum responding units to environmental perturbations are bacterial genomes[8]. More importantly, genomes are not independent microbiome features. They form ecological interactions, such as competition or cooperation, with each other and organize themselves into a higher-level structure called "guilds"[9]. Each guild is a potential functional group of bacteria in the gut ecosystem. Guild members may have widely diverse taxonomic backgrounds but thrive or decline together and thus show co-abundant behavior. Guild-level variations have been positively or negatively correlated with disease phenotypes and their members have been demonstrated as having causal role in host disease phenotypes[10,11]. Although a suite of microbiome-wide association studies (MWAS) has attempted to identify the microbiome signatures (using features such as genes, pathways, taxa, etc.) that are associated with disease phenotypes [12-15], genomes and their guild-level organization have not been extensively employed to describe the ecological structure that supports the stable provision of health-relevant functions to the host.

To this end, we suggest a genome-centric approach which is based on high-quality draft genomes assembled directly from metagenomic datasets (high-quality metagenome-assembled genomes, HQMAGs). This approach uses genomes as nodes of ecological networks and their guild-level aggregations as ecologically meaningful features for identifying microbiome signatures of chronic diseases. Furthermore, this approach is completely data-driven and unsupervised, requiring no reference databases or prior knowledge.

In this study, we hypothesized that bacteria required for providing essential health-relevant functions to the host[2] should maintain stable ecological interactions with each other to form

robust guilds[16,17]. To identify microbiome signatures that are based on stable interactions among HQMAGs, we randomized T2DM patients at baseline (M0) to receive either 3-month (M3) of high fiber intervention (W group; n = 74) or standard care (U group; n= 36) followed by a one-year follow-up (M15) in an open label, controlled trial (Fig. 1A and Fig. S1). The high fiber intervention was used to exert a positive environmental perturbation to dramatically and reversibly change the abundance of members of the gut microbiome[10,11]. Co-abundance network analysis at each of the three time points enabled us to identify genome pairs that can keep their correlations unchanged despite significant community-wide abundance changes caused by the perturbations. We found that these robust genome pairs were from 141 HQMAGs and these genomes formed two competing guilds. These two guilds were organized as the two competing ends of a robust seesaw-like network, whenever one guild increased, the other decreased in abundance. Together, these seesaw networked genomes supported machine learning models for predicting the response of a wide range of metabolic phenotypes to dietary intervention in the T2DM cohort, as well as for classifications of case and control of 12 independent metagenomic datasets from 1,816 subjects across different cohorts and various chronic diseases including T2DM, atherosclerotic cardiovascular disease (ACVD), liver cirrhosis (LC), inflammatory bowel diseases (IBD), colorectal cancer (CRC), ankylosing spondylitis (AS), schizophrenia, and Parkinson's disease (PD), suggesting that we may have identified a core microbiome signature across different chronic diseases.

Results

Reversible changes in the gut microbiota associate with reversible changes of host metabolic phenotypes

Dietary fiber intake in the U group remained unchanged throughout the study, whereas W group had a significant increase in the intake of dietary fibers from M0 to M3 and a decrease from M3 to M15 (Fig. 1B). Compared with the U group, fiber intake was significantly higher in the W group at both M3 and M15 (Fig. 1B), but energy and macronutrient consumption were similar across the study period (Fig. S2).

To investigate the gut microbial responses to the introduction and withdrawal of the high fiber intervention, we performed shotgun metagenomic sequencing on 315 fecal samples collected from 110 patients of the W and U group, among whom 95 patients provided samples at all 3 time points and 15 provided samples at M0 and M3 only (Table S2, Fig S1). To achieve genome-level resolution, we reconstructed 1,845 non-redundant high-quality draft genomes (HQMAGs, two HQMAGs were collapsed into one if the average nucleotide identity, ANI, between them was > 99%) from the metagenomic datasets. These HQMAGs accounted for more than 70% of the total reads. In the context of beta-diversity measured via the Bray-Curtis distance, the overall structure of the gut microbiota in the W group significantly changed from M0 to M3 (PERMANOVA test, $P < 0.001$) and returned to that of M0 at M15; there was no difference in the U group across the 3 timepoints (Fig. 1C, D). Similar changes in alpha-diversity based on Shannon and Simpson indices were also observed (Fig. S3). These results showed that high fiber intervention induced significant structural changes of the gut microbiota[11], however the gut microbiota reverted to baseline after the intervention was withdrawn indicating a high resilience in community structure.

To determine if host metabolic phenotypes would show similar reversible changes as the gut microbiota, we examined 43 bio-clinical parameters across the 3 time points. Hemoglobin A1c (HbA1c) in the U group showed no changes throughout the trial. The high fiber intervention

reduced the level of HbA1c in the W group from M0 to M3 by 15.22% ± 9.82% (mean ± s.d.), and such reduction was significantly bigger than what was observed in the U group. At one-year follow-up of the W group, HbA1c was significantly increased from M3 but remained lower than at M0 (Fig. 1E). The proportion of patients who achieved adequate glycemic control (HbA1c ⩽ 7%) was significantly higher in the W group (61.6 % versus 33.3% in the U group) at M3, but showed no difference between the two groups at M15 (Fig. 1F). The level of fasting blood glucose and postprandial glucose in meal tolerance test followed a similar trend as HbA1c (Fig. 1G, H). The W group also showed an alleviation of inflammation, hyperlipidemia, obesity, and T2DM complications from M0 to M3 but these parameters rebounded at one-year follow-up (Table S3). These results indicate that changes of the host metabolic phenotypes were associated with the reversible changes of the gut microbiota in response to the introduction or withdraw of the high fiber intervention.

Genome pairs with stable interactions form a seesaw-like network of two competing guilds

To facilitate the identification of genome pairs that keep their ecological interactions stable during the trial, particularly in the W group with profound microbiota and host phenotypic changes, we constructed a co-abundance network for each time point based on the abundance matrix of the HQMAGs representing the prevalent microbes. Co-abundance network is a data-driven way to investigate ecological interactions between microbes across habitats[18,19]. A total of 477 HQMAGs were selected for network construction because they were detectable in more than 75% of the samples at each time point in the W group. They also accounted for ~60% of the total abundance of the 1,845 HQMAGs. In the W group, we calculated pairwise correlations of all 113,526 possible genome pairs among these 477 prevalent HQMAGs based on their abundance

across the patients at each time point and constructed 3 co-abundance networks ($\boldsymbol{G}_{M0}$, $\boldsymbol{G}_{M3}$ and $\boldsymbol{G}_{M15}$) (Figure 2A, Table S4). The three networks were of similar order $\boldsymbol{S}$, i.e., the total number of nodes (HQMAGs), $\boldsymbol{S}_{M0}$(442), $\boldsymbol{S}_{M3}$(421), and $\boldsymbol{S}_{M15}$(429), but they varied considerably in their size $\boldsymbol{L}$, i.e., the total number of edges (correlations), $\boldsymbol{L}_{M0}$(4,231), $\boldsymbol{L}_{M3}$(2,587) and $\boldsymbol{L}_{M15}$(4,592). $\boldsymbol{L}$ in $\boldsymbol{G}_{M3}$ decreased to 61.14% of that in $\boldsymbol{G}_{M0}$ and rebounded back in $\boldsymbol{G}_{M15}$ to 108.53% of that in $\boldsymbol{G}_{M0}$. This pattern was confirmed by changes in connectance, which is defined as the proportion of realized ecological interactions among the potential ones (in undirected network, connectance= $\frac{L}{S(S-1)/2}$, range: [0,1])[20]. Connectance decreased from 0.043 in $\boldsymbol{G}_{M0}$ to 0.029 in $\boldsymbol{G}_{M3}$ and rebounded to 0.050 in $\boldsymbol{G}_{M15}$. Changes in L and connectance showed that high fiber intervention dramatically reduced the correlations among the prevalent genomes in the network. In addition, we found that the distributions of degree, i.e. the number of edges a node has, fit well with a power-law model (Fig. S4, R2 values $\boldsymbol{G}_{M0}$: 0.79, $\boldsymbol{G}_{M3}$: 0.82, $\boldsymbol{G}_{M15}$: 0.79), indicating the presence of network hubs[21]. If we define hubs as nodes that connect with more than one-fifth of the total nodes in the network (Fig. S5), we find 24 hubs, 10 of which were in $\boldsymbol{G}_{M0}$ and 20 of which in $\boldsymbol{G}_{M15}$ but none in $\boldsymbol{G}_{M3}$. These results indicate that the overall structure of the gut microbiome may have undergone profound changes during the trial, particularly, high fiber intervention resulted in the loss of interactions between genome pairs.

We considered genomes having robust and stable ecological relationship if a genome pair keeps the same ecological interaction across all three timepoints. Out of the 113,526 possible genome pairs, 92.39% had no correlations at any of the three time points, suggesting that it may be a rare event for two genomes to establish an ecological relationship (Fig. 2B). Of the 477 prevalent HQMAGs, 184 had 517 positive correlations and 118 negative correlations at all three time points. Among these 184 HQMAGs, 43 were excluded from subsequent analysis because

they had no interactions with the remaining 141 nodes (Fig. S6). The remaining 141 HQMAGs, which included 586 genome pairs with stable correlations throughout the trial were further defined as genomes with stable ecological interactions (GSEIs) and became our microbiome signature candidates. We then explored how these 141 GSEIs were connected with each other and with the rest of the nodes in $\boldsymbol{G}_{M0}$, $\boldsymbol{G}_{M3}$, and $\boldsymbol{G}_{M15}$. (Fig. S7A). The 141 GSEIs had significantly higher degree, betweenness centrality, eigenvector centrality, closeness centrality and stress centrality than the rest of the genomes in the networks (Fig. S7B-F). This finding indicates that these GSEIs exerted a relatively large amount of control over the interaction of other nodes (reflected by betweenness centrality and eigenvector centrality) and the information flow in the network (reflected by closeness centrality and stress centrality). Removing these GSEIs would lead to the collapse of the networks since on average 86.08% of the total edges would have been lost. These suggest that the 141 GSEIs can be considered as the core nodes of the networks as they were highly connected not only within themselves but also with other nodes.

These 141 GSEIs were also highly prevalent among participants, as 140 of them were in > 90%, and 104 were in 100% of the 74 individuals in the W group (Fig. S8). In addition, these 141 GSEIs were also mostly predominant members of the gut microbiota as the abundance of 111 of them was higher than the median of the 1,845 HQMAGs and accounted for 20.78% of the total sequencing reads. Based on Bray-Curtis distance, beta-diversity analysis showed significant correlations between the profiles of the 141 GSEIs and all the 1,845 HQMAGs, as evidenced by Mantel test ($R^2 = 0.62$, $P = 0.001$) and Procrustes analysis ($P = 0.001$) (Fig. S9, Fig. 1C, D). These indicate that the variations of the 141 GSEIs contributed to the major variations of the whole gut microbial community across the 3 time points.

196 Bacteria which are positively correlated with each other and show robust co-occurrence
197 behavior can be recognized as ecological guilds[9]. The 141 GSEIs organized themselves into two
198 guilds and genomes in each guild were highly interconnected with positive correlations. Fifty
199 genomes were in Guild 1 and 91 genomes were in Guild 2 (Fig.2C, Fig. S10). All the genomes in
200 Guild 1 were from the phylum Firmicutes whereas those in Guild 2 were from 5 different phyla,
201 including Firmicutes, Bacteroidota, Proteobacteria, Actinobacteriota and Fusobacteriota. The two
202 guilds were connected by negative edges only, indicating a competitive relationship. Members of
203 Guild 1 increased its abundance from M0 to M3 and then decreased from M3 to M15 while
204 members of Guild 2 showed an opposite abundance change (Fig. 2C). Thus, members within
205 each guild had robust cooperative relationships, while competitive relationships existed between
206 the two guilds (Fig. 2D). Our data showed that the two guilds of the 141 GESIs formed a stable
207 seesaw-like network that existed in all three ecological networks $\boldsymbol{G}_{M0}$, $\boldsymbol{G}_{M3}$, and $\boldsymbol{G}_{M15}$ in the W
208 group. Furthermore, the finding of the seesaw-like network in the W group at M0 suggests that
209 the existence of such microbial organization is supposed to be irrelevant to the high fiber
210 intervention in our study. Given similar overall gut microbiota structure between the W and U
211 groups at M0 and in the U group across 3 timepoints (Fig. 1C, D), we speculated that the seesaw-
212 like network can be observed in the U group across the trial. Thus, we constructed the co-
213 abundance networks based on the abundance of the 141 GESIs across the individuals in the U
214 group at each time point. 99.8%, 99.51% and 99.74% of the total edges in the co-abundance
215 networks agreed with our seesaw-like network (Fig. S11A). This suggests that the detection of
216 these seesaw networked genomes was independent of the high fiber intervention, indicating that
217 the seesaw-like network may be an inherent structure of the gut microbiome in our study.
218

Functionality of the metagenomes of the two competing guilds modulates host metabolic phenotypes

We sought to determine whether the balance between the two competing guilds could be modulated by dietary fiber and describe how the two competing guilds affects the host metabolic phenotypes. In the W group, the total abundance of Guild 1 increased and Guild 2 decreased significantly from M0 to M3. Then at M15, Guild 1 decreased to a level similar to that at M0, and Guild 2 bounced back but remained lower than that at M0. Subsequently, from M0 to M3, high fiber intervention significantly increased the Guild 1 to Guild 2 ratio. At one-year follow-up, the ratio significantly decreased and was not different from M0 (Fig. 3A). Neither the abundances of the 2 guilds nor their ratio was changed in the U group across the trial (Fig. S11B). These results showed that the changes of the balance between the two guilds composed of GSEIs were concomitant with the change patterns of dietary intake, overall gut microbiota and host phenotypes. To further validate our hypothesis that GSEIs may be essential to host health, we used the GSEIs as the selected features and applied machine leaning algorithms to explore the associations between GSEIs and each host bio-clinical parameter. Random Forest regression via leave-one-out cross-validation based on the 141 GSEIs showed 41 out of the 43 bio-clinical parameters with significant Pearson's correlation coefficient ranged from 0.11 to 0.44 (adjusted P value < 0.05) between the predicted and measured values (Fig. 3B). These results showed that the 141 genomes, as two competing guilds in a seesaw-like network, constitute an important microbiome signature for T2DM and the related metabolic phenotypes.

Next, we performed genome-centric analysis of the metagenomes of the two competing guilds to explore the genetic basis underlying the association between the dynamic changes of the seesaw networked microbiome signature and the response of the host's metabolic

phenotypes. As the balance between the two guilds can be shifted by dietary fibers, we first sought to identify carbohydrate-active enzyme (CAZy)-encoding genes and genes encoding key enzymes in short-chain fatty acids (SCFAs) production to compare the genetic capacity for carbohydrate utilization between the two guilds. Compared with genomes in Guild 2, those in Guild 1 enriched CAZy genes for arabinoxylan ($P < 0.001$), cellulose ($P < 0.01$) and had lower proportion of CAZy genes for inulin utilization ($P < 0.01$) (Fig. 3C, Table S5). There was no difference in genes for starch, pectin, and mucin utilization between the two guilds. Our previous study showed that gut microbiota benefited patients with T2DM via acetic and butyric acid production from carbohydrate fermentation[11]. Among the terminal genes for the butyrate biosynthetic pathways from both carbohydrates (i.e., *but* and *buk*) and proteins (i.e., *atoA*/*D* and *4Hbt*), the copy number of *but* was significantly higher in Guild 1 and there was no difference in the other terminal genes between the two guilds (Fig. 3C). More than one-third of the genomes in Guild 1 harbored the *but* gene while less then 5% of the genomes in Guild 2 had this gene (Fisher's exact test $P < 0.001$). Compared with Guild 2, Guild 1 also trended higher in its genetic capacity for acetate production ($P = 0.06$) but a lower genetic capacity for propionate production ($P < 0.05$) (Fig. 3C). These results showed that compared to Guild 2, Guild 1 had significantly higher genetic capacity for utilizing complex plant polysaccharides and producing acetate and butyrate.

From the perspective of pathogenicity, 21 out of the 1,845 HQMAGs encoded 750 virulence factor (VF) genes. Among the 21 VF-encoding genomes, 3 were in Guild 1 while 18 were in Guild 2. Three out of the 50 genomes in Guild 1 had one VF gene involved in antiphagocytosis. In Guild 2, 18 out of the 91 genomes encoded 747 VF genes across 15 different VF classes i.e., acid resistance, adherence, antiphagocytosis, biofilm formation, efflux

pump, endotoxin, invasion, iron uptake, manganese uptake, motility, nutritional factor, protease, regulation, secretion system, and toxin (Fig. 3C, S12A). Notably, 98.53% of all the VF genes in Guild 2 were harbored in 8 genomes (1 in *Enterobacter kobei*, 2 in *Escherichia flexneri,* 3 in *Escherichia coli* and 2 in *Klebsiella*). The highly enriched genes for virulence factors in genomes of Guild 2 ($P < 2.2\times10^{-16}$, Fisher's Exact test) indicates that this guild may play an important role in aggravating the metabolic disease phenotypes. In terms of antibiotic resistance genes (ARG), in Guild 1, only 1 genome (2.00% of the genomes in this guild) harbored a copy of an ARG related to phenicol (Fig. 3C, S12B). In Guild 2, 17 genomes (18.68% of the genomes in this guild) encode 40 ARGs for resistance to 7 different antibiotic classes i.e., aminoglycosides, beta-lactam, fosfomycin, glycopeptide, quinolone, macrolide, and tetracycline. Thus, Guild 2 may serve as a reservoir of ARGs for horizontal transfer to opportunistic pathogens. Taken together, our data showed that the two competing guilds had distinct genetic capacity with Guild 1 being potentially beneficial and Guild 2 detrimental[11].

The seesaw networked microbiome signature exists in cohorts across ethnicity and geography

We then asked that whether these 141 genomes, organized as two competing guilds in a stable seesaw-like network, may be a common microbiome signature for different diseases in other independent metagenomically studied cohorts. To answer this question, we used these 141 GSEIs in our seesaw-like network as reference genomes to perform read recruitment analysis, which is a commonly used method to estimate abundance of reference genomes[22,23] in metagenomes (Fig. S13). In an independent T2DM study[24], 32.92% of the reads were recruited and 128 of the GSEIs were detected as part of a co-abundance network based on their estimated

abundance across the T2DM patients. In this co-abundance network, 97.82% of the total edges followed the pattern in our seesaw-like network (i.e., positive edges within each guild and negative edges between the 2 guilds) (Fig. 4A), which further supported the existence of this seesaw-like network in T2DM patients. Moreover, 35.28% of the reads were recruited in the metagenomes of 136 healthy controls of the same study24, 119 of the GSEIs were constructed into a co-abundance network in which 99.45 % of the total edges agreed with our seesaw-like network (Fig. 4A). In the context of beta diversity based on Bray-Curtis distance, our microbiome signature showed significant differences (PERMAONVA test $P = 2\times10^{-4}$) between T2DM patients and the healthy controls based on the abundance matrix of the reference genomes (Fig. 4B). This suggests that the variation of this microbiome signature was associated with T2DM in this independent dataset. To further validate such associations, using the abundance matrix of the genomes in the microbiome signature as input features and the phenotype data, we constructed Random Forest regression models and found that this microbiome signature was significantly correlated with BMI, fasting insulin, and HbA1c (Fig S14). Furthermore, we developed a machine learning classifier based on a Random Forest algorithm to see if we can classify patients and control. Receiver operating characteristic curve analysis showed a moderate diagnostic power with area under the curve (AUC) of 0.70 by a leave-one-out cross-validation. Thus, we showed that our seesaw networked microbiome signature not only existed in an independent T2DM study but also maintained a similar relationship with the host metabolic phenotypes.

We then extended our hypothesis that the seesaw networked microbiome signature represents an inherent feature of human gut microbiome and the disruption of which may be related to diseases in addition to T2DM. We first performed the same validation analysis in

311 metagenomic datasets of three different types of diseases, including ACVD[25](a chronic
312 metabolic disease), LC[26] (a liver disease) and AS[27](an autoimmune disease). In ACVD patients
313 and their controls, 36.21% and 32.73% of the reads were recruited, and 134 genomes from the
314 patients and 133 genomes from the controls were constructed into co-abundance networks with
315 97.32% and 97.70% of the total edges respectively agreed with our seesaw-like network (Fig.
316 4A). 33.84%, 35.83% and 41.02% of the reads were recruited to the reference genomes in the
317 metagenomic datasets of the healthy control (the studies on LC and AS employed the same
318 control cohort), LC and AS patients respectively. 112, 113 and 113 reference genomes were
319 constructed into co-abundance networks with 99.80%, 98.81% and 98.19% of the total edges
320 agreed with our seesaw-like network in the metagenomic datasets of the healthy control, LC and
321 AS patients respectively (Fig. 4A). In the PCoA plot based on Bray-Curtis distance, our
322 microbiome signature showed significant differences (PERMANOVA test, $P < 0.001$) between
323 control and patients in all 3 datasets (Fig. 4B). In the LC study, we also used the abundance
324 matrix of the genomes in the microbiome signature as input features and the phenotype data to
325 construct Random Forest regression models and found that our microbiome signature was
326 significantly correlated with total bilirubin, albumin level, and BMI (Fig. S15). Compared with
327 the T2DM dataset[24], the Random Forest classifier based on our microbiome signature showed
328 better diagnostic power in distinguishing case from control for ACVD (AUC = 0.80), LC (AUC
329 = 0.90), and AS (AUC = 0.98) (Fig. 4C).
330 To further confirm the relevance of this microbiome signature to human diseases, we
331 estimated the abundances of the genomes from this microbiome signature in datasets from more
332 disease types and across different ethnicity and geography. These datasets included IBD
333 (American cohort and Dutch cohort), CRC (Chinese cohort, Australian cohort and German

cohort), schizophrenia (Chinese cohort), and PD (Chinese cohort). On average, 31.32% ± 4.21% (mean ± s.d.) of reads were recruited to the reference genomes in these datasets. We validated that this microbiome signature showed diagnostic power to classify case and control in the metagenomic dataset from studies on IBD (AUC = 0.71 for IBD dataset 1[28], AUC=0.91 for IBD dataset 2[29] and AUC=0.83 for IBD dataset 3[29]), CRC (AUC = 0.74 for CRC dataset 1[30], AUC = 0.75 for CRC dataset 2[31] and AUC = 0.71 for CRC dataset 3[32]), schizophrenia[33] (AUC = 0.68), and PD[34] (AUC = 0.77) (Fig. S16). In addition, we used MMUPHin[35] to correct batch effects from the different cohorts in IBD and CRC and applied leave-one-cohort-out (LOCO) analysis[36] to evaluate the universality of the diagnostic power of this microbiome signature in these two diseases. The AUC values from LOCO analysis were 0.77 to 0.84 for IBD and 0.68 to 0.70 for CRC (Fig. S17). These results showed the existence of our microbiome signature in healthy controls and various patient populations across ethnicity and geography from independent studies. The associations between the 141 genomes and host phenotypes and their discriminative power as biomarkers to classify controls vs. patients with various types of diseases indicate that these genomes, organized as two guilds in a seesaw-like network, represent a common microbiome signature associated with widely different human disease phenotypes.

Discussion

In the current study, our genome-centric, reference-free, and ecological-interaction-focused approach led to the identification of a robust seesaw-like network of two competing guilds of bacterial genomes, whose changes were associated with a wide range of host phenotypes in patients with T2DM. Moreover, random forest models based on these genomes classified case and control across a wide range of diseases, indicating that these genomes may form a novel

357 microbiome signature that exists in populations of widely different ethnicity, geography, and
358 disease status.

359 Our novel microbiome signature organizes genomes in a seesaw-like network exhibiting
360 both cooperative and competitive interactions. Though cooperative ecological networks are
361 expected to promote overall metabolic efficiency, such as the co-operative metabolism that
362 benefits the host[37], it creates dependency and the potential for mutual downfall that may bring
363 destabilizing effect on the gut microbial ecosystem. This destabilizing effect of cooperation can
364 be dampened by introducing ecological competition into the network[37]. Thus, a seesaw-like
365 network with both cooperative and competitive interactions may represent the key characteristic
366 of a stable microbiome structure[37]. Interestingly, while the seesaw-like network is stable, the
367 weight of the two ends i.e., the abundances of Guild 1 and Guild 2, are modifiable and such
368 changes are associated with host health. When large amount of complex fiber became available,
369 Guilds 1 and 2 showed no change in membership nor in the types of interactions with each other
370 but experienced dramatic shifts in guild-level abundance in a competing manner. Members in
371 Guild 1 have higher genetic capacity for degrading complex plant polysaccharides and produce
372 beneficial metabolites including SCFAs which may suppress populations of pathobionts in Guild
373 2[11]. Members of Guild 2 need to be kept low since their overgrowth may jeopardize host health
374 by increasing inflammation, etc.[38]. However, pathobionts in Guild 2 cannot be eliminated, e.g.,
375 they could serve as the necessary agents that train our immune system from early days[39,40].
376 Therefore, the balance between Guild 1 and Guild 2 becomes critical in determining whether the
377 gut microbiome supports health or aggravate diseases. This seesaw-like network between Guilds
378 1 and 2 allows the genomes in our microbiome signature to readily respond to changes of
379 external energy input to the gut microbial ecosystem and mediate its impact on host health, while

simultaneously maintains its structural integrity. Such structural integrity may be key to long-term ecological stability of the gut microbiome and its ability to provide essential health-relevant functions to the host.

Such a seesaw networked structure may have been stabilized by natural selection over a long history of co-evolution between microbiomes and their hosts[16 41]. A selection pressure may have been exerted by dietary fibers that interact directly with gut bacteria as external energy source[42,43]. Studies on coprolites showed that dietary fiber intake was much higher in ancient humans and only reduced significantly in the past 150 years[44,45] (130 g/d of plant fiber intake in prehistoric diet[46] vs. a median intake of 12–14 g/d in the modern American diet[47]). Such a high fiber intake over evolutionary history may have favored beneficial bacteria in Guild 1 due to their higher genetic capacity to utilize plant polysaccharides as an external energy supply, enabling them to gain competitive advantage over pathobionts in Guild 2[48]. Akin to tall trees as the foundation species for a closed forest, Guild 1 may work as the "foundation guild" for stabilizing a healthy gut microbiome and keeping the pathobionts at bay[49]. The dominance of Guild 1 over Guild 2 can increase host fitness as shown by the epidemiologically and clinically proven health benefits of dietary fibers in both preventing and alleviating a wide range of chronic conditions[11,43,50,51].

Moreover, the seesaw networked microbiome signature may be considered as part of the core gut microbiome in humans[52,53], since 1) they are commonly shared among populations across ethnicity and geography; 2) they show temporal stability not only in membership but also in their interactions with each other and the host; 3) they make up about 10% of the gut microbiome membership but are disproportionally important for shaping the ecological community; 4) they support the provision of essential health-relevant functions to the host; and

5) such a core microbiome organized in a seesaw-like network may have been established over a long history of co-evolution.

The fact that this seesaw-like network can be detected in other independent metagenomic datasets and is shown correlated with different diseases indicates that this core microbiome signature could be an evolutionarily conserved ecological structure and may be fundamentally important to human health recovery and maintenance. In addition, our seesaw-like network demonstrated stable relationships both internally within the network and externally with multiple host clinical markers, suggesting that genome-based bacterial guilds may serve as robust disease biomarkers. Within the seesaw-like network, it is the imbalance between the two competing guilds that may play a role as the common biological basis for many human diseases. Targeting this core microbiome signature to restore and maintain dominance of the beneficial guild over the detrimental guild could help reduce disease risk or alleviate symptoms, thus opening a new avenue for chronic diseases management and prevention.

Materials and Methods

Clinical Experiment

Study design[11]: This clinical trial, conducted at the Qidong People's Hospital (Jiangsu, China), examined the effect of a high fiber diet in free-living conditions in a cohort of individuals clinically diagnosed T2DM (QIDONG). The study protocol was approved by Ethics Committee of Shanghai General Hospital (2014KY104), and the study was conducted in accordance with the principles of the Declaration of Helsinki. All participants provided written informed consent. The trial was registered in the Chinese Clinical Trial Registry (ChiCTR-IPC-14005346). The study design and participant flow are shown in Fig. S1.

T2DM patients of the Chinese Han ethnicity were recruited for the study (age: 37 - 70 years; HbA1c: 6.5% - 12.0%. More detailed description of inclusion and exclusion criteria were shown in Chinese Clinical Trial registry (http://www.chictr.org.cn).

Patients received either a high-fiber diet (WTP diet) as the treatment group (W group) or the usual care (Usual diet) as the control group (U group) for 3 months. Total caloric and macronutrients prescriptions were based on age-specific Chinese Dietary Reference Intakes (Chinese Nutrition Society, 2013). The WTP diet, based on wholegrains, traditional Chinese medicinal foods and prebiotics, included three ready-to-consume pre-prepared foods[11]. The usual care included standard dietary and exercise advice that was made according to the Chinese Diabetes Society guidelines for T2DM[54]. Patients in W group were provided with the WTP diet to perform a self-administered intervention at home for three months, while patients in U group accepted the usual care. W group stopped WTP diet intervention at the end of the third month (at M3). Then W and U continued a one-year follow-up (M15). A meal-based food frequency questionnaire and 24-h dietary recall were used to calculate nutrient intake based on the China Food Composition 2009[55]. Patients in both groups continued with their antidiabetic medications according to their physician prescriptions (Table S1).

Before a 2-week run-in period, all participants attended a lecture on diabetes intervention and improvements and received diabetes education and metabolic assessments. 119 eligible individuals were enrolled based on the inclusion and exclusion criteria and assigned into two groups in a 2:1 ratio (n = 79 in W group, n = 40 in U group) determined by SAS software.

Physical examinations were carried out at M0, M3, and M15 in Qidong People's Hospital (Jiangsu, China). Sample collection instructions were provided to the participants at the day before. The participants provided the feces and first early morning urine as requested. After

collecting fasting venous blood sample, a 3-h meal tolerance test (Chinese buns containing 75 g of available carbohydrates; MTT test) was conducted and the postprandial venous blood samples at 30, 60, 120, and 180 min were collected. All the blood samples were centrifuged at 3000 rpm for 20 min at 4℃ after standing at room temperature for 30 min to obtain serum. The fasting blood serum were divided into two parts, one used for hospital tests and the other used for lab tests. The feces, urine, and serum samples were stored in dry ice immediately then transported to lab and frozen at -80℃. Subsequently, anthropometric markers and diabetic complication indexes were measured. Ewing test[56] and 24-h dynamic electrocardiogram were conducted to estimate diabetic autonomic neuropathy (DAN). B-mode carotid ultrasound was conducted to estimate atherosclerosis. Michigan Neuropathy Screening Instrument[57] was conducted to estimate diabetic peripheral neuropathy (DPN). In addition, A meal-based food frequency questionnaire and the 24-h dietary review were recorded for nutrient intake calculation. The drug use was self-reported and presented in table S1.

The fasting venous blood was used to measure HbA1c, fasting blood glucose, fasting insulin, fasting C-Peptide, C-reactive protein (CRP), blood routine examination, blood biochemical examination and five analytes of thyroid. The venous blood samples at 30, 60, 120, and 180 min of MTT were used to measure the postprandial blood glucose, insulin, and C-Peptide. The fasting early morning urine was used to measure the routine urine examination and urinary microalbumin creatinine ratio. The measurements above were completed at Qidong People's Hospital. Fasting venous blood was used to quantify TNF-α (R&D Systems, MN, USA), lipopolysaccharide-binding protein (Hycult Biotech, PA, USA), leptin (P&C, PCDBH0287, China) and adiponectin (P&C, PCDBH0016, China) by enzyme-linked immunosorbent assays (ELISAs) at Shanghai Jiao Tong University.

The homeostatic model assessments of insulin resistance (HOMA-IR) and islet β-cell function (HOMA-β) were calculated based on fasting blood glucose (mmol/L) and fasting C-Peptide (pmol/L)[58]: HOMA-IR = 1.5 + FBG * Fasting-C-Peptide / 2800;

HOMA-β = 0.27 * Fasting-C-Peptide / (FBG - 3.5). Glomerular Filtration Rate was estimated by formula GFR (ml/min per 1.73 m^2) = 186 * $Scr^{-1.154}$ * $age^{-0.203}$ * 0.742 (if female) * 1.233 (if Chinese)[59], where Scr (serum creatinine) is in mg/dl and age is in years.

Gut microbiome analysis

Metagenomic sequencing. DNA was extracted from fecal samples using the methods as previously described[10]. Metagenomic sequencing was performed using Illumina Hiseq 3000 at GENEWIZ Co. (Beijing, China). Cluster generation, template hybridization, isothermal amplification, linearization, and blocking denaturing and hybridization of the sequencing primers were performed according to the workflow specified by the service provider. Libraries were constructed with an insert size of approximately 500 bp followed by high-throughput sequencing to obtain paired-end reads with 150 bp in the forward and reverse directions. Table S3 shows the number of raw reads of each sample.

Data quality control. Prinseq[60] was used to: 1) trim the reads from the 3′ end until reaching the first nucleotide with a quality threshold of 20; 2) remove read pairs when either read was < 60 bp or contained "N" bases; and 3) de-duplicate the reads. Reads that could be aligned to the human genome (H. sapiens, UCSC hg19) were removed (aligned with Bowtie2[61] using --reorder --no-hd --no-contain --dovetail). Table S3 shows the number of high-quality reads of each sample for further analysis.

De novo assembly, abundance calculation, and taxonomic assignment of genomes. De novo assembly was performed for each sample by using IDBA_UD[62] (--step 20 --mink 20 --maxk 100 --min_contig 500 --pre_correction). The assembled contigs were further binned using MetaBAT[63] (--minContig 1500 --superspecific -B 20). The quality of the bins was assessed using CheckM[64]. Bins had completeness > 95%, contamination < 5% and strain heterogeneity < 5% were retained as high-quality draft genomes (Table S6). The assembled high-quality draft genomes were further dereplicated by using dRep[65]. DiTASiC[66], which applied kallisto for pseudo-alignment[67] and a generalized linear model for resolving shared reads among genomes, was used to calculate the abundance of the genomes in each sample, estimated counts with P-value > 0.05 were removed, and all samples were downsized to 36 million reads (One sample with read mapping ratio < 25%, which could not be well represented by the high quality genomes, were removed in downstream analysis). Taxonomic assignment of the genomes was performed by using GTDB-Tk[68] (Table S7).

Gut microbiome functional analysis. Prokka[69] was used to annotate the genomes. KEGG Orthologue (KO) IDs were assigned to the predicted protein sequences in each genome by HMMSEARCH against KOfam using KofamKOALA[70]. Antibiotic resistance genes were predicted using ResFinder[71] with default parameters. The identification of virulence factors were based on the core set of Virulence Factors of Pathogenic Bacteria Database (VFDB[72], download July 2020). The predicted proteins sequences were aligned to the reference sequence in VFDB using BLASTP (best hist with E-value < 1e-5, identity > 80% and query coverage > 70%). Genes encoding carbohydrate-active enzymes (CAZys) were identified using dbCAN (releasee

6.0)[73], and the best-hit alignment was retained. Genes encoding formate-tetrahydrofolate ligase, propionyl-CoA: succinate-CoA transferase, propionate CoA-transferase, 4Hbt, AtoA, AtoD, Buk and But were identified as described previously[11].

Gut microbiome network construction and analysis. In W group, prevalent genomes shared by more than 75% of the samples at every timepoint were used to construct the co-abundance network at each timepoint. Fastspar[75], a rapid and scalable correlation estimation tool for microbiome study, was used to calculate the correlations between the genomes with 1,000 permutations at each time point based on the abundances of the genomes across the patients and the correlations with $P \leqslant 0.001$ were retained for further analysis. The networks were visualized with Cystoscape v3.8.1[76]. The layout of the nodes and edges was determined by Edge-weighted Spring Embedded Layout using the correlation coefficient as weights. The links between the nodes are treated as metal springs attached to the pair of nodes. The correlation coefficient was used to determine the repulsion and attraction of the spring[76]. The layout algorithm sets the position of the nodes to minimize the sum of forces in the network. We defined robust stable edges as the unchanged positive/negative correlations between the same two genomes across all the 3 networks at M0, M3, and M15. Stable genome pairs were clustered based on robust positive (set as 1) and negative (set as -1) edges with average clustering. We used iTOL[77], an online tool for display, manipulation, and annotation for various trees, to integrate and visualize the clustering tree, taxonomy information, and abundance changes of the 141 genomes.

Validation in independent cohorts. Twelve independent metagenomic datasets were downloaded from SRA or ENA database. The group information was collected from the

corresponding papers or from curatedMetagenomicData[78] (Table S8). DiTASiC was used to recruit reads and estimate the abundance of the 141 genomes in each sample, estimated counts with P-value > 0.05 were removed and further converted to relative abundance divided by the total number of reads. To reduce false positive in the validation dataset, relative abundance < 0.001% were further removed. A random forest classification model to classify case and control was constructed based on the estimated abundances of the genomes in each dataset with leave-one-out cross-validation.

MMUPHin[35] was used to adjust the estimated abundances of the genomes by correcting batched effects from the different cohorts in IBD and CRC studies. Random forest classification models with leave-one-cohort-out analysis were further performed on the adjusted abundance matrix[36].

Datasets from 4 studies were included to validate the commonality of the seesaw-like network. These datasets were from 136 control and 136 T2DM individuals in Qin et al., 2012[24]; 171 control and 214 atherosclerotic cardiovascular disease individuals in Jie et al., 2017[25]; 83 control and 84 liver cirrhosis individuals in Qin et al., 2014[26]; and 83 control and 97 ankylosing spondylitis individuals in Wen et al., 2017[27]. Fastspar was used to calculate the correlations between the genomes with 1,000 permutations and the correlations with $P \leqslant 0.001$ were remained for constructing the networks. 30 repeat 5-fold cross-validation was used and the correlations shared by more than 95% of the 150 networks constructed from the cross-validation process were remained in the final network.

Statistical Analysis

Statistical analysis was performed in the R environment (R version3.6.1). Friedman test followed by Nemenyi post-hoc test was used for intra-group comparisons. Mann-Whitney test (two-sided) was used for comparisons between W and U at the same time point. Pearson Chi-square tests was performed to compare the differences of categorical data between groups or timepoints. PERMANOVA test (9,999 permutations) was used to compare the groups of gut microbiota structure. P value less than 0.05 was accepted as statistical significance.

Mann-Whitney test (two-sided) and Fisher's exact test (two sided) were used to compare the functions between Guild 1 and Guild 2. Random Forest with leave-one-out cross-validation was used to perform regression and classification analysis based on this microbiome signature and clinical parameters/groups.

Acknowledgements

This work was supported by grants from National Natural Science Foundation of China (31930022, 81871091, 81870594 and 81870596), the National Key Research and Development Project (2019YFA0905600), Clinical Research Plan of SHDC [No. SHDC2020CR1016B], the Project of Songjiang District Municipal Health Commission (0702N18003), the School of Environmental and Biological Sciences and the New Jersey Institute for Food, Nutrition, and Health (seed pilot grant cycle 1, awarded in 2019), Canadian Institute for Advanced Research and Notitia Biotechnologies Company.

Conflict of interest

Liping Zhao is a co-founder of Notitia Biotechnologies Company.

Reference

1 Dominguez-Bello, M. G., Godoy-Vitorino, F., Knight, R. & Blaser, M. J. Role of the microbiome in human development. *Gut* **68**, 1108-1114, doi:10.1136/gutjnl-2018-317503 (2019).

2 Kundu, P., Blacher, E., Elinav, E. & Pettersson, S. Our Gut Microbiome: The Evolving Inner Self. *Cell* **171**, 1481-1493, doi:10.1016/j.cell.2017.11.024 (2017).

3 O'Hara, A. M. & Shanahan, F. The gut flora as a forgotten organ. *Embo Rep* **7**, 688-693, doi:10.1038/sj.embor.7400731 (2006).

4 Koh, A. & Backhed, F. From Association to Causality: the Role of the Gut Microbiota and Its Functional Products on Host Metabolism. *Mol Cell* **78**, 584-596, doi:10.1016/j.molcel.2020.03.005 (2020).

5 Sanna, S. *et al.* Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nat Genet* **51**, 600-+, doi:10.1038/s41588-019-0350-x (2019).

6 Meijnikman, A. S., Gerdes, V. E., Nieuwdorp, M. & Herrema, H. Evaluating Causality of Gut Microbiota in Obesity and Diabetes in Humans. *Endocr Rev* **39**, 133-153, doi:10.1210/er.2017-00192 (2018).

7 Levin, S. A. Ecosystems and the biosphere as complex adaptive systems. *Ecosystems* **1**, 431-436 (1998).

8 Zhang, C. & Zhao, L. Strain-level dissection of the contribution of the gut microbiome to human metabolic disease. *Genome Med* **8**, 41, doi:10.1186/s13073-016-0304-1 (2016).

9 Wu, G., Zhao, N., Zhang, C., Lam, Y. Y. & Zhao, L. Guild-based analysis for understanding gut microbiome in human health and diseases. *Genome Medicine* **13**, 22, doi:10.1186/s13073-021-00840-y (2021).

10 Zhang, C. H. *et al.* Dietary Modulation of Gut Microbiota Contributes to Alleviation of Both Genetic and Simple Obesity in Children. *Ebiomedicine* **2**, 968-984, doi:10.1016/j.ebiom.2015.07.007 (2015).

11 Zhao, L. *et al.* Gut bacteria selectively promoted by dietary fibers alleviate type 2 diabetes. *Science* **359**, 1151-1156 (2018).

12 Tierney, B. T., Tan, Y., Kostic, A. D. & Patel, C. J. Gene-level metagenomic architectures across diseases yield high-resolution microbiome diagnostic indicators. *Nat Commun* **12**, 2907, doi:10.1038/s41467-021-23029-8 (2021).

13 Wang, J. & Jia, H. Metagenome-wide association studies: fine-mining the microbiome. *Nat Rev Microbiol* **14**, 508-522, doi:10.1038/nrmicro.2016.83 (2016).

14 Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A. & Alm, E. J. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat Commun* **8**, 1784, doi:10.1038/s41467-017-01973-8 (2017).

15 Jackson, M. A. *et al.* Gut microbiota associations with common diseases and prescription medications in a population-based cohort. *Nat Commun* **9**, 2655, doi:10.1038/s41467-018-05184-7 (2018).

16 Foster, K. R., Chluter, J. S., Oyte, K. Z. C. & Rakoff-Nahoum, S. The evolution of the host microbiome as an ecosystem on a leash. *Nature* **548**, 43-51, doi:10.1038/nature23292 (2017).

17 Sommer, F., Anderson, J. M., Bharti, R., Raes, J. & Rosenstiel, P. The resilience of the intestinal microbiota influences health and disease. *Nat Rev Microbiol* **15**, 630-638, doi:10.1038/nrmicro.2017.58 (2017).

18 Ma, B. *et al.* Earth microbial co-occurrence network reveals interconnection pattern across microbiomes. *Microbiome* **8**, 82, doi:10.1186/s40168-020-00857-2 (2020).

19 Bauer, E. & Thiele, I. From Network Analysis to Functional Metabolic Modeling of the Human Gut Microbiota. *mSystems* **3**, doi:10.1128/mSystems.00209-17 (2018).

20 Poisot, T. & Gravel, D. When is an ecological network complex? Connectance drives degree distribution and emerging network properties. *PeerJ* **2**, e251, doi:10.7717/peerj.251 (2014).

21 Barabasi, A. L. Network science. *Philos Trans A Math Phys Eng Sci* **371**, 20120375, doi:10.1098/rsta.2012.0375 (2013).

22 Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nat Biotechnol* **39**, 499-509, doi:10.1038/s41587-020-0718-6 (2021).

23 Reji, L., Cardarelli, E. L., Boye, K., Bargar, J. R. & Francis, C. A. Diverse ecophysiological adaptations of subsurface Thaumarchaeota in floodplain sediments revealed through genome-resolved metagenomics. *ISME J*, doi:10.1038/s41396-021-01167-7 (2021).

24 Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55-60, doi:10.1038/nature11450 (2012).

25 Jie, Z. *et al.* The gut microbiome in atherosclerotic cardiovascular disease. *Nat Commun* **8**, 845, doi:10.1038/s41467-017-00900-1 (2017).

26 Qin, N. *et al.* Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513**, 59-64, doi:10.1038/nature13568 (2014).

27 Wen, C. *et al.* Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. *Genome Biol* **18**, 142, doi:10.1186/s13059-017-1271-6 (2017).

28 Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655-662, doi:10.1038/s41586-019-1237-9 (2019).

29 Franzosa, E. A. *et al.* Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat Microbiol* **4**, 293-305, doi:10.1038/s41564-018-0306-4 (2019).

30 Yu, J. *et al.* Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* **66**, 70-+, doi:DOI 10.1136/gutjnl-2015-309800 (2017).

31 Feng, Q. *et al.* Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat Commun* **6**, 6528, doi:10.1038/ncomms7528 (2015).

32 Wirbel, J. *et al.* Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat Med* **25**, 679-689, doi:10.1038/s41591-019-0406-6 (2019).

33 Zhu, F. *et al.* Metagenome-wide association of gut microbiome features for schizophrenia. *Nat Commun* **11**, 1612, doi:10.1038/s41467-020-15457-9 (2020).

34 Qian, Y. W. *et al.* Gut metagenomics-derived genes as potential biomarkers of Parkinson's disease. *Brain* **143**, 2474-2489 (2020).

35 Ma, S. *et al.* Population Structure Discovery in Meta-Analyzed Microbial Communities and Inflammatory Bowel Disease. *bioRxiv*, 2020.2008.2031.261214, doi:10.1101/2020.08.31.261214 (2020).

36 Wu, Y. *et al.* Identification of microbial markers across populations in early detection of colorectal cancer. *Nat Commun* **12**, 3063, doi:10.1038/s41467-021-23265-y (2021).

37 Coyte, K. Z., Schluter, J. & Foster, K. R. The ecology of the microbiome: Networks, competition, and stability. *Science* **350**, 663-666, doi:10.1126/science.aad2602 (2015).

38 Kamada, N., Seo, S. U., Chen, G. Y. & Nunez, G. Role of the gut microbiota in immunity and inflammatory disease. *Nat Rev Immunol* **13**, 321-335, doi:10.1038/nri3430 (2013).

39 Vatanen, T. *et al.* Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans. *Cell* **165**, 1551, doi:10.1016/j.cell.2016.05.056 (2016).

40 Bach, J. F. The hygiene hypothesis in autoimmunity: the role of pathogens and commensals. *Nat Rev Immunol* **18**, 105-120, doi:10.1038/nri.2017.111 (2018).

41 Risely, A. Applying the core microbiome to understand host–microbe systems. *Journal of Animal Ecology* **89**, 1549-1558 (2020).

42 Makki, K., Deehan, E. C., Walter, J. & Backhed, F. The Impact of Dietary Fiber on Gut Microbiota in Host Health and Disease. *Cell Host Microbe* **23**, 705-715, doi:10.1016/j.chom.2018.05.012 (2018).

43 Reynolds, A. *et al.* Carbohydrate quality and human health: a series of systematic reviews and meta-analyses. *The Lancet* **393**, 434-445 (2019).

44 Eaton, S. B. The ancestral human diet: what was it and should it be a paradigm for contemporary nutrition ? *P Nutr Soc* **65**, 1-6, doi:10.1079/Pns2005471 (2006).

45 Jew, S., AbuMweis, S. S. & Jones, P. J. Evolution of the human diet: linking our ancestral diet to modern functional foods as a means of chronic disease prevention. *J Med Food* **12**, 925-934, doi:10.1089/jmf.2008.0268 (2009).

46 Spiller, G. A. & Amen, R. J. *Topics in dietary fiber research*. (Springer, 1978).

47 Thompson, H. J. & Brick, M. A. Perspective: Closing the dietary fiber gap: An ancient solution for a 21st century problem. *Advances in Nutrition* **7**, 623-626 (2016).

48 Deehan, E. C. *et al.* Modulation of the gastrointestinal microbiome with nondigestible fermentable carbohydrates to improve human health. *Microbiology spectrum* **5**, 5.5. 04 (2017).

49 Prevey, J. S., Germino, M. J. & Huntly, N. J. Loss of foundation species increases population growth of exotic forbs in sagebrush steppe. *Ecol Appl* **20**, 1890-1902, doi:10.1890/09-0750.1 (2010).

50 Anderson, J. W. *et al.* Health benefits of dietary fiber. *Nutr Rev* **67**, 188-205, doi:10.1111/j.1753-4887.2009.00189.x (2009).

51 Kaczmarczyk, M. M., Miller, M. J. & Freund, G. G. The health benefits of dietary fiber: beyond the usual suspects of type 2 diabetes mellitus, cardiovascular disease and colon cancer. *Metabolism* **61**, 1058-1066, doi:10.1016/j.metabol.2012.01.017 (2012).

52 Risely, A. Applying the core microbiome to understand host-microbe systems. *J Anim Ecol* **89**, 1549-1558, doi:10.1111/1365-2656.13229 (2020).

53 Berg, G. *et al.* Microbiome definition re-visited: old concepts and new challenges. *Microbiome* **8**, 103, doi:10.1186/s40168-020-00875-0 (2020).

54 Society, C. D. China guideline for type 2 diabetes (2013 edition). *Chin J Diabetes* **22** (2014).

55 yuexin, Y., guangya, W. & xingchang, P. China Food Composition (Book 1, Beijing Medical Univ. Press, ed. 2, 2009). (2009).

56 Ewing, D. & Clarke, B. Diagnosis and management of diabetic autonomic. *British Medical Journal* **285** (1982).

57 Feldman, E. L. *et al.* A Practical Two-Step Quantitative Clinical and Electrophysiological Assessment for the Diagnosis and Staging of Diabetic Neuropathy. *Diabetes Care* **17**, 1281-1289 (1994).

58 Li, X., Zhou, Z. G., Qi, H. Y., Chen, X. Y. & Huang, G. Replacement of insulin by fasting C-peptide in modified homeostasis model assessment to evaluate insulin resistance and islet beta cell function. *Zhong Nan Da Xue Xue Bao Yi Xue Ban* **29**, 419-423 (2004).

59 Ma, Y. C. *et al.* Modified glomerular filtration rate estimating equation for Chinese patients with chronic kidney disease. *J Am Soc Nephrol* **17**, 2937-2944, doi:10.1681/ASN.2006040368 (2006).

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

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

62 Peng, Y., Leung, H. C., Yiu, S. M. & Chin, F. Y. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420-1428, doi:10.1093/bioinformatics/bts174 (2012).

63 Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165, doi:10.7717/peerj.1165 (2015).

64 Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**, 1043-1055, doi:10.1101/gr.186072.114 (2015).

65 Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J* **11**, 2864-2868, doi:10.1038/ismej.2017.126 (2017).

66 Fischer, M., Strauch, B. & Renard, B. Y. Abundance estimation and differential testing on strain level in metagenomics data. *Bioinformatics* **33**, i124-i132, doi:10.1093/bioinformatics/btx237 (2017).

67 Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525-527, doi:10.1038/nbt.3519 (2016).

68 Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, doi:10.1093/bioinformatics/btz848 (2019).

69 Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069, doi:10.1093/bioinformatics/btu153 (2014).

70 Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251-2252, doi:10.1093/bioinformatics/btz859 (2020).

71 Zankari, E. *et al.* Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother* **67**, 2640-2644, doi:10.1093/jac/dks261 (2012).

72 Liu, B., Zheng, D., Jin, Q., Chen, L. & Yang, J. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res* **47**, D687-D692, doi:10.1093/nar/gky1080 (2019).

73 Yin, Y. *et al.* dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res* **40**, W445-451, doi:10.1093/nar/gks479 (2012).

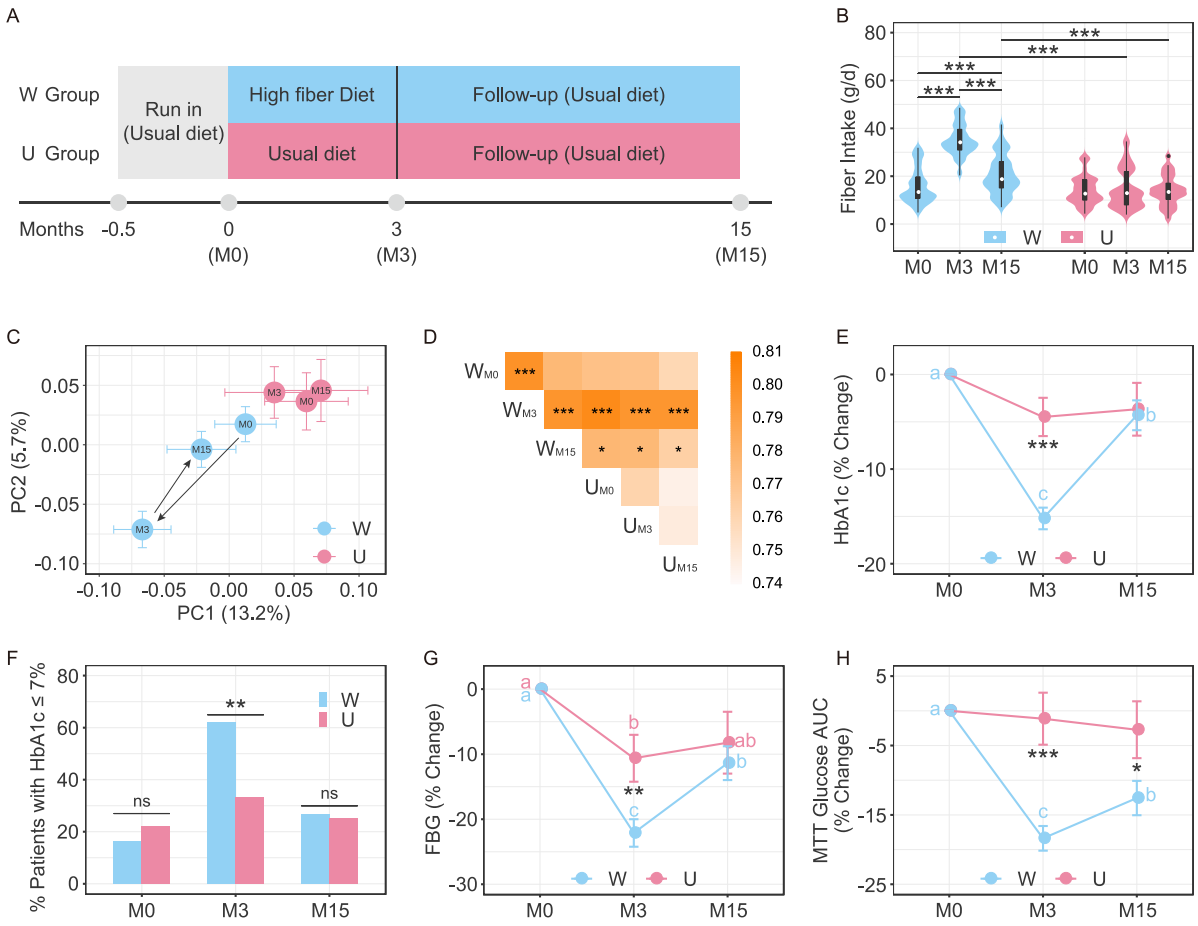
74 Bortolaia, V. *et al.* ResFinder 4.0 for predictions of phenotypes from genotypes. *J Antimicrob Chemother* **75**, 3491-3500, doi:10.1093/jac/dkaa345 (2020).

75 Watts, S. C., Ritchie, S. C., Inouye, M. & Holt, K. E. FastSpar: rapid and scalable correlation estimation for compositional data. *Bioinformatics* **35**, 1064-1066, doi:10.1093/bioinformatics/bty734 (2019).

76 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504, doi:10.1101/gr.1239303 (2003).

77 Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* **47**, W256-W259, doi:10.1093/nar/gkz239 (2019).

78 Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nat Methods* **14**, 1023-1024, doi:10.1038/nmeth.4468 (2017).



781

782

783

784 **Fig.1 Reversible changes of gut microbiota associates with reversible shifts of metabolic phenotypes in patients**
785 **with T2DM.** (A) Study design. Before Run-in, written informed consent, questionnaire of personal information and
786 measuring HbA1c at screening. After Run-in, medical checkup and sample collection at baseline (M0), three months
787 after on the high fiber intervention or usual diet (M3) and one year after the high fiber intervention stopped (M15).
788 (B) Changes of fiber intake. (C) Global changes of the gut microbiome as shown by the principal coordinate analysis
789 based on the Bray-Curtis distance for the 1845 genomes and (D) Average Bray-Curtis distance between the groups.
790 PERMANOVA test (9,999 permutations) was performed to compare the groups. * $P < 0.05$ and *** $P < 0.001$. The
791 color of the square showed the magnitude of average Bray-Curtis distance. (E) Change of HbA1c, (F) The percentage
792 of participants with adequate glycemic control, (G) Fasting blood glucose, and (H) The glucose area under the curve
793 (AUC) in a meal tolerance test (MTT). For (E), (G) and (H), data shown as percent changes from baseline (± S.E.M).
794 Friedman test followed by Nemenyi post-hoc test was used for comparison in the same group, compact letters reflect
795 significance ($P < 0.05$). n = 67 in W group and n = 28 in U group. Mann-Whitney test (two-sided) was used for
796 comparison between W and U at the same time point, * $P < 0.05$, ** $P < 0.01$ and *** $P < 0.001$. n = 74 in W (M0)
797 (For panel H, n=72), n = 74 in W (M3), n = 67 in W (M15), n = 36 in U (M0), n = 36 in U (M3) and n = 28 in U (M15).

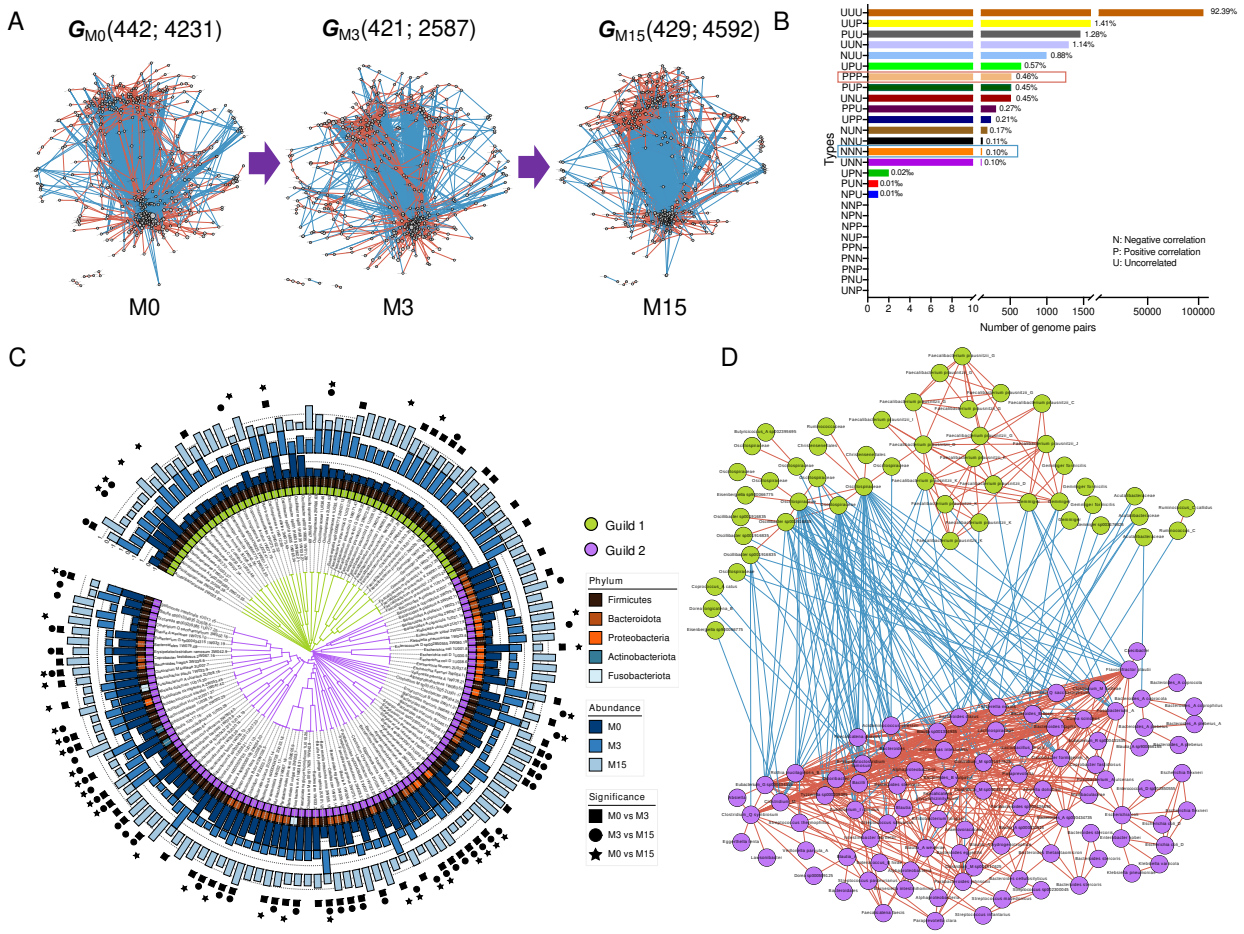


Fig. 2. Two competing guilds of bacteria constitute a robust seesaw-like network despite the profound global changes in the gut microbial ecosystem induced by introduction and withdrawal of the high fiber intervention. (A) Co-abundance networks of the prevalent genomes in W group at M0, M3 and M15 during the trial, denoted as $\boldsymbol{G}_{M0}$(442; 4231), $\boldsymbol{G}_{M3}$(421; 2587) and $\boldsymbol{G}_{M15}$(429; 4592), numbers in parenthesis are order and size of the network. The correlations between genomes were calculated using FastSpar, n = 67 patients. All significant correlations with $P \leq 0.001$ were included. Edges between nodes represent correlations. Red and blue colors indicate positive and negative correlations, respectively. Node size indicates the average abundance of the genomes. The layout of the nodes and edges was determined by Edge-weighted Spring Embedded Layout with correlation efficient as weight. (B) The distribution of different types of correlations of the genome pairs during the trial. The 3 letters show the correlations of the genome pairs at M0, M3 and M15 subsequently. Stable correlations, NNN and PPP, were highlighted (C) Average clustering of the 141 nodes based on their robust positive and negative correlations showed two clusters (green and purple range). The bar plots show the abundance changes of each node throughout the trial, which is expressed as median abundance with Z-score transformation. The differences of each node over time were tested using the Friedman test followed by Nemenyi post-hoc test. $P < 0.05$ was considered as significant. This panel was plotted using iTOL. (D) The seesaw-like network with the 141 nodes in two polarizing clusters. Edges between nodes represent correlations. Red and blue colors indicate positive and negative correlations, respectively. For (C) and (D), the color of the node represents the members in the two guilds: green for Guild 1 and purple for Guild 2.

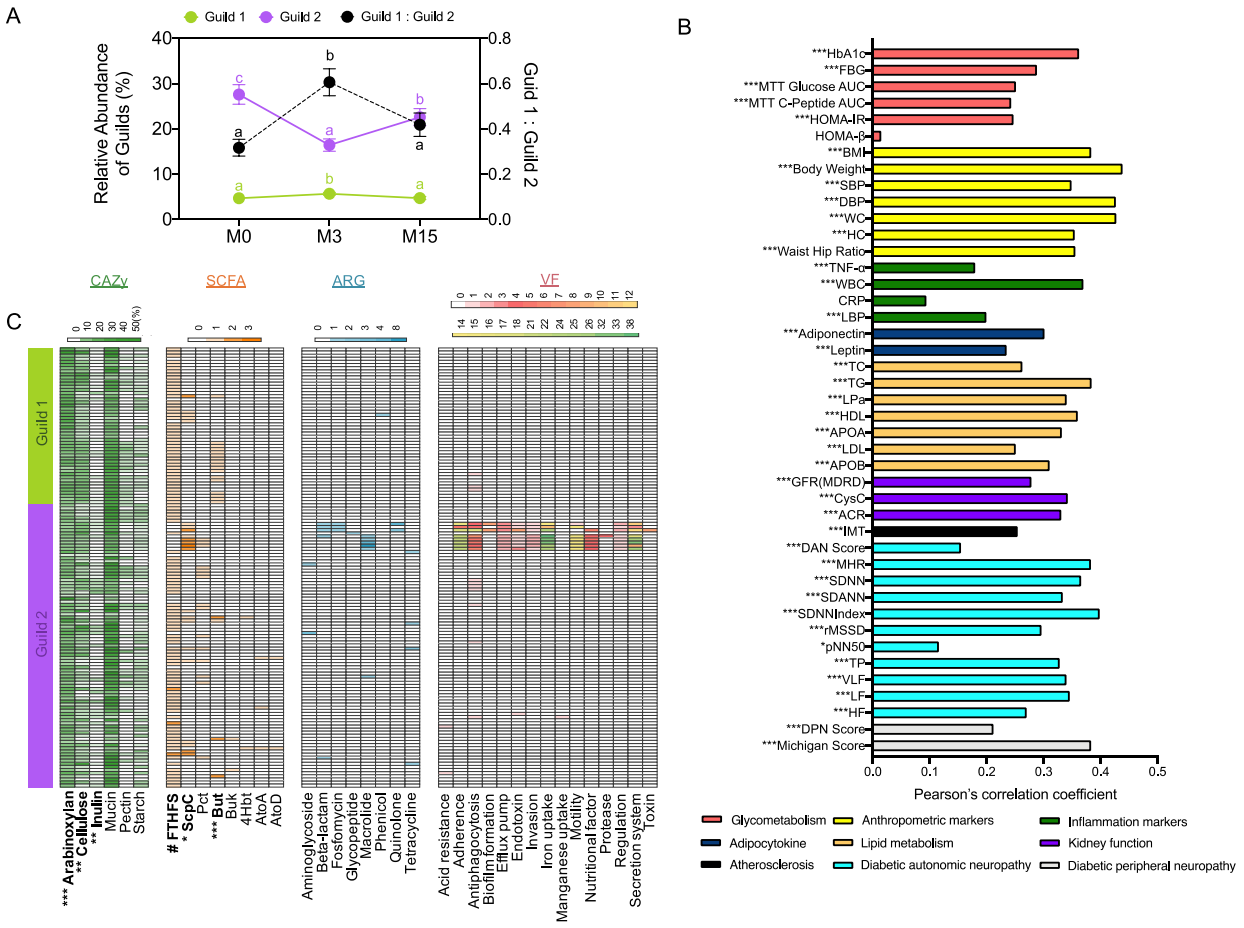


Fig. 3. The balance between the two competing guilds in the seesaw-like network was associated with the metabolic health of patients with type 2 diabetes. (A) Change of the total abundance of Guild 1, Guild 2, and their ratio across the trial in the W group. Friedman test followed by Nemenyi test was used to analyze the difference between time points. Compact letters reflect the significance at $P < 0.05$. (B) Random Forest regression with leave-one-out cross-validation was used to explore the associations between the 141 genomes and the clinical parameters. The bar plot shows the Pearson's correlations coefficient between the predicted and measured values. The asterisk before the parameter's name shows the significance of the Pearson's correlations. P values were adjusted by Benjamini & Hochberg's method. * adjusted $P < 0.05$, ** adjusted $P < 0.01$ and adjusted *** $P < 0.001$. BMI, body mass index; SBP, systolic blood pressure; DBP, diastolic blood pressure; WC, waist circumference; HP, hip circumference; TNF-α, tumor necrosis factor-α; WBC, white blood cell count; CRP, C-reactive protein; LBP, lipopolysaccharide-binding protein; TC, total cholesterol; TG, triglyceride; Lpa, lipoprotein a; HDL, high-density lipoprotein; APOA, apolipoprotein A; LDL, low-density lipoprotein; APOB, apolipoprotein B; GFR (MDRR), glomerular filtration rate; CysC, Cystatin C; ACR, urinary microalbumin to creatinine ratio; IMT, intima-media thickness; DAN, diabetic autonomic neuropathy score; MHR, mean heart rate; SDNN, standard deviation of NN intervals; SDANN, standard deviation of the average NN intervals calculated over 5 minutes; SDNNIndex, mean of standard deviation of NN intervals for 5-minute segments; rMSSD, root-mean-square of the differences of successive NN intervals; pNN50, percentage of the interval differences of successive NN intervals greater than 50 ms; TP, total power; VLF, very low frequency power; LF, low frequency power; HF, high frequency power; DPN, diabetic peripheral neuropathy score. (C) Differences in genetic capacity of carbohydrate substrate utilization (CAZy), short-chain fatty acid production (SCFA), number of antibiotic resistance genes (ARG) and number of virulence factor genes (VF). (C) The heatmaps show the proportion (CAZy) or gene copy numbers (SCFA, ARG and VF) of each category in each genome. For carbohydrate substrate utilization, CAZy genes were predicted in each genome. The proportion of CAZy genes for a particular substrate was calculated as the number of the CAZy genes involved in its utilization divided by the total number of the CAZy genes. Arabinoxylan-related CAZy families: CE1, CE2, CE4, CE6, CE7, GH10, GH11, GH115,

GH43, GH51, GH67, GH3 and GH5; cellulose-related: GH1, GH44, GH48, GH8, GH9, GH3 and GH5; inulin-related: GH32 and GH91; mucin-related families: GH1, GH2, GH3, GH4, GH18, GH19, GH20, GH29, GH33, GH38, GH58, GH79, GH84, GH85, GH88, GH89, GH92, GH95, GH98, GH99, GH101, GH105, GH109, GH110, GH113, PL6, PL8, PL12, PL13 and PL21;pectin-related: CE12, CE8, GH28, PL1 and PL9; starch-related: GH13, GH31 and GH97. For short chain fatty acid production, FTHFS: formate-tetrahydrofolate ligase for acetate production; ScpC: propionyl-CoA succinate-CoA transferase and Pct: propionate-CoA transferase for propionate production; But: Butyryl–coenzyme A (butyryl-CoA): acetate CoA transferase, Buk: butyrate kinase, 4Hbt: butyryl- CoA: 4-hydroxybutyrate CoA transferase, Ato: butyryl-CoA:acetoacetate CoA transferase (AtoA: alpha subunit, AtoD: beta subunit) for butyrate production. Mann-Whitney test (two-sided) was used to analyze the difference between Guild 1 and Guild 2. # $P < 0.1$, * $P < 0.05$, ** $P < 0.01$ and *** $P < 0.001$. Guild 1 (green bar): n = 50, Guild 2 (purple bar): n = 91.

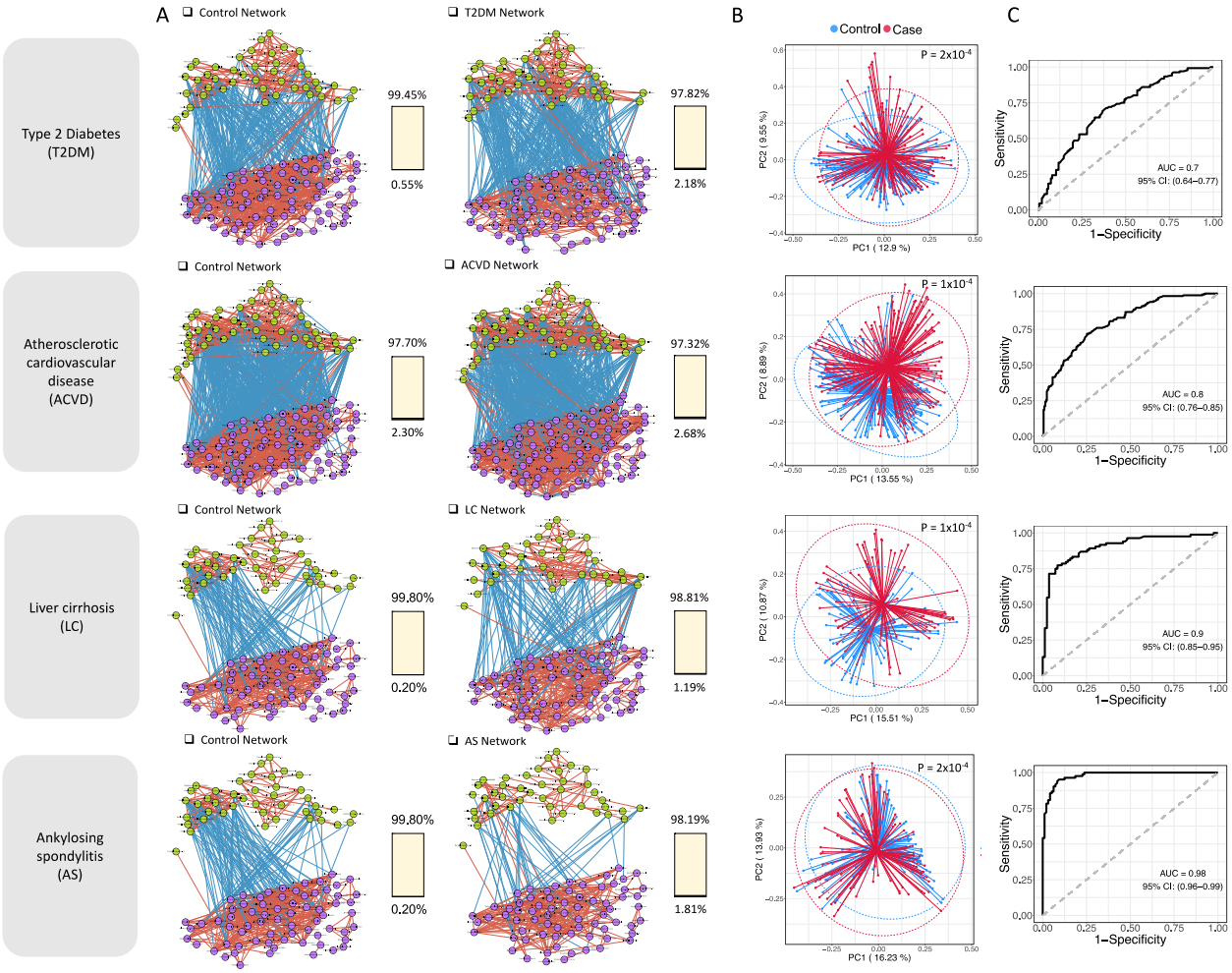


Fig.4. The seesaw networked microbiome signature exists in other independent human cohorts and supports classification models for different diseases. (A) Members of the two competing Guilds in the seesaw networked microbiome signature showed similar ecological interactions in four independent human gut metagenomic datasets. The correlations between the genomes were calculated using FastSpar. All significant correlations ($P \leqslant 0.001$) belonged to seesaw-like network (positive correlations within Guilds and negative correlations between Guilds) were included. Lines between nodes represent correlations, and red and blue colors indicate positive and negative correlations, respectively. The color of the node represents the members in the two competing guilds: green for Guild 1 and purple for Guild 2. The percentage of correlations followed the pattern in the seesaw networked microbiome signature (i.e., positive edges within each guild, negative edges between the 2 guilds) was in yellow, and the ratio of correlations that were negative within each guild and positive between the guilds was in black of the 100% stacked bar. (B) The composition of the c microbiome signature was different between control and patients in each dataset as shown in the Principal Coordinates Analysis plot based on Bray-Curtis distance. 95% confidence ellipses were projected for control and patients respectively. The p values of the PERMANOVA test were indicated. (C) The microbiome signature supports classification models for the four different diseases. The area under the ROC curve (AUC) of the Random Forest classifier based on the 141 genomes in the microbiome signature to classify control and patients in each dataset. Leave-one-out cross validation was applied. Type 2 diabetes (T2DM): Control n = 136, T2D n=136; Atherosclerotic cardiovascular disease (ACVD): Control n = 171 and ACVD n = 214; Liver cirrhosis (LC): control n = 83 and LC n =84; Ankylosing spondylitis: Control n = 83 and AS n = 97.