Supporting Information

Deterministic assembly processes shaping habitat-specific glycoside hydrolase composition

List:

Figure S1-S13, Data S1-S4.

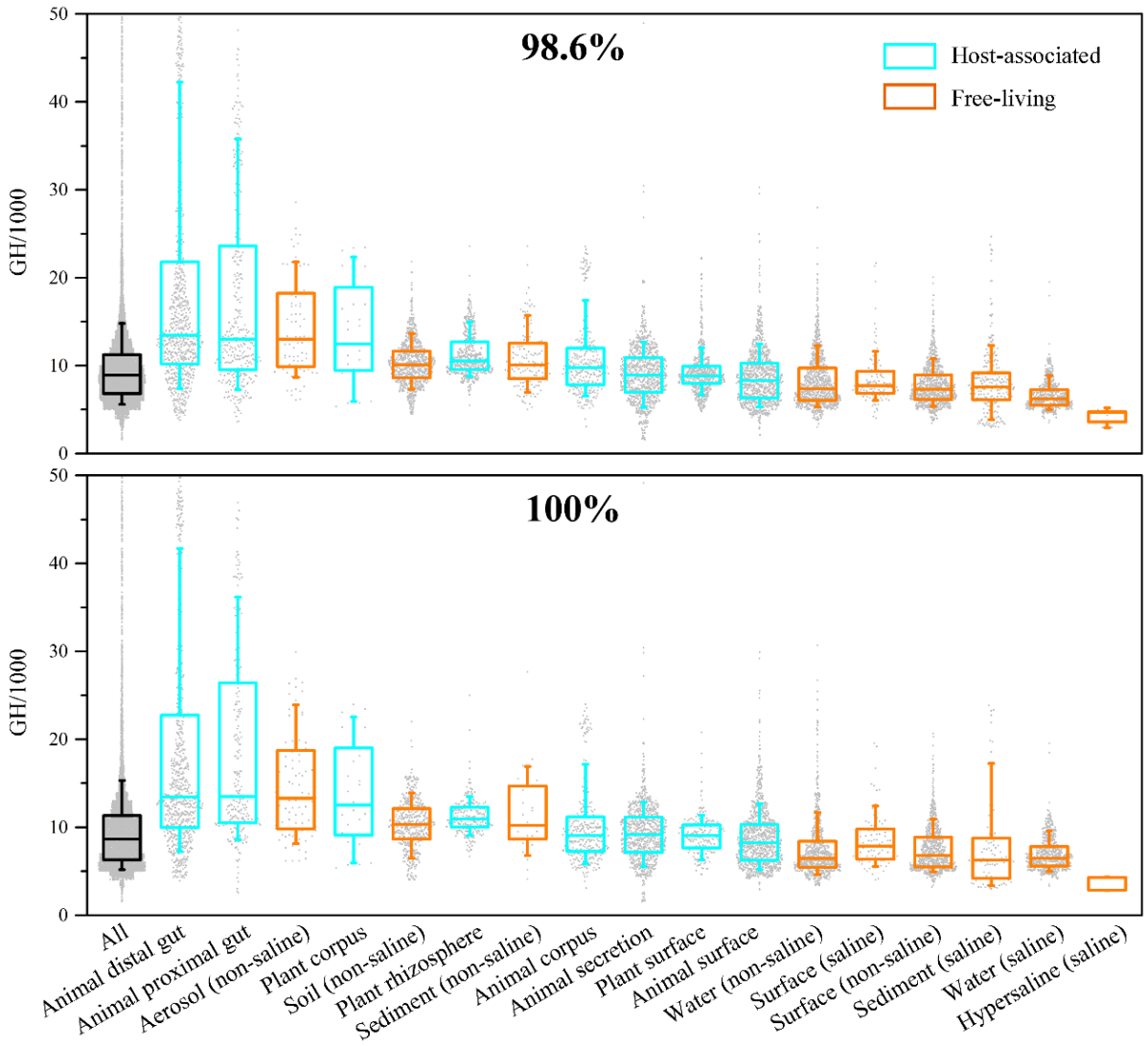

Figure S1. Differences in GH abundance within prokaryotic communities in different environments. The OTUs corresponding to the CAZy genomes were determined based on a 16S rRNA-V4 region identity of greater than 98.6% or 100%. The GH abundance was calculated separately for the two sample sets, with each grey point representing an EMP sample. For the boxplots, the middle line indicates the median, the box represents the 25th-75th percentile, and the whisker indicates the 10th-90th percentile of observations. EMPO classified 17 microbial environments (level 3) into free-living or host-associated (level 1), and saline or non-saline (for free-living) or animal or plant (for host-associated) (level 2). Host-associated environments are represented in cyan, while free-living environments are represented in orange.

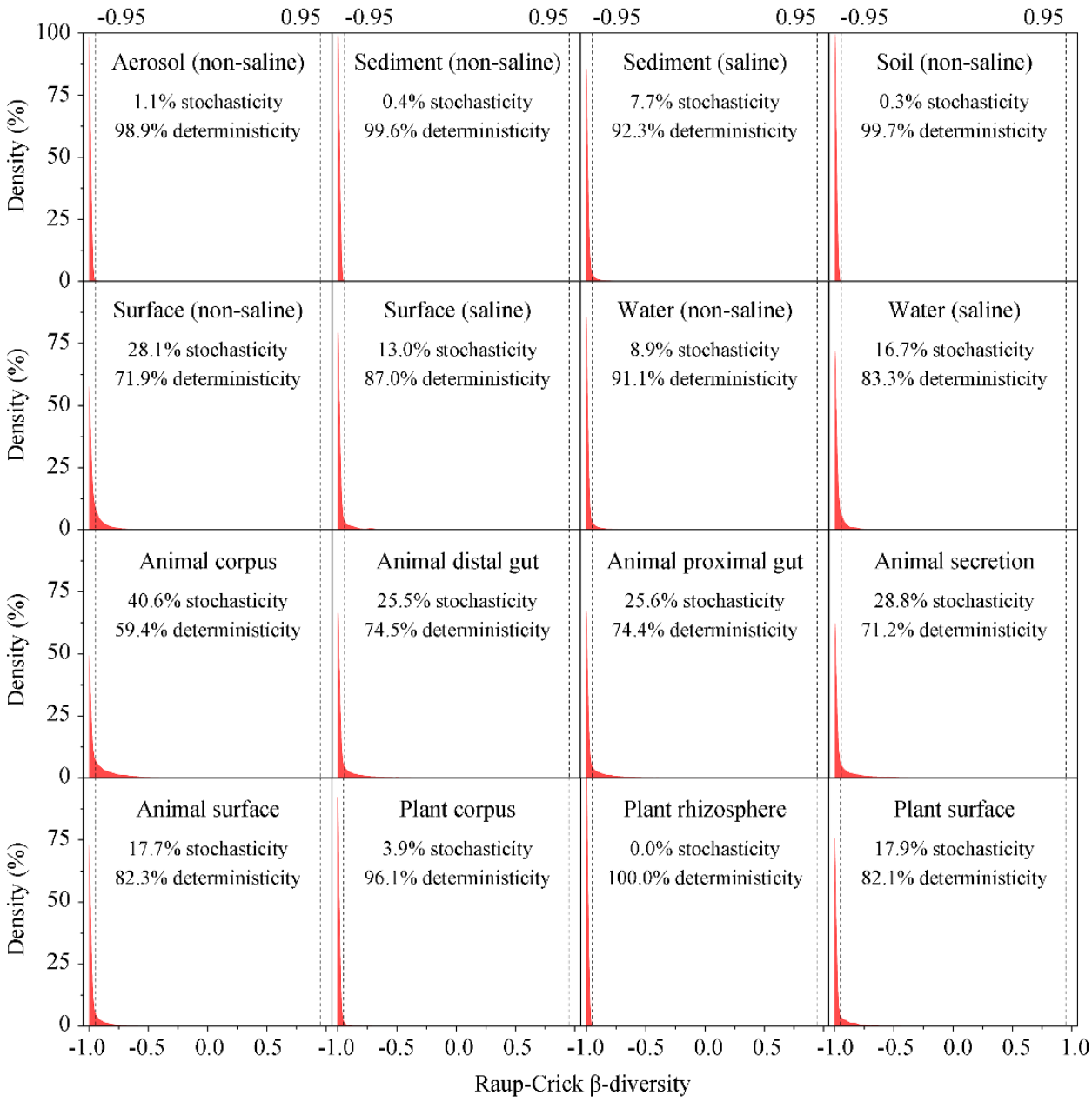

Figure S2. Deterministic processes dominate the assembly of GH communities. Pairwise Raup-Crick between communities indicates that the GH genes are either more similar (-0.95 > Raup-Crick > -1) or less similar (0.95 < Raup-Crick < 1) to each other than expected by random chance (1000 randomizations). Any other value of Raup-Crick indicates that the GH communities are stochastically assembled (-0.95 < Raup-Crick < 0.95). The vertical dashed lines mark the positions of -0.95 and 0.95 in panels. The contribution ratios of deterministic and stochastic processes are also marked in panels.

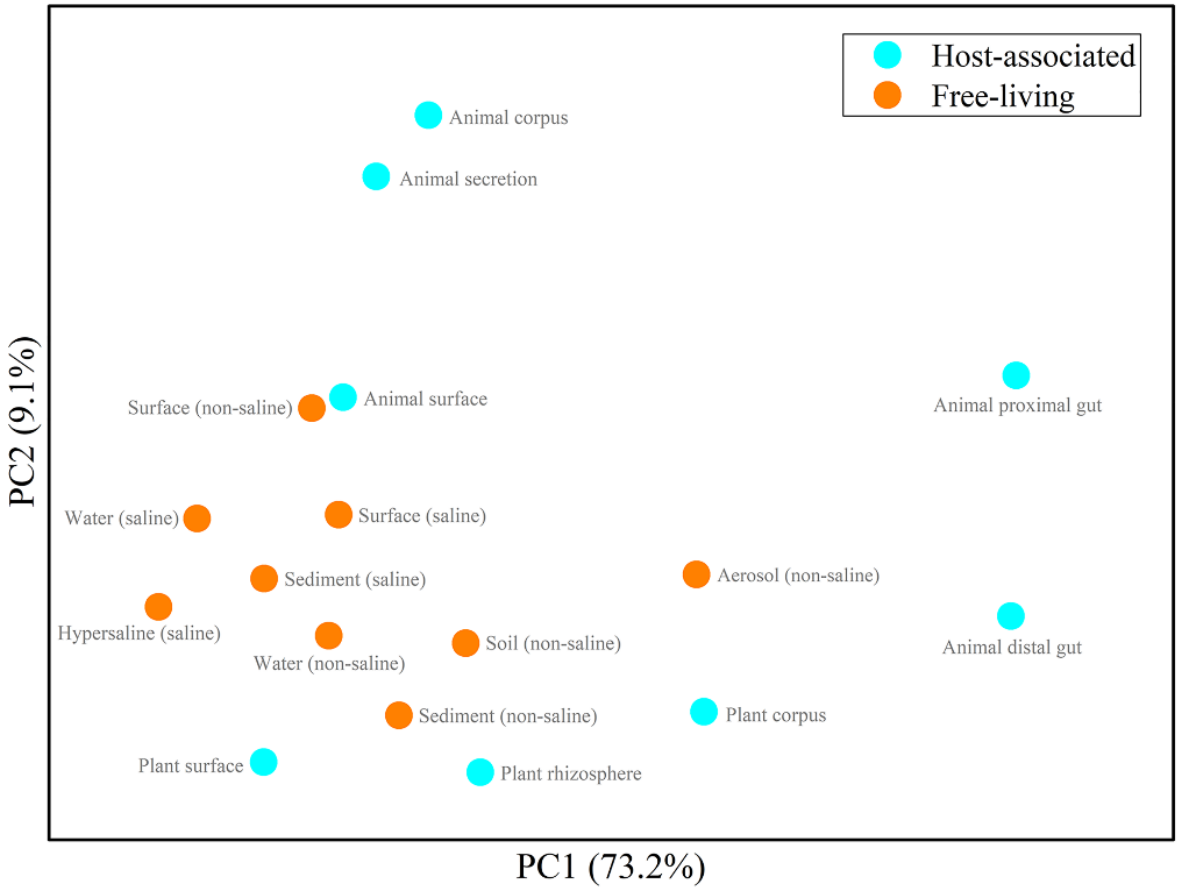

Figure S3. Similarity of GH composition across various environments. Principal coordinate analysis (PCoA) of the GH distribution characteristics in EMPO level 3 environments was performed using the PAST3 software based on the Bray-Curtis method.

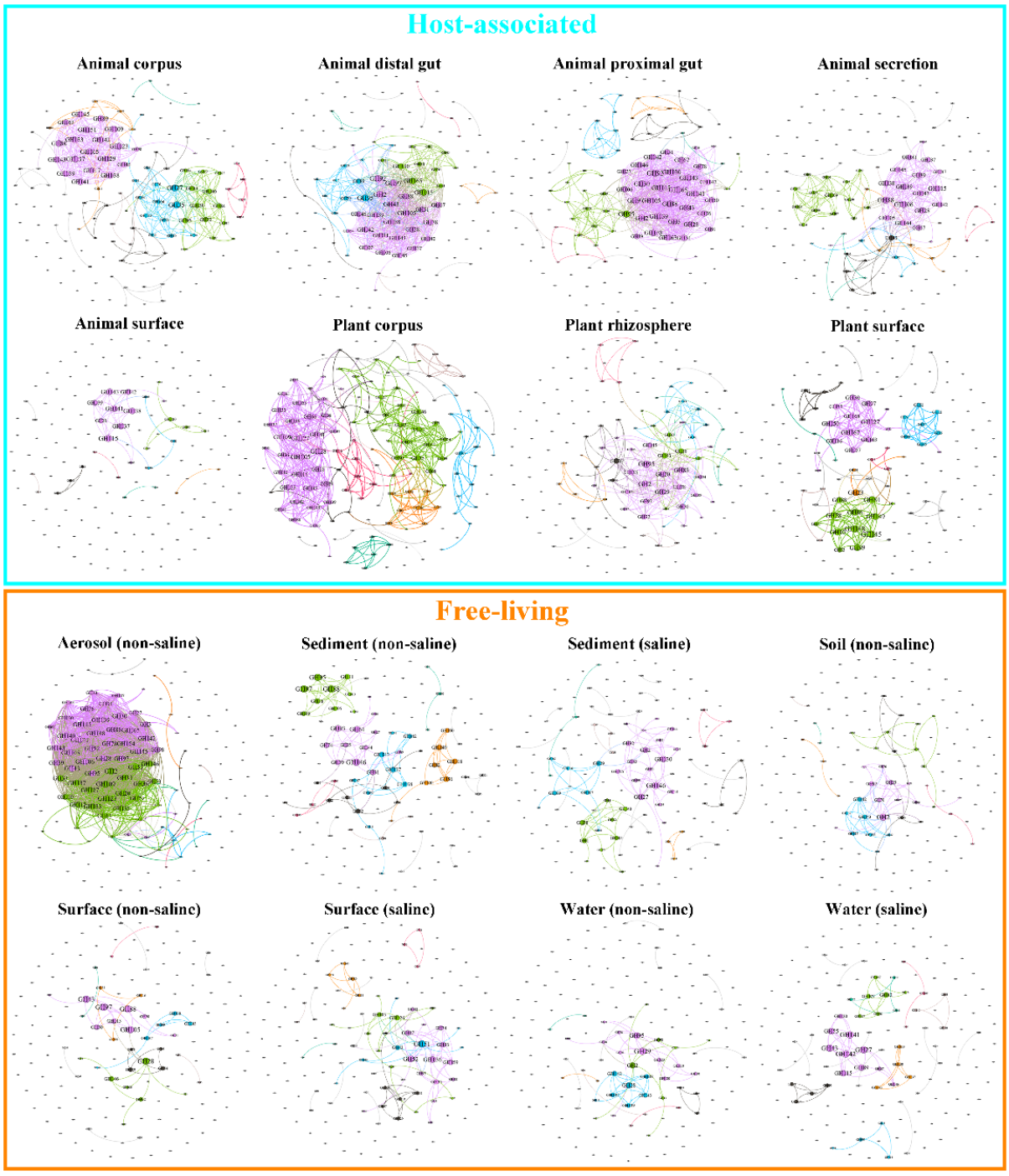

Figure S4. A co-occurrence network diagram exhibiting the correlations between various GH families within the environment. The EMPO classified environmental samples into 17 microbial environments. The hypersaline (saline) environmental samples were too few to be displayed. A co-occurrence network was constructed by Gephi, and the pairwise Spearman's ρ and p values were determined using the "psych" package in R. Values of Spearman's $\rho > 0.8$ and $p < 0.01$ were considered to indicate valid relationships (connecting lines in the network diagram).

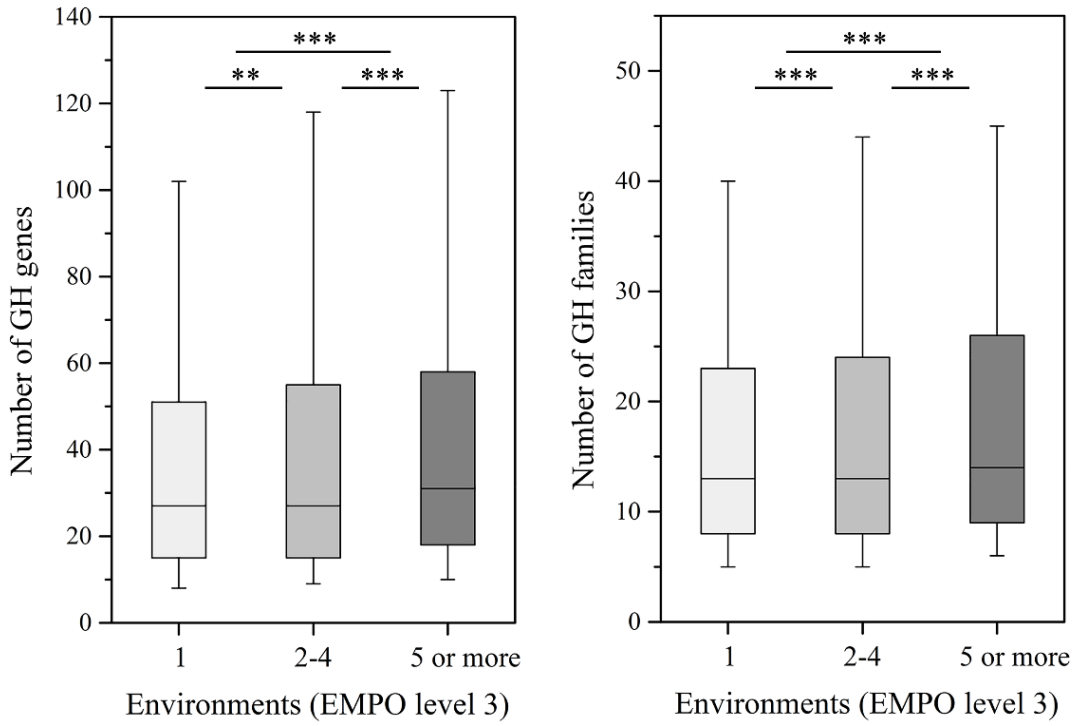

Figure S5. Potential impact of prokaryotic-encoded GHs on their environmental adaptability. Comparisons between bins were analyzed using the Wilcoxon rank test. ** $p < 0.01$, *** $p < 0.001$.

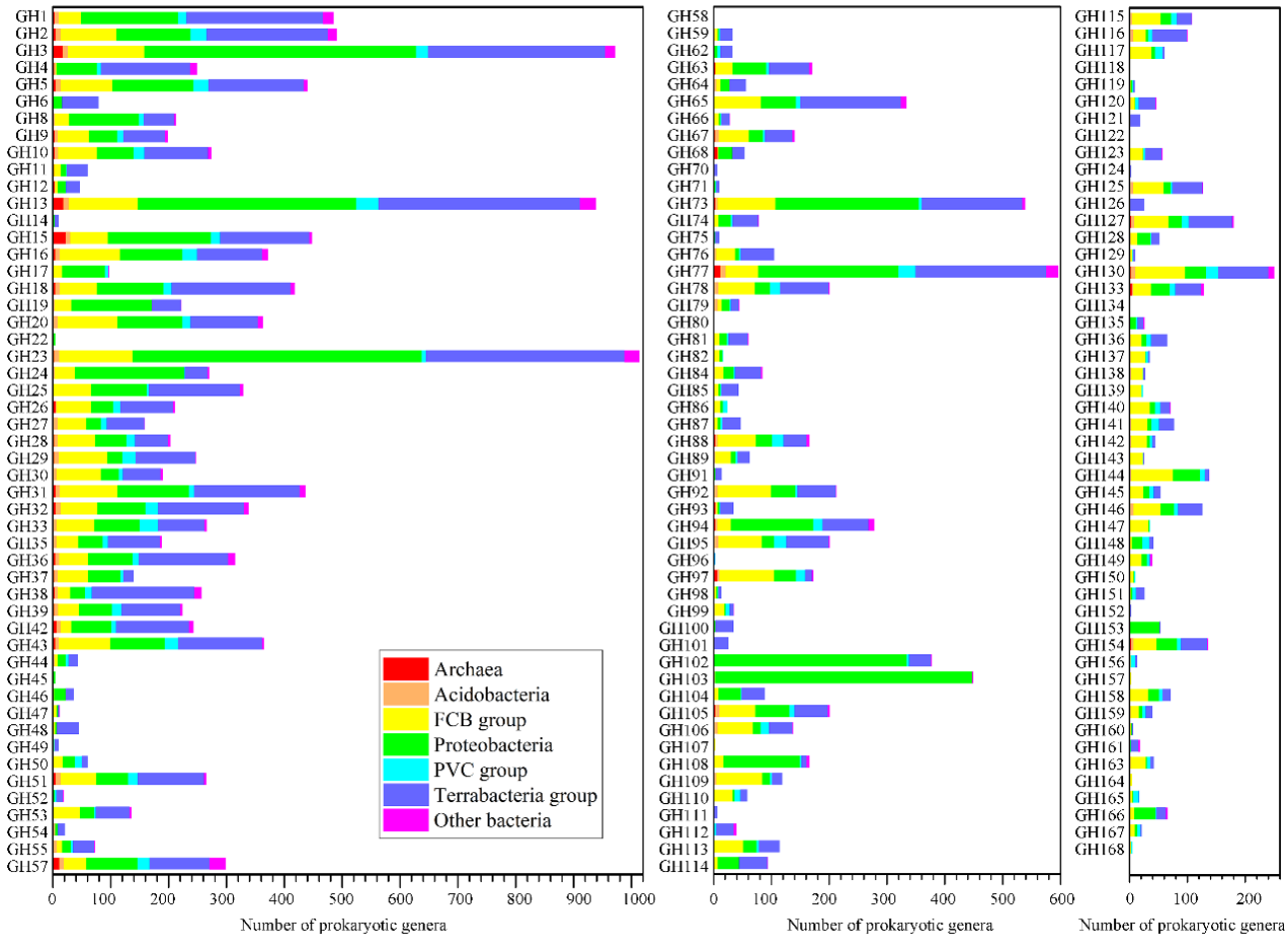

Figure S6. Taxonomic source statistics of GH families in all EMP samples. The number of prokaryotic genera encoding GHs in EMP samples was counted using the NCBI Taxonomy database. The different colors in the figure indicate the prokaryotic genera from distinct taxonomic groups.

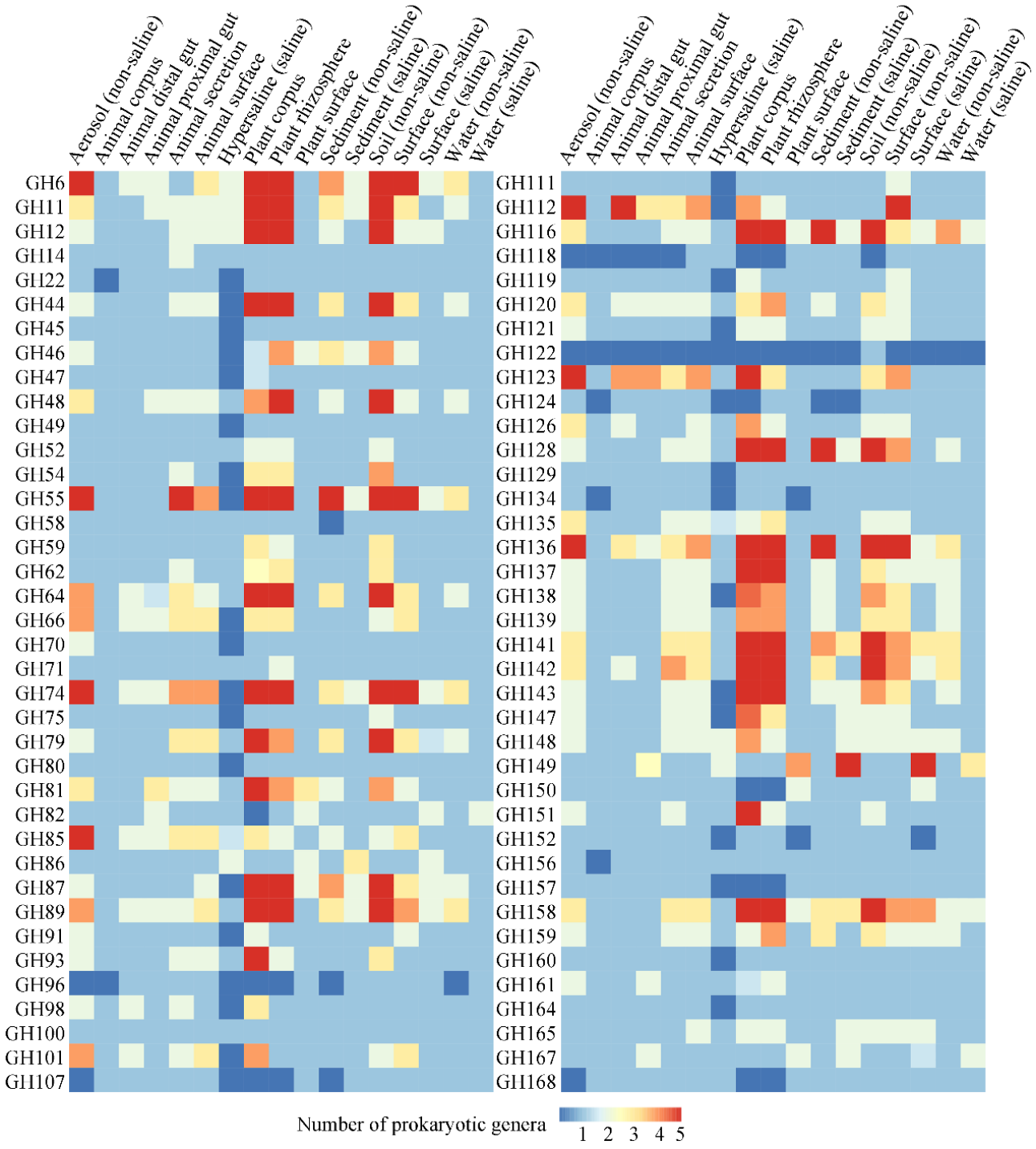

Figure S7. Diverse taxonomic sources of GH families in prokaryotic communities. Statistical analyses were carried out at EMPO level 3. The colors in the figure indicate the median number of source prokaryotic genera of GH families in all EMP samples from the same environment, with increasing values from blue to red. Only 76 GH families with an occupancy of less than 0.75 in all EMP samples are displayed, and the other 76 GH families are shown in Figure 6.

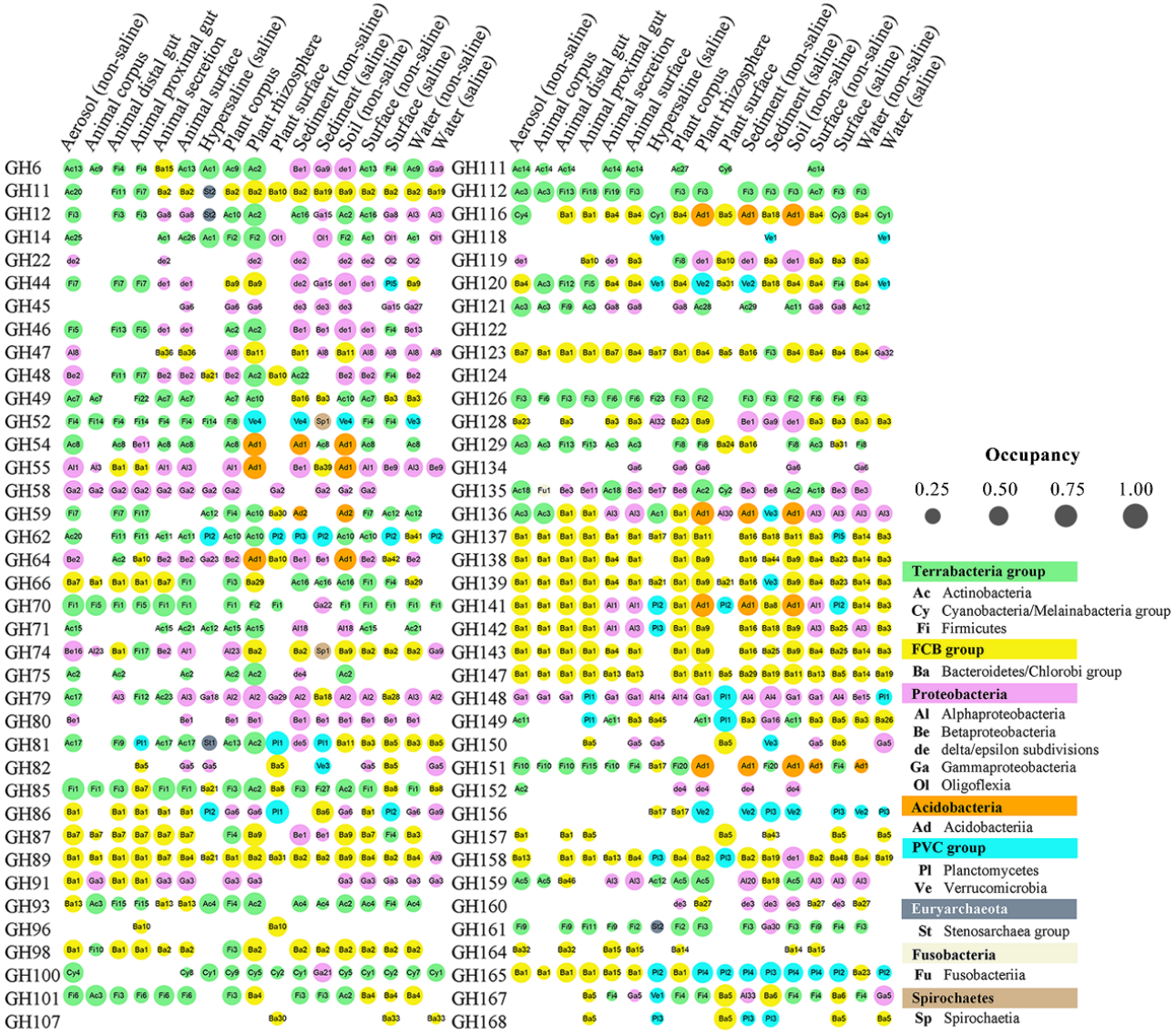

Figure S8. Most common prokaryotic genera of GH families in EMPO environments. The most common prokaryotic genus was defined as the genus that appeared in the most samples and had an occupancy of at least 0.05 in all samples. The EMP samples were classified into 17 microbial environments by EMPO. Different colors represent different taxonomic groups. The circle size indicates the occupancy of the genus in all samples within a single environment. Only 76 GH families with an occupancy of less than 0.75 in all EMP samples are displayed, and the other 76 GH families are shown in Figure 7. Specific information on the prokaryotic genera can be found in Figure S9.

Proteobacteria

Alphaproteobacteria

Al1 *Sphingomonas*
Al2 *Bradyrhizobium*
Al3 *Sphingobium*
Al4 *Sinorhizobium*
Al5 *Celeribacter*
Al6 *Sulfitobacter*
Al7 *Halocynthiibacter*
Al8 *Brevundimonas*
Al9 *Caulobacter*
Al10 *Ruegeria*
Al11 *Nordella*
Al12 *Pseudolabrys*
Al13 *Agrobacterium*
Al14 *Ensifer*
Al15 *Paracoccus*
Al16 *Erythrobacter*
Al17 *Pseudohalocynthiibacter*
Al18 *Ancylobacter*
Al19 *Bartonella*
Al20 *Devosia*
Al21 *Methylosinus*
Al22 *Rhodobacteraceae*
Al23 *Sphingosinithalassobacter*
Al24 *Acetobacter*
Al25 *Boseongicola*
Al26 *Commensalibacter*
Al27 *Fluviibacterium*
Al28 *Granulibacter*
Al29 *Methylobacterium*
Al30 *Novosphingobium*
Al31 *Pelagibaca*
Al32 *Rhodovulum*
Al33 *Sphingopyxis*
Al34 *Beijerinckia*
Al35 *Labrenzia*
Al36 *Mesorhizobium*
Al37 *Rhodobacter*
Al38 *Starkeya*

Betaproteobacteria

Be1 *Mitsuaria*
Be2 *Janthinobacterium*
Be3 *Acidovorax*
Be4 *Methylibium*
Be5 *Lautropia*
Be6 *Rhodoferax*
Be7 *Snodgrassella*
Be8 *Delftia*
Be9 *Methylophilus*
Be10 *Neisseria*
Be11 *Paraburkholderia*
Be12 *Bordetella*
Be13 *Candidatus Nitrotoga*
Be14 *Cupriavidus*
Be15 *Herbaspirillum*
Be16 *Massilia*
Be17 *Ralstonia*
Be18 *Alicycliphilus*
Be19 *Betaproteobacteria*
Be20 *Dechloromonas*
Be21 *Rhizobacter*

delta/epsilon subdivisions

de1 *Corallococcus*
de2 *Labilithrix*
de3 *Pseudenhygromyxa*
de4 *Minicystis*
de5 *Chondromyces*
de6 *Desulfobacula*
de7 *Geobacter*
de8 *Pelobacter*

Gammaproteobacteria

Ga1 *Pseudomonas*
Ga2 *Escherichia*
Ga3 *Raoultella*
Ga4 *Acinetobacter*
Ga5 *Pseudoalteromonas*
Ga6 *Cellvibrio*
Ga7 *Gilliamella*
Ga8 *Xanthomonas*
Ga9 *Microbulbifer*
Ga10 *Alteromonas*
Ga11 *Kineobactrum*
Ga12 *Methylocaldum*
Ga13 *Haemophilus*
Ga14 *Stenotrophomonas*
Ga15 *Teredinibacter*
Ga16 *Vibrio*
Ga17 *Colwellia*
Ga18 *Dyella*
Ga19 *Salinimonas*
Ga20 *Spiribacter*
Ga21 *Thioalkalivibrio*
Ga22 *Azotobacter*
Ga23 *Candidatus Methylospira*
Ga24 *Granulosicoccus*
Ga25 *Halomonas*
Ga26 *Immundisolibacter*
Ga27 *Legionella*
Ga28 *Marichromatium*
Ga29 *Marinomonas*
Ga30 *Methylomonas*
Ga31 *Salinivibrio*
Ga32 *Serratia*
Ga33 *Marinobacter*

Oligoflexia

Ol1 *Halobacteriovorax*
Ol2 *Proteobacteria*

Terrabacteria group

Actinobacteria

Ac1 *Herbiconiux*
Ac2 *Streptomyces*
Ac3 *Bifidobacterium*
Ac4 *Leifsonia*
Ac5 *Pseudarthrobacter*
Ac6 *Corynebacterium*
Ac7 *Actinomyces*
Ac8 *Curtobacterium*
Ac9 *Mycolicibacterium*
Ac10 *Micromonospora*
Ac11 *Cellulomonas*
Ac12 *Microbacterium*
Ac13 *Nocardioides*
Ac14 *Schaalia*
Ac15 *Kineococcus*
Ac16 *Tetrasphaera*
Ac17 *Brachybacterium*
Ac18 *Rothia*
Ac19 *Candidatus Planktophila*
Ac20 *Nocardiopsis*
Ac21 *Rhodococcus*
Ac22 *Acidothermus*
Ac23 *Brevibacterium*
Ac24 *Candidatus Aquiluna*
Ac25 *Geodermatophilus*
Ac26 *Marmoricola*
Ac27 *Pauljensenia*
Ac28 *Planctinospora*
Ac29 *Verrucosispora*
Ac30 *Arthrobacter*
Ac31 *Ilumatobacter*
Ac32 *Mycobacterium*
Ac33 *Nonomuraea*

Chloroflexi

Ch1 *Caldilinea*
Ch2 *Chloroflexus*
Ch3 *Roseiflexus*

Cyanobacteria/Melainabacteria group

Cy1 *Synechococcus*
Cy2 *Chondrocystis*
Cy3 *Cyanobium*
Cy4 *Microcoleus*
Cy5 *Nostoc*
Cy6 *Acaryochloris*
Cy7 *Anabaena*
Cy8 *Chroococcidiopsis*
Cy9 *Leptolyngbya*
Cy10 *Calothrix*
Cy11 *Oscillatoria*
Cy12 *Thermoleptolyngbya*

Deinococcus-Thermus

Dt1 *Meiothermus*

Firmicutes

Fi1 *Streptococcus*
Fi2 *Bacillus*
Fi3 *Clostridium*
Fi4 *Paenibacillus*
Fi5 *Lactobacillus*
Fi6 *Enterococcus*
Fi7 *Ruminococcus*
Fi8 *Cohnella*
Fi9 *Lachnospiraceae*
Fi10 *[Ruminococcus]*
Fi11 *Herbinix*
Fi12 *Roseburia*
Fi13 *[Clostridium]*
Fi14 *Anoxybacillus*
Fi15 *Blautia*
Fi16 *Lactococcus*
Fi17 *[Eubacterium]*
Fi18 *Intestinimonas*
Fi19 *Lachnoanaerobaculum*
Fi20 *Paeniclostridium*
Fi21 *Staphylococcus*
Fi22 *Desulfitobacterium*
Fi23 *Exiguobacterium*
Fi24 *Pseudobutyrivibrio*
Fi25 *Romboutsia*
Fi26 *Selenomonas*
Fi27 *Turicibacter*

Thermotogae

Thermotogae

Th1 *Fervidobacterium*
Th2 *Pseudothermotoga*

Fusobacteria

Fusobacteriia

Fu1 *Leptotrichia*
Fu2 *Fusobacterium*

Spirochaetes

Spirochaetia

Sp1 *Spirochaeta*

Aquificae

Aquificae

Aq1 *Thermocrinis*

FCB group

Bacteroidetes/Chlorobi group

Ba1 *Bacteroides*
Ba2 *Paraflavitalea*
Ba3 *Flavobacterium*
Ba4 *Pedobacter*
Ba5 *Polaribacter*
Ba6 *Flavobacteriaceae*
Ba7 *Prevotella*
Ba8 *Maribacter*
Ba9 *Niastella*
Ba10 *Aquimarina*
Ba11 *Chitinophaga*
Ba12 *Psychroserpens*
Ba13 *Hymenobacter*
Ba14 *Mucilaginibacter*
Ba15 *Capnocytophaga*
Ba16 *Niabella*
Ba17 *Petrimonas*
Ba18 *Prolixibacteraceae*
Ba19 *Urechidicola*
Ba20 *Tenacibaculum*
Ba21 *Flavivirga*
Ba22 *Dokdonia*
Ba23 *Flavisolibacter*
Ba24 *Cellulophaga*
Ba25 *Cyclobacterium*
Ba26 *Formosa*
Ba27 *Cytophaga*
Ba28 *Euzebyella*
Ba29 *Pseudobacter*
Ba30 *Wenyingzhuangia*
Ba31 *Zobellia*
Ba32 *Alistipes*
Ba33 *Bernardetia*
Ba34 *Oceanihabitans*
Ba35 *Pontibacter*
Ba36 *Sphingobacterium*
Ba37 *Chryseobacterium*
Ba38 *Cloacibacterium*
Ba39 *Draconibacterium*
Ba40 *Emticicia*
Ba41 *Filimonas*
Ba42 *Kordia*
Ba43 *Labilibaculum*
Ba44 *Mariniflexile*
Ba45 *Muriicola*
Ba46 *Odoribacter*
Ba47 *Tannerella*
Ba48 *Winogradskyella*
Ba49 *Algoriphagus*
Ba50 *Arcticibacterium*
Ba51 *Gramella*
Ba52 *Ignavibacteriae*
Ba53 *Pseudarcicella*
Ba54 *Rimella*

Gemmatimonadetes

Ge1 *Gemmatimonas*

PVC group

Planctomycetes

Pl1 *Mariniblastus*
Pl2 *Roseimaritima*
Pl3 *Planctomycetes*
Pl4 *Gemmata*
Pl5 *Bythopirellula*
Pl6 *Isosphaera*
Pl7 *Limnoglobus*

Verrucomicrobia

Ve1 *Coraliomargarita*
Ve2 *Opitutus*
Ve3 *Verrucomicrobia*
Ve4 *Lacunisphaera*
Ve5 *Akkermansia*

Acidobacteria

Acidobacteriia

Ad1 *Candidatus Solibacter*
Ad2 *Acidisarcina*
Ad3 *Edaphobacter*
Ad4 *Granulicella*

Blastocatellia

Bl1 *Chloracidobacterium*

Vicinamibacteria

Vi1 *Luteitalea*

Euryarchaeota

Stenosarchaea group

St1 *Halorubrum*
St2 *Halorhabdus*
St3 *Haloarcula*

Nitrospirae

Nitrospira

Ni1 *Nitrospira*

Figure S9. Specific information on the most common prokaryotic genera. Based on NCBI Taxonomy statistics, different colors represent prokaryotic genera from different taxonomic groups, and subgroups are shown in bold. The genus codes correspond to those shown in Figure 7, Figure S8, and Figure S12.

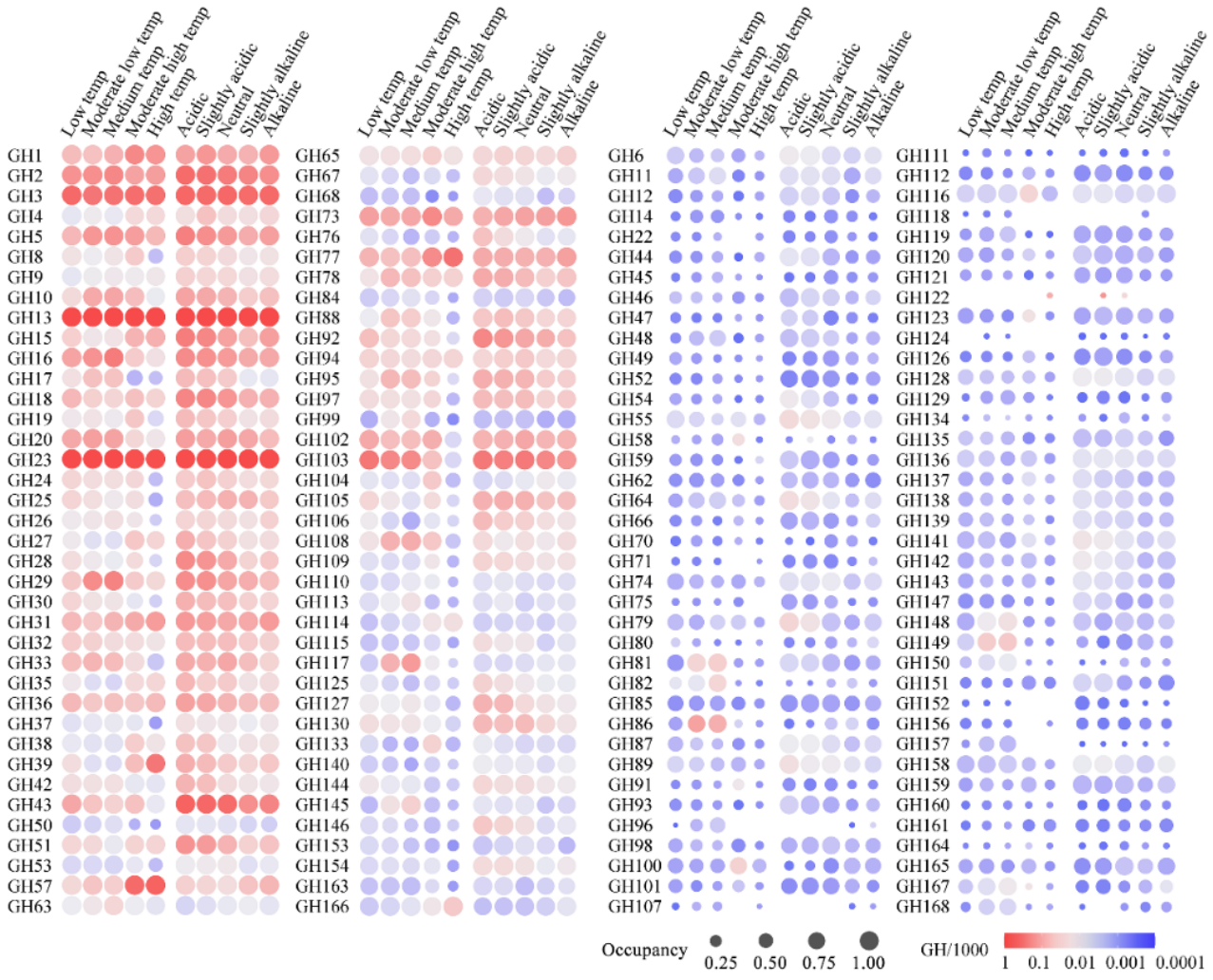

Figure S10. Environmental temperature and pH affect the occupancy and abundance of GH families in prokaryotic communities. Statistical analyses were carried out based on 2381 EMP samples with recorded temperature information and 1183 samples with recorded pH values. The EMP samples were categorized into 5 groups according to environmental temperature and pH: low temp (≤ 10°C), moderate low temp (> 10°C and ≤ 20°C), medium temp (> 20°C and < 30°C), moderate high temp (≥ 30°C and < 45°C), and high temp (≥ 45°C); acidic (≤ 5), slightly acidic (> 5 and ≤ 6.5), neutral (> 6.5 and < 7.5), slightly alkaline (≥ 7.5 and < 9) and alkaline (≥ 9). The occupancy of each GH family in samples within different temperature or pH groups is represented by the circle size. Colors represent the median abundance of the GH family across all EMP samples where the family was detected within the same group, with values shown as GH/1000, increasing from blue to red.

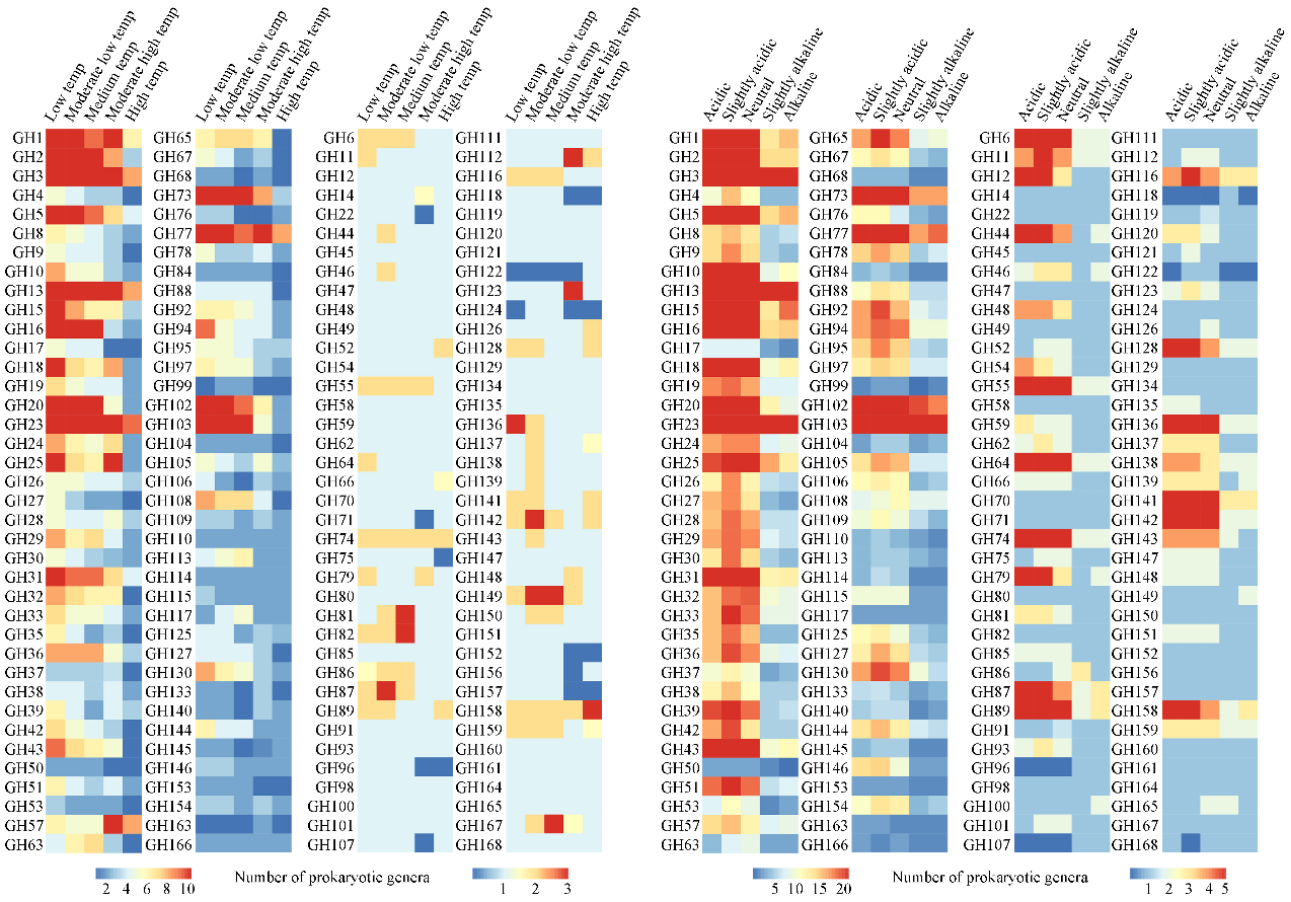

Figure S11. Diverse taxonomic sources of GH families in prokaryotic communities under different environmental temperature and pH conditions. The EMP samples were categorized into 5 groups according to environmental temperature and pH. The colors in the figure indicate the median number of source prokaryotic genera of GH families in all EMP samples from the same environment, with increasing values from blue to red.

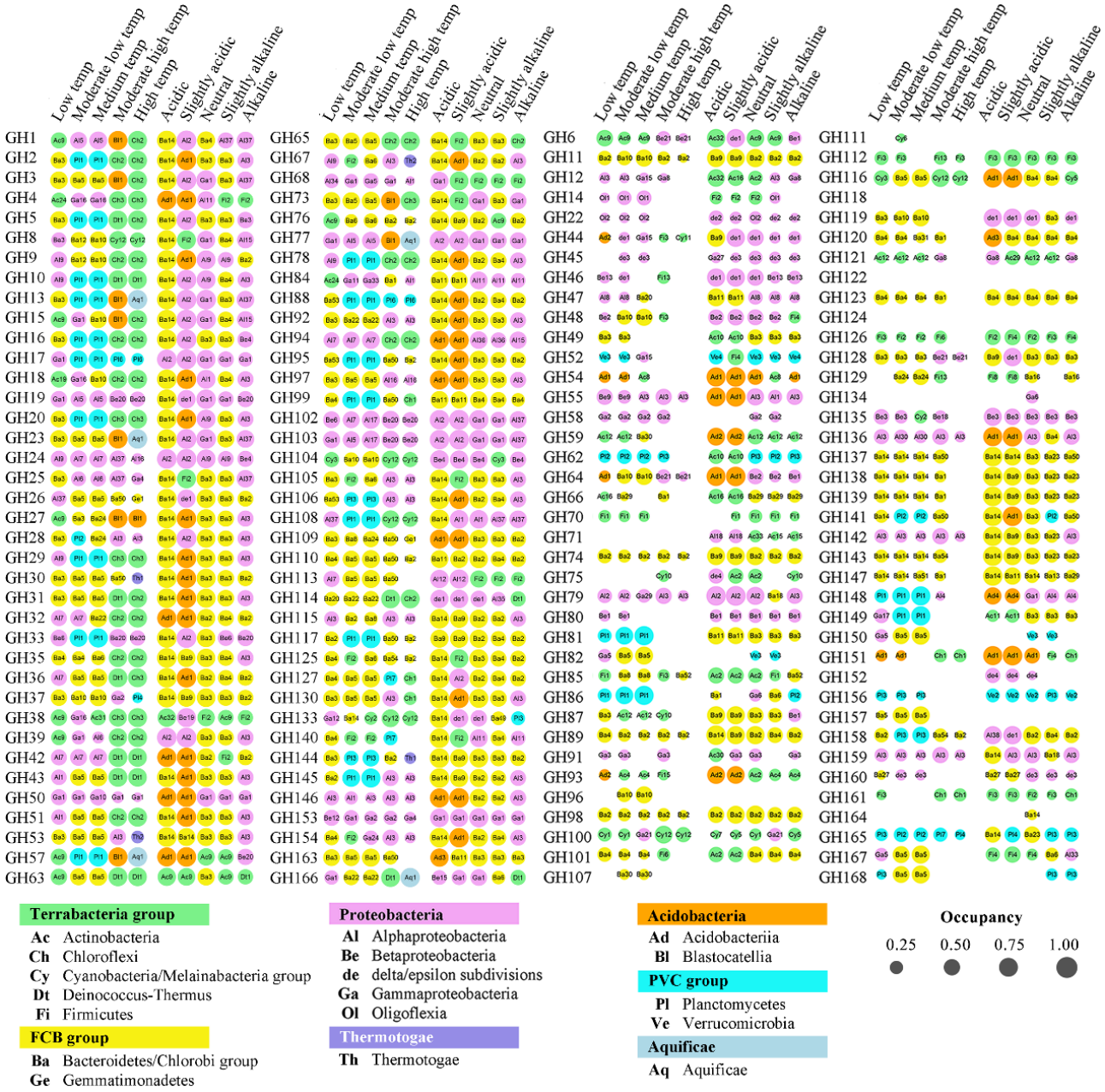

Figure S12. Most common prokaryotic genera of GH families in various environments. The most common prokaryotic genus was defined as the genus that appeared in the most samples and had an occupancy of at least 0.05 in all samples. The EMP samples were divided into 5 groups according to environmental temperatures and pH. Different colors represent different taxonomic groups. The circle size indicates the occupancy of the genus in all samples from a single environment. Specific information on the prokaryotic genera can be found in Figure S9.

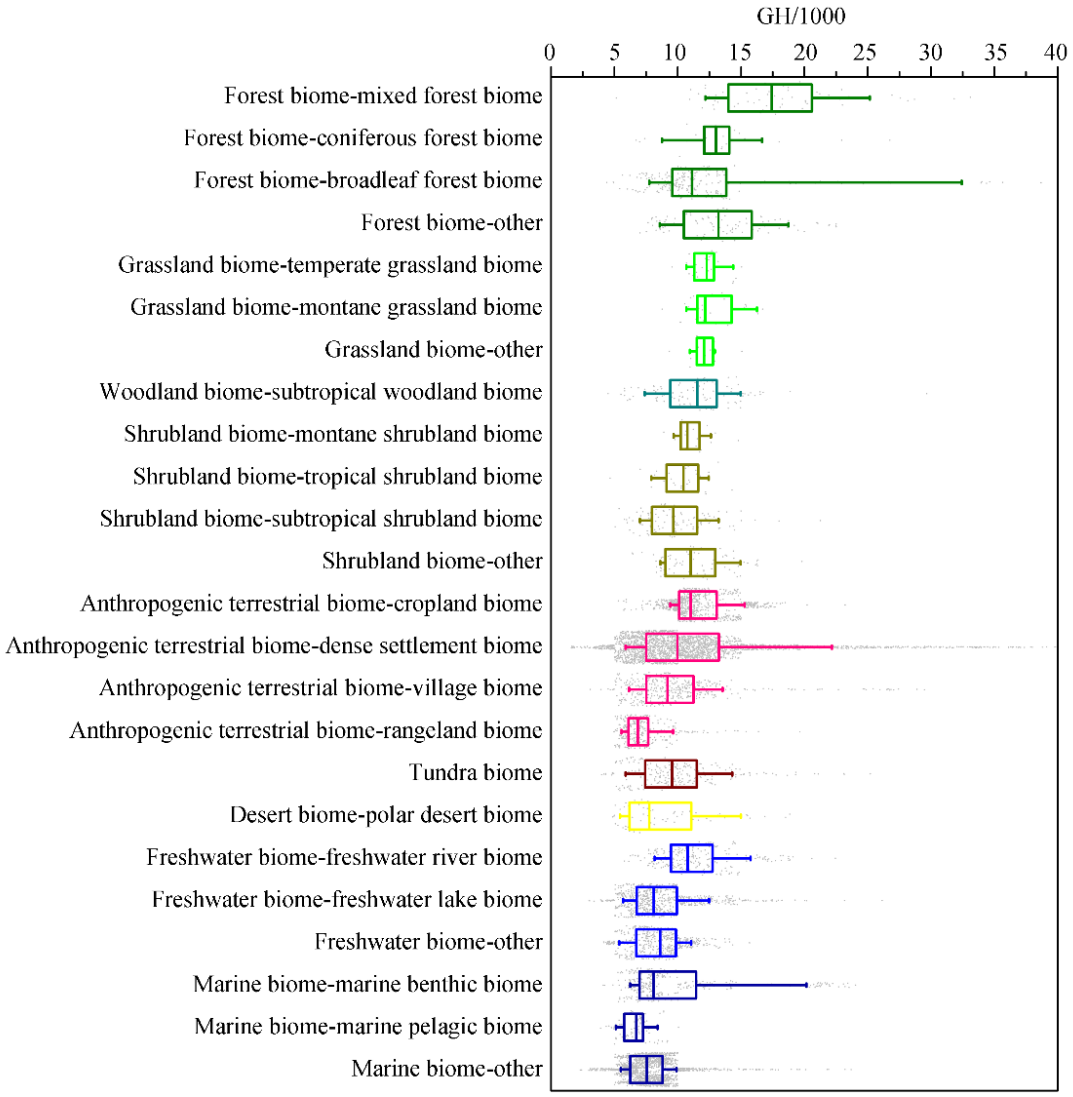

Figure S13. Variations in the abundance of GHs encoded by prokaryotic communities across different ENVO environments. The GH abundance is expressed as the GH/1000 value, i.e., the number of GH genes per thousand prokaryotic genes in a sample. Each grey point represents an EMP sample. For the boxplots, the middle line indicates the median, the box represents the 25th-75th percentile, and the whisker indicates the 10th-90th percentile of observations.

Data S1. Data on GH abundance in EMP samples.

Data S2. Abundance of 152 GH families in EMP samples.

Data S3. Correlations of different GH families in various environments.

Data S4. Information on taxonomic sources of GH families in prokaryotic communities.