Supplementary Figures for

Population-level gut microbiome and its associations with environmental factors and metabolic disorders in Southwest China

Qianyu Qu[1†], Qingyu Dou[2†], Zhejun Xiang[1], Bin Yu[1,3], Lili Chen[1], Zhenxin Fan[4], Xing Zhao[1], Shujuan Yang[1*], Peibin Zeng[1*]

[1]West China School of Public Health and West China Fourth Hospital, Sichuan University, Chengdu, China

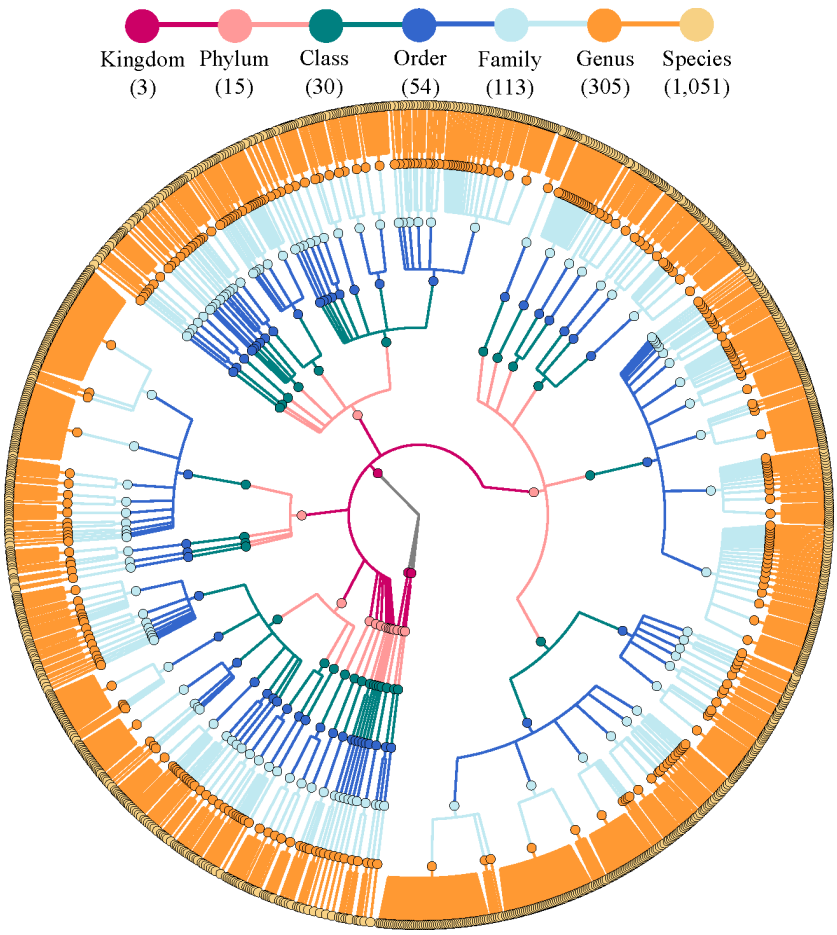
[2]National Clinical Research Center of Geriatrics, Geriatric Medicine Center, West China Hospital, Sichuan University, Chengdu, China

[3]Institute for Disaster Management and Reconstruction, Sichuan University-The Hong Kong Polytechnic University, Chengdu, China

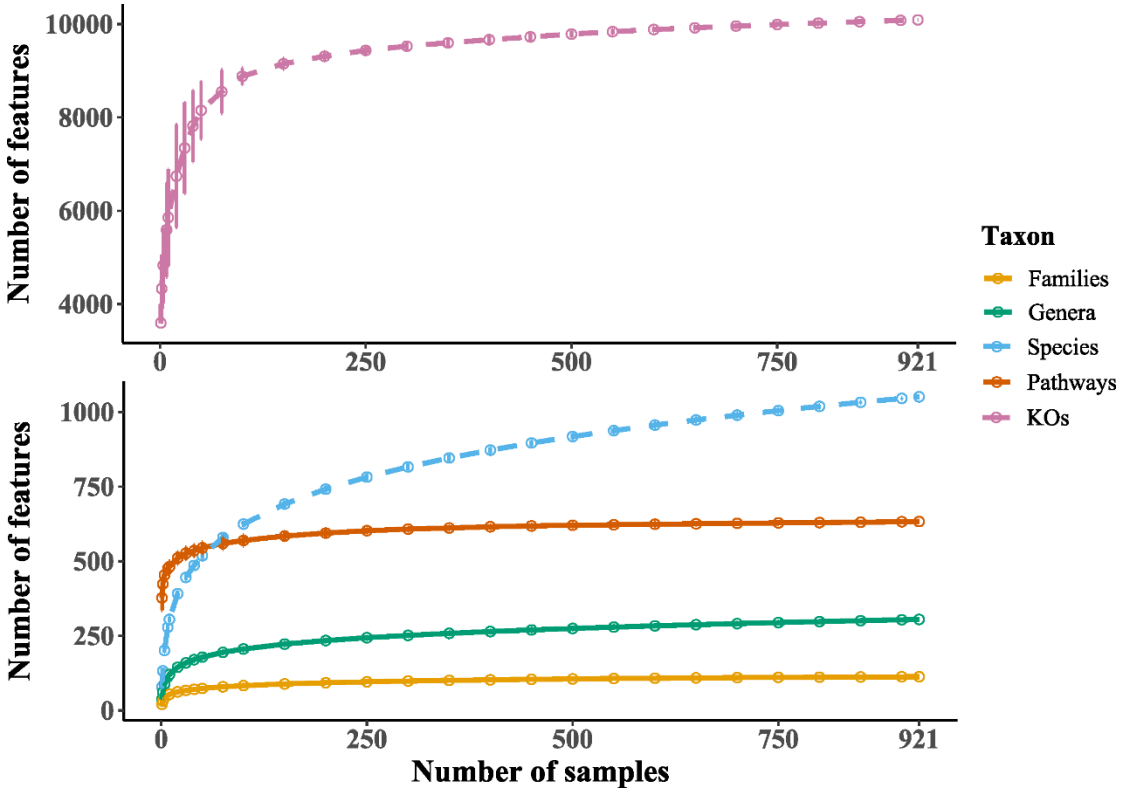
[4]College of Life Sciences, Sichuan University, Chengdu, China

[*]Correspondence: Peibin Zeng: zengpeibin@live.cn; Shujuan Yang: rekiny@126.com

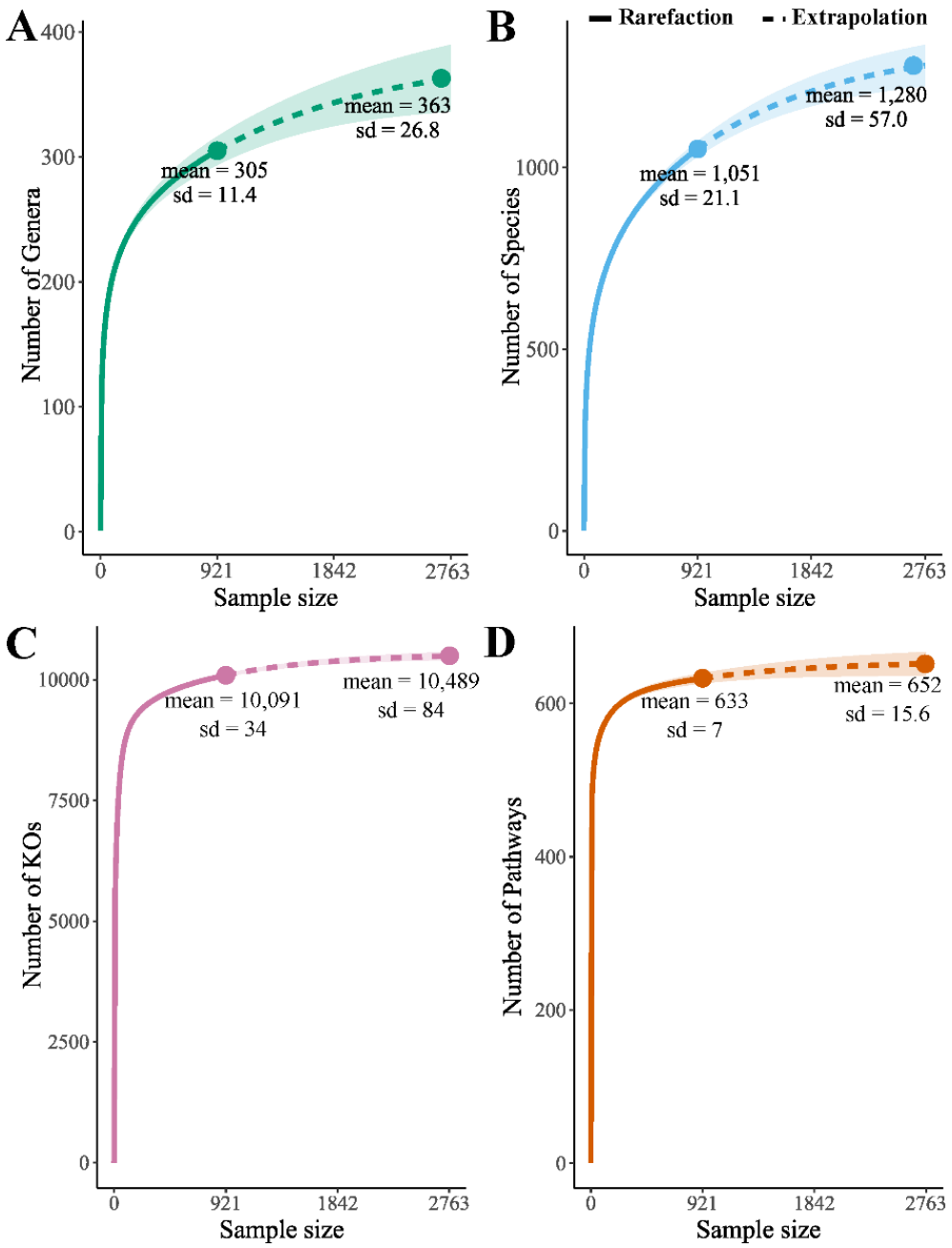
†Authors have contributed equally to this work and share the first authorship



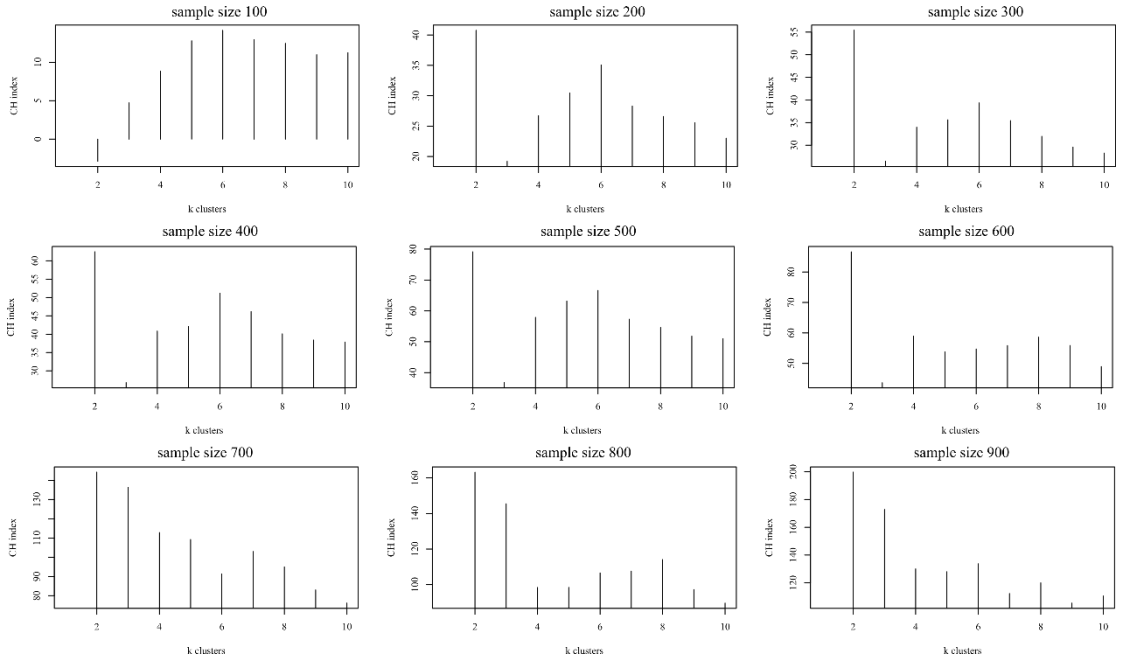
Supplementary Figure 1 | The taxonomic tree of microbial taxonomies predicted by MetaPhlAn3. The dataset contained 1,051 species (1040 bacteria, 6 archaea, and 5 eukaryotes) belonging to 3 kingdoms, 15 phyla, 30 classes, 54 orders, 113 families, and 305 genera. Each coloured circle represents a taxonomic entity, and different colours indicate different taxonomy levels. From the inner to outer circles, the taxonomic levels range from kingdoms to species.



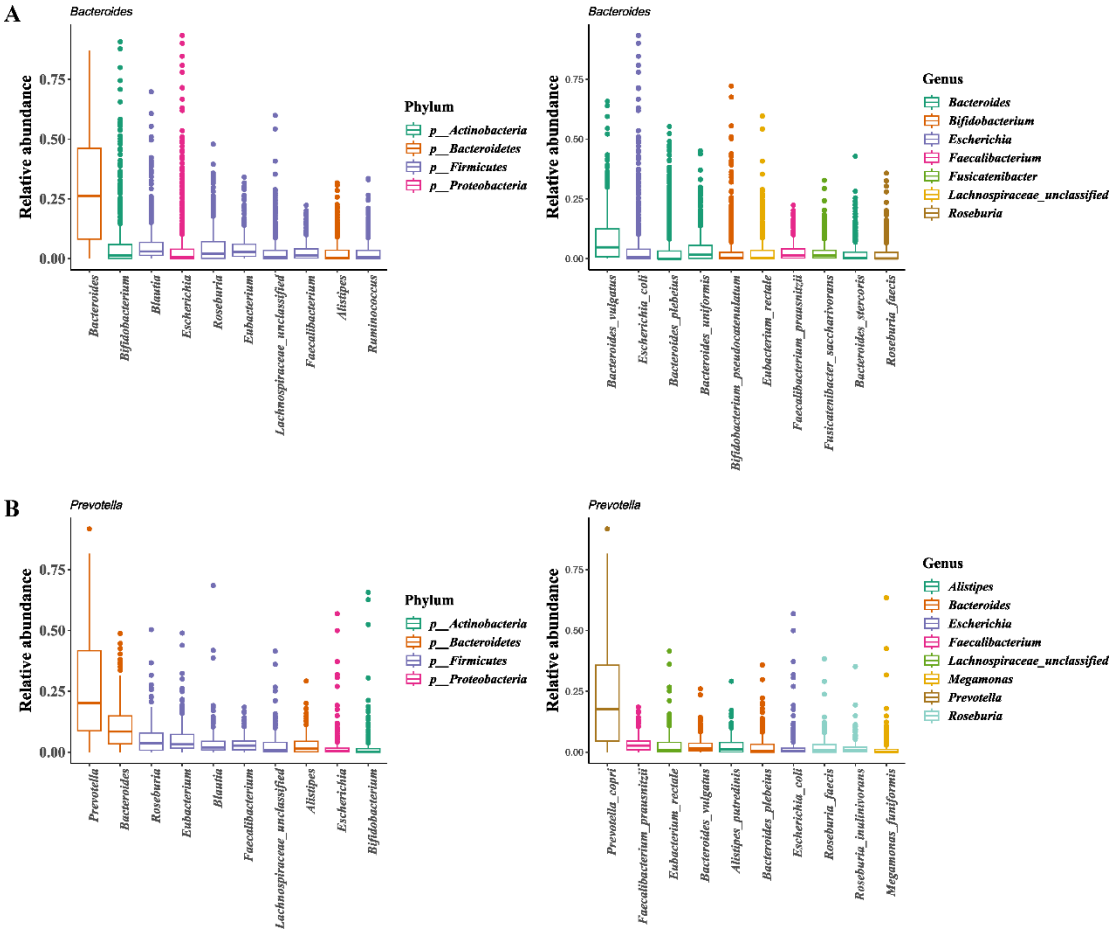
Supplementary Figure 2 | Species accumulation curve. Estimate microbiome feature richness by performing 100 random permutations on abundance data of taxa, KOs, and pathways. Sparsification methods were used to calculate and plot species accumulation curves.



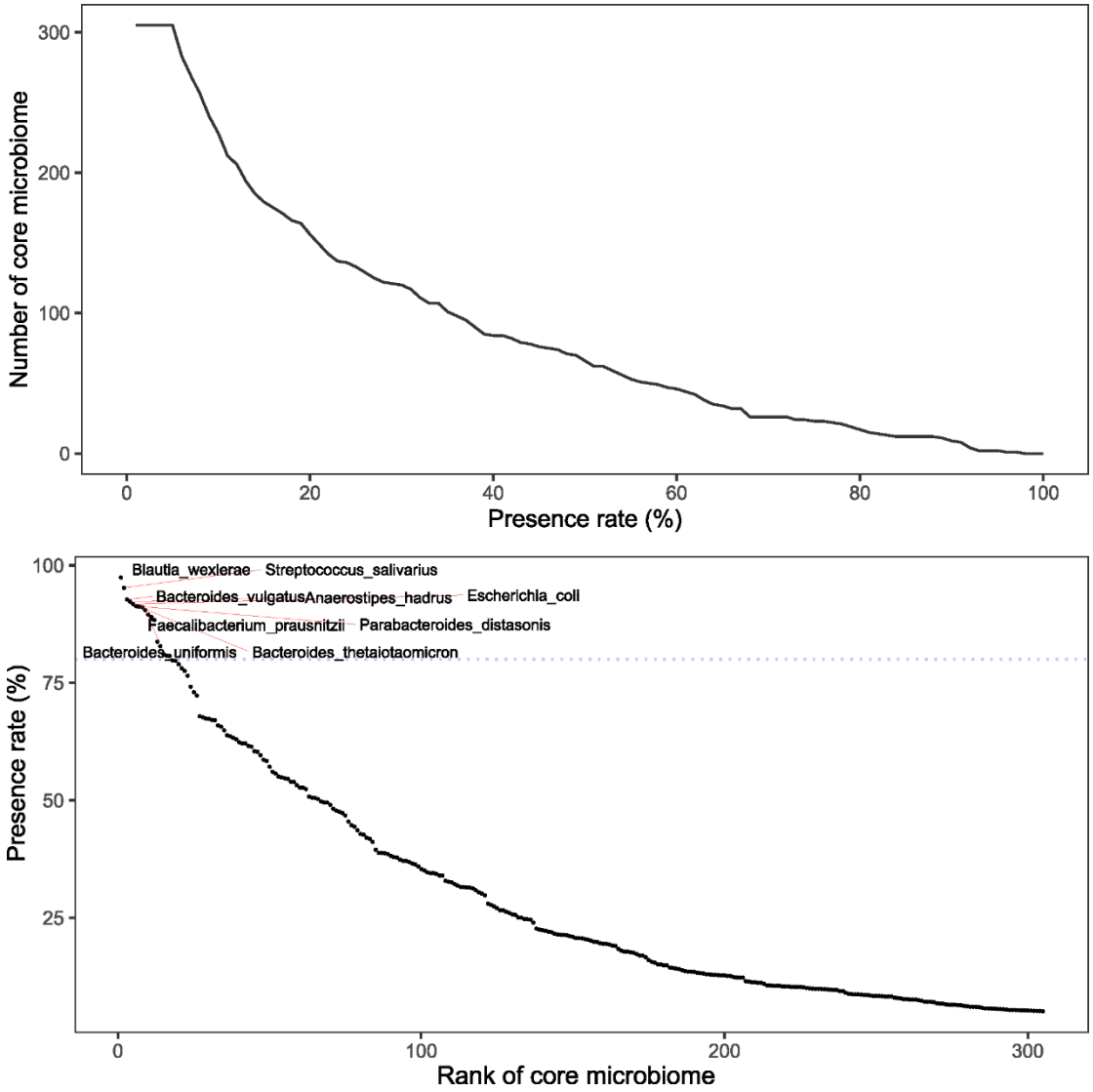
Supplementary Figure 3 | Estimation of the total number of microbiome features in the population in Southwest China. The rarefaction and extrapolation sampling curve for **(A)** genera, **(B)** species, **(C)** KOs, and **(D)** pathways. The extrapolated part of the rarefaction curve is shown dotted. The SD of the estimate is shaded and the asymptotic richness estimate is shown.



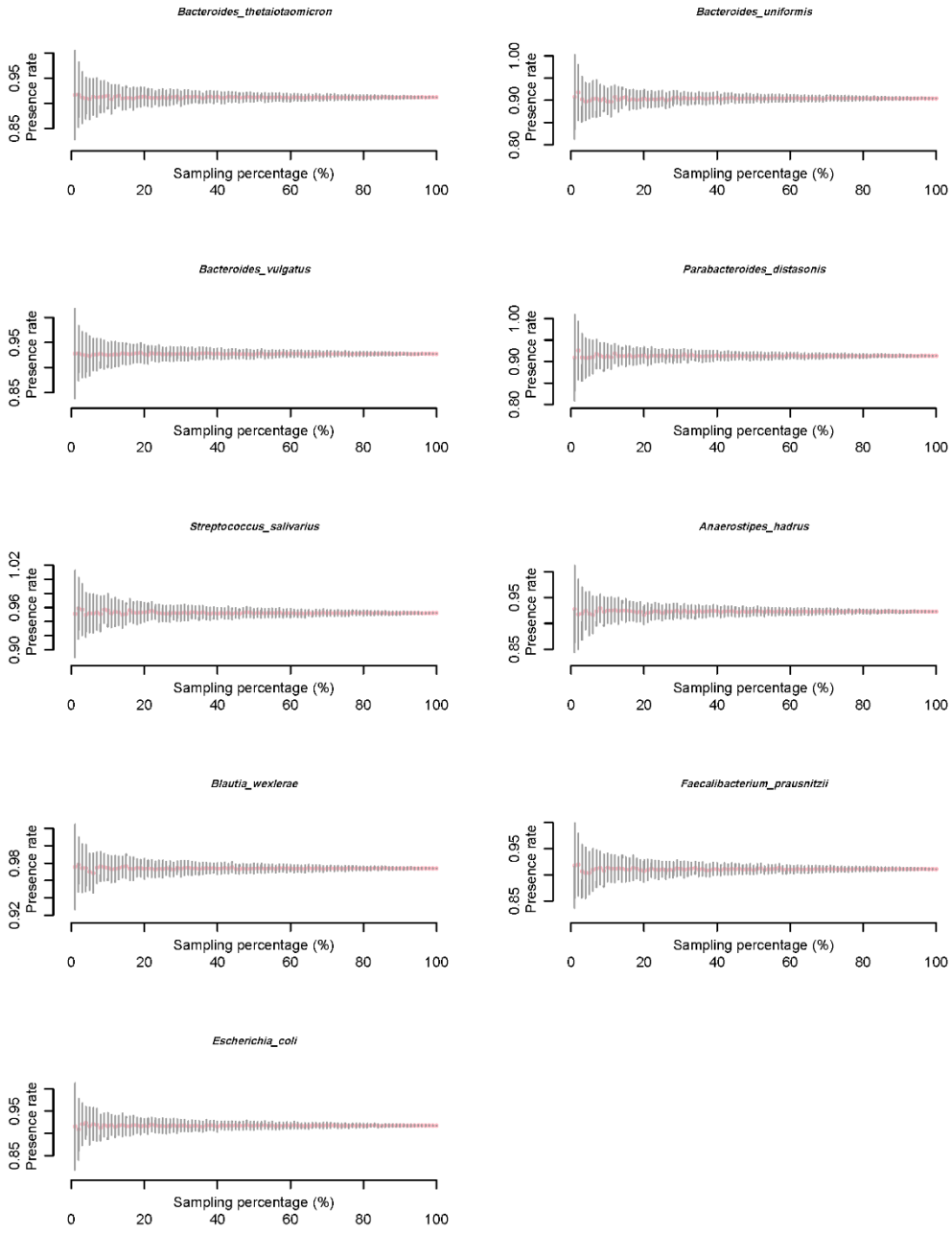
Supplementary Figure 4 | The Calinski-Harabasz (CH) index of clustering under different sample sizes. A repeated enterotype analysis on randomly down-sampled datasets to evaluate the effect of sample size changes on the overall clustering. Samples were clustered based on relative abundances at the genus level using the Jensen-Shannon divergence (JSD) distance and the Partitioning Around Medoids (PAM) clustering algorithm. The Calinski-Harabasz (CH) index was used to assess the optimal number of clusters in different sample sizes. The CH index under different sample sizes was present using the function *plot* in R.



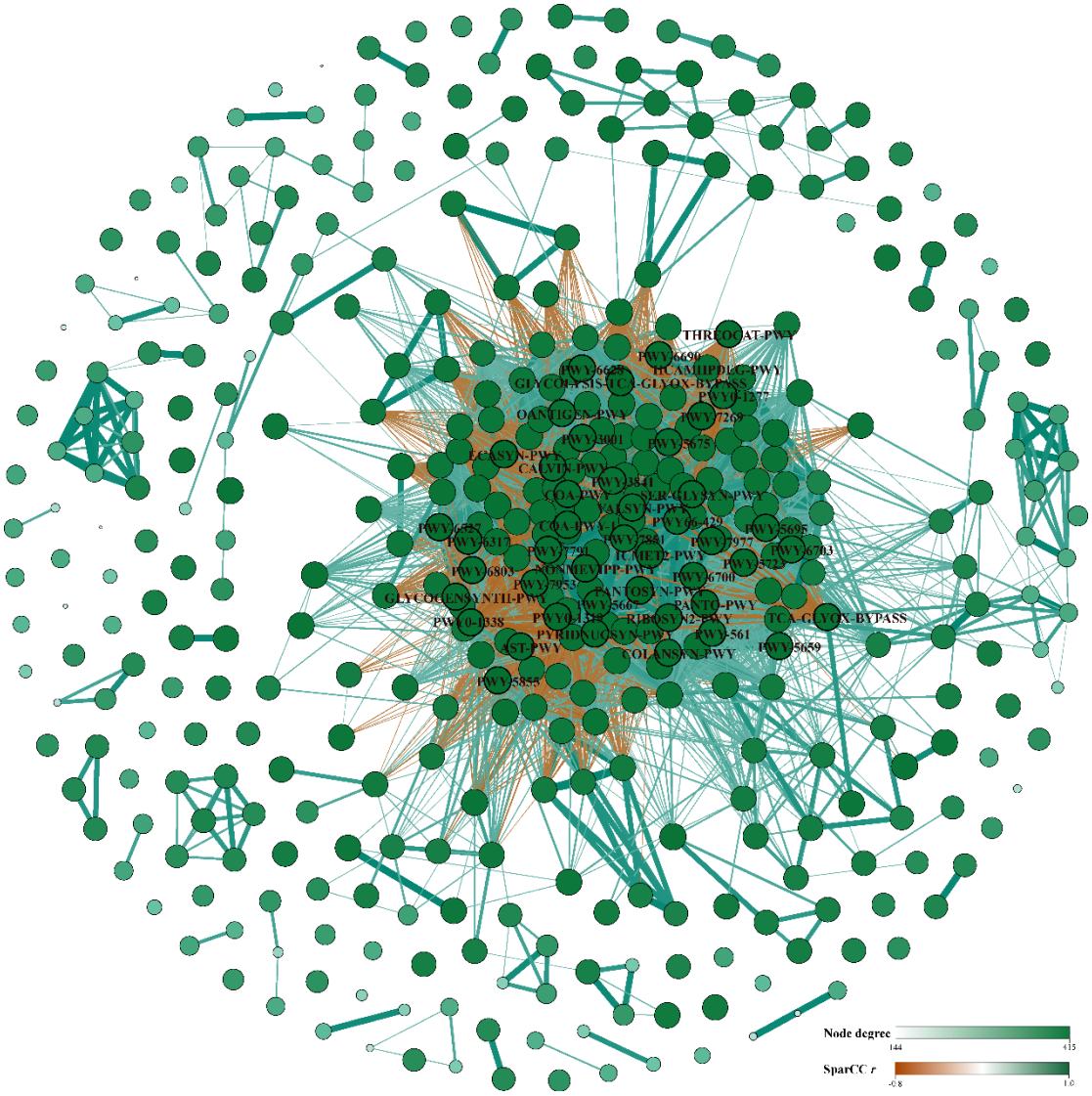
Supplementary Figure 5 | The distribution of genera or species across different enterotypes. Relative abundance variation box plot for the 10 most abundant genera or species in **(A)** cluster 1 and **(B)** cluster 2**.** Genera are coloured by their respective phyla, and species are coloured by their respective genera.



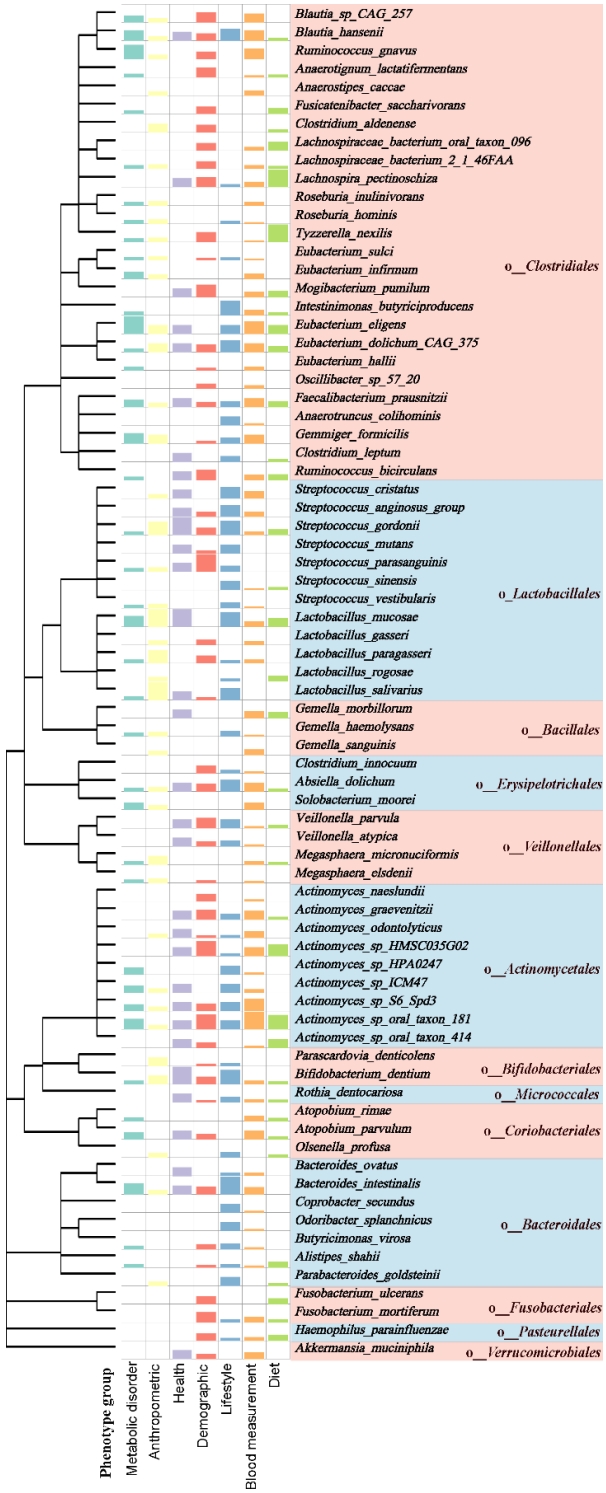
Supplementary Figure 6 | Core species under different prevalence rates. A bootstrapping-based selection approach was used to identify the core microbiota. By subsampling the cohort with sampling ratios of 1% to 100%, the prevalences of each species at different subsampling levels were obtained. Microbial features with a prevalence ≥ 90% were defined as the core microbiome.



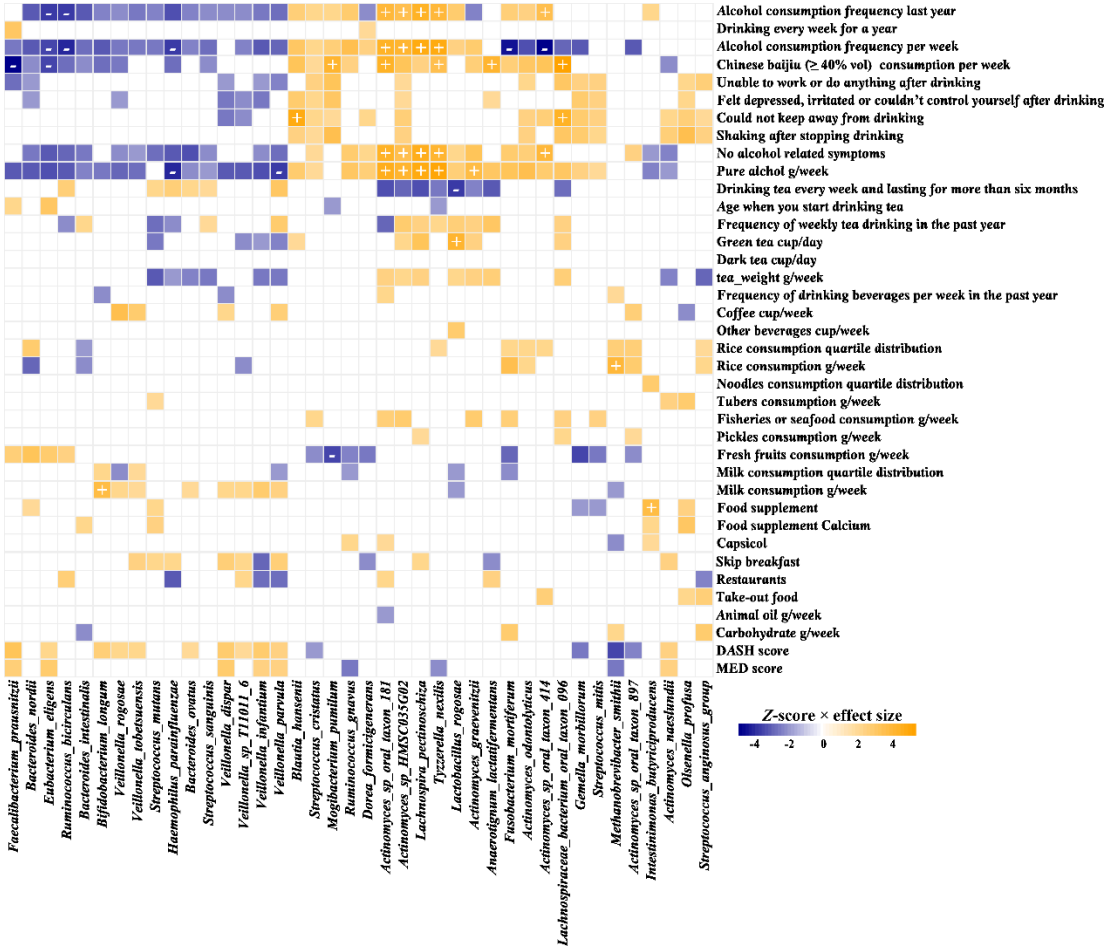
Supplementary Figure 7 | The prevalence of nine core species under different sampling ratios.



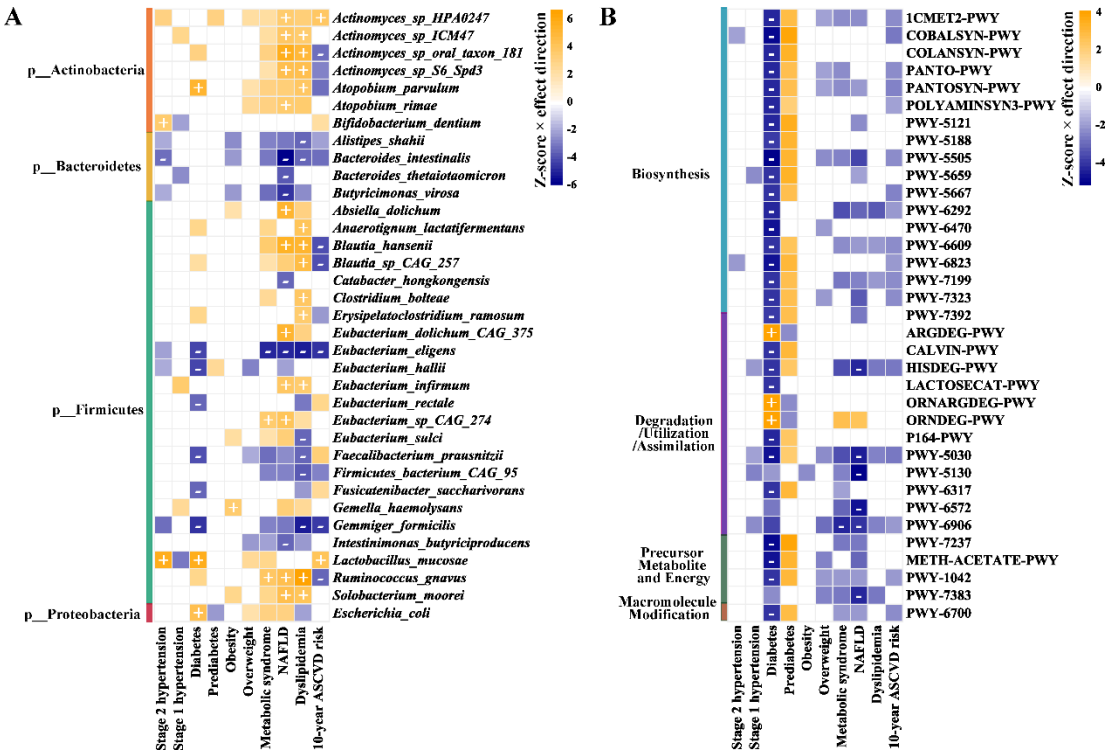
Supplementary Figure 8 | Microbial MetaCyc pathways co-abundance networks. In total, 79,297 significant (FDR < 0.05) co-abundance relationships were identified between 430 MetaCyc pathways. Only showed the edges with $|r| > 0.60$ to simplify the figure. The size of the nodes represents the level of degrees, and the thickness of the edges represents the absolute value of Pearson *r*. Created by Gephi.



Supplementary Figure 9 | Overview of microbiome and phenotype associations. Figure shows the number of significant associations (FDR < 0.05) between each phenotype group and species (only shown the species with significant associations number ≥ 4), clustered by taxonomy. Bar heights represent the number of associations relative to the maximal number of associations for the phenotype group.



Supplementary Figure 10 | Heat map of microbial species associated with Diet phenotypes. Top 40 microbial species with the highest number of associations are clustered by association *Z*-score (multivariate linear regression of CLR-transformed relative abundance of taxa, correcting for annual household income, city, defecation frequency, sequencing batch, and sampling month) using hierarchical clustering. Associations are coloured by direction of effect (blue, negative; orange, positive), with associations significant at study-wide FDR < 0.05 marked with plus and minus for positive and negative correlations, respectively. Coloured associations without a label indicate nominally significant associations ($P < 0.05$).



Supplementary Figure 11 | Heat map of microbial species associated with metabolic disorders and health status. The top 35 microbial species and pathways with the highest number of associations are clustered by association *Z*-score (multivariate linear regression of CLR-transformed relative abundance of taxa, correcting for city, defecation frequency, sequencing batch, and sampling month) using hierarchical clustering for **(A)** species and **(B)** pathways. Associations are coloured by direction of effect (blue, negative; orange, positive), with associations significant at study-wide FDR < 0.05 marked with plus and minus for positive and negative correlations, respectively. Coloured associations without a label indicate nominally significant associations ($P < 0.05$).