Comparative genomic analysis reveals metabolic flexibility of Woesearchaeota

Wen-Cong Huang1,7, Yang Liu1,7, Xinxu Zhang1, Cui-Jing Zhang1, Dayu Zou1,2, Shiling Zheng3, Wei Xu4, Zhuhua Luo4,5, Fanghua Liu3,6 & Meng Li1

The archaeal phylum Woesearchaeota, within the DPANN superphylum, includes phylogenetically diverse microorganisms that inhabit various environments. Their biology is poorly understood due to the lack of cultured isolates. Here, we analyze datasets of Woesearchaeota 16S rRNA gene sequences and metagenome-assembled genomes to infer global distribution patterns, ecological preferences and metabolic capabilities. Phylogenomic analyses indicate that the phylum can be classified into ten subgroups, termed A–J. While a symbiotic lifestyle is predicted for most, some members of subgroup J might be host-independent. The genomes of several Woesearchaeota, including subgroup J, encode putative [FeFe] hydrogenases (known to be important for fermentation in other organisms), suggesting that these archaea might be anaerobic fermentative heterotrophs.
The known scope of archaeal diversity has noticeably expanded in recent years after the discovery of novel lineages, enabled by the development of bioinformatics methodologies, the continually generated sequencing data, and cultivation. The expansion of the archaeal tree of life has changed the picture of the ecological and evolutionary importance of archaean. For example, we now know that Thaumarchaeota, ammonia oxidizers detected in aquatic and terrestrial environments, participate in the global nitrogen cycle; that some members of Bathyarchaeota, archaean from a non-euryarchaean lineage, exhibit methanogenic characteristics; that the genomic content of Asgard archaean sheds light on the origin of eukaryotes; and that the DPANN (Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, Nanohaloarchaeota) archaean, a proposed monophyletic group of enigmatic archaean, are typically small cells harboring reduced genomes with a limited metabolic repertoire. Notably, few DPANN members have been successfully enriched in co-culture and reportedly rely upon their hosts to proliferate. For instance, Nanoarchaeum equitans is an obligate endosymbiont of the host, Ignicoccus hospitalis, which provides growth factors, lipids, amino acids, and probably ATP, to N. equitans. However, in a recent study, researchers generated single amplified genomes affiliated with DPANN lineage using fluorescence-activated cell sorting. They found minor heterogeneous DNA sources from potential hosts in the genomes, which raised the possibility that most DPANN archaean might not lead a symbiotic lifestyle in subsurface environments.

The Woesearchaeota phylum (formerly Euryarchaeota DHVEG-6) was proposed within the DPANN superphylum in 2015. Woesearchaeota are ubiquitous residents of various environments (e.g., groundwater, soil, marine sediments, hydrothermal vents, and freshwater sediments) where they may shape the surroundings and impact global biogeochemical cycles, interacting or not with other archaean. For instance, based on genome-resolved metagenomic analysis, Castelle et al. proposed that Woesearchaeota AR20 may lead a symbiotic lifestyle, and are involved in anaerobic carbon and hydrogen cycles. Later, Castelle and Banfield reported that some Woesearchaeota may employ the bacterial methylerythritol phosphate (MEP) pathway transferred from Firmicutes to synthesize isopentenyl pyrophosphate and dimethylallyl diphosphate precursors for cell membrane assembly. Investigation of approximately 1000 genomes reconstructed from several metagenomics-based studies revealed that a Woesearchaeota genome, with features suggesting a fermentation-based lifestyle, encodes a near-complete glycolysis pathway, components of a potential metal-reducing respiratory pathway involved in iron metabolism, and a cytoplasmic MvhD-HdrABC complex (F420-reducing quinol oxidase) functioning in the final step of methanogenic pathways. Collectively, these studies suggest intriguing metabolic diversity among the Woesearchaeota.

Liu et al. reported a co-occurrence pattern between the operational taxonomic units (OTUs) affiliated with Woesearchaeota 16S rRNA genes, and those of Methanomicrobia and Methanobacteria, which indicated possible interactions between members of these groups. Based on these findings, the authors proposed that Woesearchaeota probably provide substrates for H2/CO2-utilizing methanogens and acetate-utilizing methanogens in return for amino acids and other compounds, to compensate for their own metabolic deficiencies. Meanwhile, a positive correlation of the Woesearchaeota relative abundance and bacterial community was reported in a study based on 16S rRNA gene amplicon sequences, suggesting possible interactions between these microbes.

Despite the above-mentioned glimpses into the metabolic potential of Woesearchaeota archaean and interactions with other organisms, their ecological patterns, metabolic diversity, and evolutionary history remain unclear. To address that, here, we retrieved the Woesearchaeota 16S rRNA gene sequences from the Earth Microbiome Project (EMP) datasets. We then analyzed genomes of Woesearchaeota from different environments, including 152 metagenome-assembled genomes (MAGs), with 49 MAGs reported for the first time in the current study. Our analyses help us to understand the global distribution patterns of Woesearchaeota in different biotopes, and shed new light on the metabolism and evolutionary history of Woesearchaeota diversification.

Results

Ecological patterns of Woesearchaeota distribution. To survey the global distribution and abundance of Woesearchaeota, we selected 2,163 16S rRNA gene libraries from EMP amplicon datasets, using the criteria: minimum sequencing depth of 30,000 and minimum relative abundance of Woesearchaeota in the libraries of 0.1%. These libraries were sampled from 11 distinct biotope types worldwide, with the relative Woesearchaeota abundance ranging from 0.1 to 3.9% (Fig. 1a). The relative abundance of Woesearchaeota in saltmarshes was significantly higher than that in other biotopes (p < 0.01, post hoc test after ANOVA; Supplementary Fig. 1). The alpha diversity (PD_whole_tree value) of Woesearchaeota was significantly higher in saltmarshes, freshwater, and mangrove than in other biotopes, with the lowest value in the sand samples (Fig. 1b). The t-distributed stochastic neighbor embedding (t-SNE) analysis of the similarity of Woesearchaeota community matrix (the unweighted UniFrac metric) revealed that the saline biotopes were separated from plant and non-saline biotopes (Fig. 1c; F = 173.78, p = 0.001, R² = 0.138, 999-permutations PERMANOVA test). Further, communities in different biotopes were significantly different from each other (Fig. 1d; F = 103.72, p = 0.001, R² = 0.325). Finally, by testing correlations between the available physicochemical parameters and beta diversity of Woesearchaeota communities using the Mantel test, we discovered that salinity levels were significantly correlated with Woesearchaeota community composition (r = 0.295, p = 0.001, n = 167) (Supplementary Table 1). These observations indicate that salinity influences Woesearchaeota community and distribution.

Woesearchaeota genome dataset. We reconstructed 49 metagenome-assembled genomes (MAGs) affiliated with Woesearchaeota. These MAGs were recovered from a wide range of environments, including various water depths of the Yap trench, intertidal mangrove sediments of Mai Po Nature Reserve (Hong Kong), Futian Mangrove Nature Reserve (Shenzhen), seagrass sediments of Swan Lake Nature Reserve (Rongcheng) and sediments of Fuilong River estuary (Fujian). The MAGs reconstructed here were estimated to be 77.7% (median) complete and 1.5% (median) contaminated (Supplementary Fig. 2). These MAGs were combined with 103 publicly available genomes (estimated completeness ≥50% and contamination ≤10%), resulting in a dataset of 152 MAGs (Supplementary Data 1).

Distinct Woesearchaeota subgroups and their metabolic potential. Maximum-likelihood phylogenetic tree based on the carefully selected single-copy orthologs (see “Methods”), with most nodes with ultrafast bootstrap (UFBOOT) values ≥95% and Shimodaira and Hasegawa-like approximate likelihood ratio test (SH-aLRT) values ≥80%, divided Woesearchaeota into ten phylogenetically distinct subgroups (subgroups A–J; Fig. 2). Most subgroups are monophyletic in the phylogenetic trees based on the 16S rRNA gene and 15 ribosomal proteins (see Supplementary Notes). In six subgroups (C–J), the average estimated
genome size exceeded 1 Mbp. In contrast, members of subgroup A, deeply rooted within Woese archaeota in the phylogenetic tree, had the smallest median estimated genomic size (0.98 Mbp, Fig. 2; Supplementary Fig. 3). By contrast, the relationship between the phylogenetic group and genomic GC content was not pronounced. The genomic GC content in several subgroups greatly varied internally (Fig. 2; Supplementary Fig. 3). For example, estimated genome sizes of archaea in subgroup G, range from 0.88 to 1.36 Mbp (average 1.31 ± 0.33 Mbp), with the GC content ranging from 30 to 59%.

The 16S rRNA gene sequences identified in the MAGs allow us to link the subgroups defined here with previously described clusters (Fig. 2; Supplementary Fig. 4; Supplementary Table 2). Subgroup A, C, E, H, I, and J contain sequence representatives in
Fig. 1 Global distribution of Woesearchaeota. a Global distribution of Woesearchaeota with at least 0.1% relative abundance, based on 2,163 16S rRNA gene amplicon datasets. Note that some libraries have the same coordinate and thus are overlapping. The world map was generated using R package “maps” v3.3.0, in R v3.6.0[97]. b Box plot of phylogenetic diversity whole-tree index of Woesearchaeota in 11 biotopes. The median is shown in a thick black bar. The first and third quartile are shown respectively in the lower and upper bound of the box. The whiskers correspond to the 1.5 interquartile range from the bounds. The outlier is marked by dots. Data points from samples with sizes of less than or equal to ten are shown. The number of replicates used for different biotopes are as follows: freshwater: 629, cropland: 590, freshwater sediment: 365, marine sediment: 293, sand: 174, mangrove: 46, soil: 30, hydrothermal: 15, saltmarshes: 10, saline environment: 8, marine water: 3. c Beta diversity plots of Woesearchaeota community based on 2163 16S rRNA gene amplicon datasets. Beta diversity was calculated by the unweighted UniFrac metric method and tested with a 999-permutation PERMANOVA test. The dots are colored according to the biotope. Source data are provided as a Source Data file.

Fig. 2 Phylogenetic tree of Woesearchaeota. The maximum likelihood tree was generated based on the concatenated 50% top-ranked orthologs (n = 38, 10,896 sites) with IQ-TREE (v2.0.7) using LG + F + R + G60 model. The tree includes 152 Woesearchaeota genomes and is rooted by 30 Pacearchaeota genomes. 49 MAGs generated in this study are marked by the star. Black dots denote nodes with UFBOOT support values ≥95% and SH-aLRT values ≥80%. Scale bar shows the average number of substitutions per site. The minimum number of orthologs in Woesearchaeota MAGs is 21. J072 and Yap5000.bin19 were ungrouped as they did not form well-supported clade (UFBOOT < 95% and SH-aLRT <80%). Source data are provided as a Source Data file.
Woesearchaeota (Woese-3, Woese-4, Woese-14b, Woese-14a, Woese-24, and Woese-21a, respectively. In addition, subgroup G has sequence representatives in a monophyletic clade in the 16S rRNA gene tree, including Woese-8, Woese-10, Woese-9, Woese-6, Woese-18, and Woese-20, indicating a large diversity in subgroup G that remains unexplored. Next, we used the linked sequence clusters to probe the approximate distribution of subgroups in the 2163 16S RNA gene libraries. The analysis showed that subgroup I, J, and G were detected among all biotopes investigated (with >73% occurrence in the studied libraries; Supplementary Data 2), indicating high ecological adaptability of these lineages (Supplementary Fig. 5).

We next inferred the metabolic potential of 152 genomes (≥50% completeness and ≤10% contamination) representing the 10 Woesearchaeota subgroups. We confirmed some previously described metabolic features of Woesearchaeota16–18, namely: (1) while all genomes appear to lack complete electron-transport chains, including NADH dehydrogenase and complexes II–IV of the oxidative phosphorylation chain, some encode complex V (V-type ATPase) (Fig. 3a; Supplementary Data 3); (2) most genomes encode few components of tricarboxylic acid (TCA) cycle; (3) some genomes (YT1_182 and Yap2000.bin4.8) are missing complete glycolysis (Fig. 3b; Supplementary Data 4). Notably, two genomes (YT1_182 and Yap2000.bin4.8) from subgroup J exhibited the potential for complete glycolysis (Fig. 3b; Supplementary Data 4).

We then used the Carbohydrate-active Enzymes (CAZY) database to assess Woesearchaeota capacity to degrade complex carbon sources. The analysis revealed that 78 genomes, including YT1_182 and Yap2000.bin4.8 from subgroup J (Fig. 3b), encode at least one copy of alpha-amylase (GH57), supporting potential for starch degradation among Woesearchaeota. Most Woesearchaeota members (mainly from subgroups B, C, D, H, I, and J) appear to use pyruvate-2-oxidid-ferreredoxin oxidoreductase (porA/porB) to decarboxylate pyruvate for acetyl-CoA synthesis. By contrast, some members (mainly from subgroups A and J) appear to employ pyruvate dehydrogenase, which is generally found among aerobes (mostly Eukarya and Bacteria). Genes for phosphate acetyltransferase (pta) and acetate kinase (ackA), which catalyze the synthesis of acetate (the Pta-AckA pathway), are present in many subgroups, except for subgroups A, B, C, and I. With the presence of acetyl-CoA synthetase (acsA), pta, and ackA, MAGs such as YT1_182 and Yap2000.bin4.8 may be able to regulate electron flow by consuming or generating acetate (Fig. 3b). However, further fermentation into ethanol might be disabled because the gene for aldehyde dehydrogenase is apparently missing (Fig. 3b). In addition, other than alcohol and acetate, many MAGs were also predicted to encode lactate dehydrogenase (Supplementary Data 3). Apart from glucose, using extracellular DNA as a growth substrate is also possible for Woesearchaeota, as intermediates of nucleotide degradation, such as glycerate-3-phosphate, could be channeled into the second half glycolytic pathway. For example, some members from subgroups A, G, H, and J encode a complete nucleoside degradation pathway, including AMP phosphorylase (deoA), ribose-1,5-bisphosphate isomerase (R15Pi), and ribulose 1,5-bisphosphate carboxylase (rbCL), in which the main form of rbCL among Woesearchaeota represents class III-3b (Supplementary Fig. 6). Many pathways identified herein in Woesearchaeota require the participation of cofactors, such as NAD+/NADH, required for glycolysis, and CoA, needed in the Ack-Pta pathway. Nonetheless, the biosynthetic pathways for cofactors (e.g., cobalamin, CoA, and thiamine) were rarely complete in these genomes (Supplementary Data 3). Collectively, these observations indicate that limited metabolic potential in carbohydrate metabolism is common among Woesearchaeota.

H₂ metabolism might play roles in the generation of proton-motive forces in subgroup J. Hydrogenases, the key enzymes in hydrogen metabolism, can employ H₂ as reducing power or H⁺ as oxidants to dissipate excessive reductants in cells. They were commonly found among different Woesearchaeota subgroups and appeared to be important for Woesearchaeota. In the placement of a well-defined electron transport chain (Complex I –IV), hydrogenase could re-oxidize NADH, NADPH and reduced ferredoxin by reducing H₂ to molecular hydrogen. Specifically, we identified putative cytosolic [NiFe] hydrogenases (group 3B), which evolve H₂ by coupling the oxidation of NADPH and may be reversible in the genomes of subgroup B, D, I, and G (Supplementary Fig. 7). The hydrogenase is also encoded in genomes of multiple diverse bacterial and archaeal phyla (e.g., Proteobacteria and Euryarchaeota)21. Remarkably, [FeFe] hydrogenases, the most efficient enzymes for catalytic hydrogen turnover that are typically found in bacteria and eukaryotes21,22, were identified in Woesearchaeota, i.e., subgroup B, E, G, H, and J. Phylogenetic analysis of their catalytic domain indicated that they belong to group A [FeFe] hydrogenase (Fig. 4a). Further examination into the genetic organization of group A [FeFe] hydrogenases encoded in Woesearchaeota showed that most of them are trimeric, containing a catalytic subunit (H-cluster), a nuoF-like, and a nuoE-like gene (Supplementary Fig. 8). The nuoF-like gene encodes a 4Fe–4S cluster binding domain, an electron carrier flavin mononucleotide (FMN)-binding domain, and a soluble ligand-binding domain. Therefore, the trimeric [FeFe] hydrogenases in Woesearchaeota belong to [FeFe] hydrogenase group A3,20,21, which can reversibly bifurcate electrons from H₂ to ferredoxin and NAD+. Cysteine residues of Woesearchaeota A3 [FeFe] hydrogenases that bind the metal ions are conserved, suggesting the presence of active sites (Fig. 4c). Phylogenetically, the catalytic subunit of these hydrogenases formed a monophyletic group with some other DPANN archaea (i.e., Micrarchaeota, Aerinigmarchaeota, and Nanothaloaarchaeota) and are closely related to the hydrogenases in Thermotogae and Bacteroidetes (Supplementary Fig. 9).

Additionally, we identified a potential membrane-bound complex named _Rhodobacter_ nitrogen fixation (Rnf) electron transport complex in four members of subgroup J (i.e., YT1_182, Yap2000.bin4.8, YT1_767, and MP5_5_87). This complex could serve as a respiratory enzyme that couples the oxidation of reduced ferredoxin to reduce NAD⁺. The free energy of this exergonic reaction could be used to translocate sodium ion or proton out of cells, thereby generating a potential gradient. The putative V-type ATP synthase could then harness the gradient to conserve energy by making ATP. The Rnf complex operon in the four genomes consisted of six subunits (rnfCDEGAB), resembling the organization of the operon in _Acetobacterium woodii_.23 Phylogenetic analyses of these six genes indicated that they might have been transferred from bacteria (Supplementary Figs. 10–15). The electrochemical gradient could also be generated by a putative pyrophosphate-energized sodium pump (hppA), which utilizes the energy of pyrophosphate hydrolysis to move Na⁺/H⁺ across the membrane (Fig. 3b; Supplementary Data 4).

The four MAGs in subgroup J also have genes encoding a putative butyryl-CoA dehydrogenase (Bcd) adjacent to genes encoding electron transfer flavoprotein subunits EtfA and EtfB. These genes may constitute an EtfAB/Bcd complex, which catalyzes short-chain acyl CoA transformation to short-chain trans-2,3-dehydroacyl-CoA, with NADH as the electron donor, ferredoxin as the negative redox potential acceptor, and short-chain acyl CoA as the positive potential acceptor21. Phylogenetic analysis indicated Bcd genes in the four MAGs are most related to Firmicutes (Supplementary Fig. 16). However, the absence of the
Fig. 3 Metabolic comparison of Woesearchaeota subgroups. a Occurrence of a gene of interest in different subgroups of Woesearchaeota. The occurrence was calculated in percent across the total number of genomes included in each subgroup based on the presence/absence table. Raw data are available in Supplementary Data 3. b Reconstructed metabolic pathways of two MAGs from subgroup J. YT1_182 (estimated to be 94.61% complete and 1.96% contaminated) and Yap2000.bin4.8 (estimated to be 94.61% complete and 3.92% contaminated). A detailed list of genes encoded by these two MAGs can be found in Supplementary Data 4. Source data are provided as a Source Data file.
**Fig. 4 Diversity of [FeFe] hydrogenases in Woesearchaeota.**

- **a** Phylogeny analysis of the catalytic domain of [FeFe] hydrogenase sequences (n = 856, trimmed alignment length = 356) based on LG + G + C10 model using IQ-Tree (v1.6.8). Black dots denote nodes with UFBOOT values ≥95%. Scale bar shows the average number of substitutions per site. Woesearchaeota sequences are colored in red embedded in the group A sequences.
- **b** Probability plot of the occurrence of each amino acid of the P1, P2, and P3 motif of [FeFe] hydrogenase group A3. Amino acids are colored based on their hydrophobicity score. Gray box is added to distinguish the letters V and I from Y. Positions of alignment are indicated below the amino acids.
- **c** Conservation of the P1, P2, and P3 motif of selected amino acid sequences in comparison to *Thermotoga maritima* (WP_004081677.1), *Desulfovibrio fructosovorans* (WP_005990603.1), and *Clostridium pasteurianum* (WP_023973103.1). The difference to the reference sequences is colored by their hydrophobicity score and conserved site is colored in gray. The position of amino acids flanking the P1, P2, and P3 motif in the alignment is indicated at the bottom of the plots. Source data are provided as a Source Data file.

FAD-binding domain in EtfAB renders electron bifurcation unlikely in this complex24.

To summarize, our analyses indicate that hydrogenases may play important roles in the metabolism of Woesearchaeota. In subgroup J, like YT1_182, they might employ hydrogenase and Rnf complex and other electron bifurcation complex to balance the reducing pool (NADH and ferredoxin), especially reducing equivalents from the glycolytic pathway. Energy could be conserved by substrate-level phosphorylation and coupling of the Rnf complex with ATP synthase. Therefore, based on the metabolic analysis, we predict that members of subgroup J may have a heterotrophic lifestyle with fermentative metabolism. ** Biosynthetic capacity differentiates subgroup J from other Woesearchaeota.** Compared with other Woesearchaeota subgroups, subgroup D and J Woesearchaeota appeared to harbor expanded gene inventories related to the biosynthesis of amino acids, nucleotides, and isoprenoids (Fig. 3a). In terms of amino acid biosynthesis, members of subgroups D and J harbor various genes for the de novo synthesis of arginine, serine, proline, leucine, and threonine (Fig. 3a; Supplementary Data 3). Members of the subgroup J were also found to encode genes required to synthesize tryptophan (Fig. 3a). Further, some members of the subgroup J (e.g., YT1_182 and Yap00000048.8) encode genes necessary for the de novo synthesis of proline, valine, and leucine.
Still, the pathway for the synthesis of complex amino acids, such as aromatic amino acids, is incomplete (Fig. 3b). Some amino acids like arginine might be translocated into cells through an amino acid/polyamine transporter (Fig. 3b).

The difference in the occurrence of genes for the purine and pyrimidine biosynthesis between subgroups was not as striking as that in the occurrence of genes for amino acid biosynthesis (Fig. 3a). Nonetheless, of note, genomes from subgroups D and J harbor many genes for nucleotide biosynthesis. For example, we recovered nearly complete pathways for the purine and pyrimidine synthesis from YT1_182 and Yap2000.bin4.8 (Fig. 3b; Supplementary Data 4).

Furthermore, only the bacterial MEP pathway for the synthesis of isoprenoid precursors (isopentenyl diphosphate and dimethylallyl diphosphate) existed in Woesearchaeota genomes. More interestingly, the pathway appeared to be confined to subgroups F and J, and this pathway was present in nearly half of the members of subgroup J, including YT1_182 and Yap2000.bin4.8 (Fig. 3). Most members of subgroup J have fused genes encoding 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase (ispD) and 2-C-methyl-D-erythritol 2,4-cyclophosphate synthase (ispF) domains. Phylogenetic analyses of genes in the pathway indicated that they are closely related to sequences of different bacterial organisms (Supplementary Figs. 17–22). For example, ispD/F is most related to Candidatus Neelsonbacteria (Supplementary Fig. 19), while 1-deoxy-D-xylulose 5-phosphate reductoisomerase (dxr) and 4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase (ispG) of subgroup J are most related to Firmicutes (Supplementary Figs. 18, 21). Nevertheless, these archaea might be unable to synthesize membrane lipids because of the absence of two crucial genes, namely geranylgeranylglycerol phosphate synthase and digeranylgeranylglycerol phosphate synthase.

**Evolutionary history of subgroup J.** Given members of subgroup J have more genes involved in isoprenoid and amino-acid biosynthesis and some may have a distinct lifestyle (i.e., heterotrophic fermentative metabolism) from other Woesearchaeota, we sought to reconstruct the gene family evolutionary history of subgroup J, hoping to gain insights into the lifestyle transformation. As data completeness could affect the accuracy of ancestral reconstruction, we selected 47 Woesearchaeota genomes for further analysis according to the criterion (>79% completeness and <5% contamination). This criterion was set according to the quality metric of the first circular complete MAG reconstructed in Castelle et al.6, Woesearchaeota AR20, whose completeness was estimated to be 79.17% in this study. On the other hand, we opted to use the gene-tree-aware approach implemented by ALE which incorporates a probabilistic method to account for the estimated missing fraction of genomes. Taxonomic sampling also affects the accuracy of ancestral reconstruction. Here, 47 representative MAGs were selected according to their quality metrics. They covered nine subgroups except for subgroup F. We then clustered 65,396 genes from 47 Woesearchaeota genomes and 5 Pacearchaeota genomes into 4,562 gene families. Gene tree samples were inferred for the 2,320 gene families with ≥4 sequences. They were probabilistically reconciled with the maximum likelihood tree inferred for the 52 genomes. Phylogenetic placement of 47 Woesearchaeota genomes in the tree was in good agreement with their positions in the tree used for subgroup assignment. It was also well supported, with most placements achieving maximal support (Supplementary Fig. 23). We chose to report ancestral events and gene copies numbers with a minimum threshold of 0.3. This threshold is relaxed and believed to be able to detect ancestral events blurred by noises from alignment, tree reconstruction, and reconciliation25. The ancestral reconstruction result is summarized in Fig. 5a.

The analysis predicted that a recent major expansion of subgroup J genomes occurred recently, multiple times (node 87 to node 82, node 82 to node 79, node 82 to node 73, and node 79 to node 74) (Fig. 5a). Nearly half of the genes predicted to gain at nodes 82 and 79 are related to metabolism (Fig. 5b). Further, at node 82, a large proportion of acquired genes was related to amino acid transport and metabolism (approximately 16%), and nucleotide transport and metabolism (approximately 10%). According to the minimum frequency threshold, the predicted gene content gains mainly covered 110 origins events (de novo genes or transfers from unsampled genomes, 43% of gains) and 83 intra-transfers events (transfers from sampled genomes, 32% of gains). At node 79, the acquired genes were mainly related to amino acid transport and metabolism (approximately 11%), and energy production and conversion (approximately 13%) and mainly occurred by 38 origins (15% of the gains) and 84 intra-transfers (33% of the gains). Interestingly, genes involved in isoprenoid synthesis (e.g., dxs, ispE, ispDF, ispG, and ispH), amino acid biosynthesis (e.g., leuA, hisG, and trpG), and nucleotide synthesis (e.g., purE and pyrE) were predicted to be gained at node 82. Our analyses suggest that half of those apparent gene gains occurred through origination events, possibly by acquisition from bacteria (Supplementary Data 5; Supplementary Figs. 17–21; Supplementary Figs. 24–38). Genes involved in energy production and conversion including Rnf complex and EtfAB/Bcd complex were apparently acquired at node 79 through origination events, possibly transferred from bacteria as well (Supplementary Figs. 10–16). Furthermore, approximately 10% and 7% of genes gained at node 73 were related to amino acid transport and metabolism, and nucleotide transport and metabolism, respectively (Supplementary Data 5). These apparent gain events may have contributed to the enhanced metabolic potential of subgroup J.

**Discussion.** Woesearchaeota are one of the most ubiquitously distributed lineages within the DPANN superphylum18,27. In the current study, we retrieved the Woesearchaeota 16S rRNA gene sequences from different biotopes, to understand the global distribution patterns of Woesearchaeota, and collated 103 Woesearchaeota genomes with 49 new ones reconstructed in this study to gain insights into the metabolism and evolutionary history of Woesearchaeota diversification.

We here first updated the previously published 16S rRNA gene-based division of Woesearchaeota subgroups18, by inclusion and robust phylogenomic inference from newly reconstructed genomes (see Supplementary Note 1). We anticipated that the newly defined Woesearchaeota subgroups A–J would facilitate understanding of their global distribution pattern. We showed that Woesearchaeota abundance is relatively low in natural environments, although their roles in maintaining the community stability, considering their high diversity, might be important28,29. Based on the phylogenetic diversity index analysis, Woesearchaeota are more abundant and more diverse in saltmarshes and mangroves than in other biotopes. Sand and soil appear to be the least preferred habitats of Woesearchaeota. Finally, the approximate link between genome subgroups and 16S rRNA gene sequence clusters indicated subgroup G, I, and J may have high ecological adaptability. However, more 16S rRNA gene sequences from genomes are needed to accurately navigate subgroup distribution in the environment.

Despite the wide distribution across 11 distinct biotopes, Woesearchaeota appear to share some common metabolic
abilities, including the lack of complete TCA cycle and electron transport chains (complexes I−IV), and the minimal capacity to synthesize biomolecules, such as nucleotides, precursors for isoprenoids, amino acids, and vitamins. Further, a complete non-oxidative phase of the pentose phosphate pathway and some genes involved in fermentation, such as \textit{ldh}, were recovered from most genomes. Hence, being consistent with the previous analyses, these features highlight a conspicuous metabolic deficiency of Woesearchaeota, and indicate that most of these archaea might mainly lead an anaerobic and parasitic/fermentation-based lifestyle. How these organisms acquire nutrients and essential building blocks remains unclear because of the lack of pure cultures. These archaea may be intimately associated with other microorganisms, akin to the relationship between \textit{N. equitans} and \textit{I. hospitalis}, so as to obtain necessary cellular metabolites, such as amino acids, nucleotides, and lipids. The
diversity-generating retroelements abounding in DPANN (highly represented in Woesearchaeota) may also aid the adaption to such a lifestyle\(^{30}\). Despite the above unifying traits, the metabolic potentials of Woesearchaeota subgroups vary pronouncedly. Specifically, subgroup J microbes appear to harbor genes involved in the biosynthesis of amino acids and nucleotides more frequently than the other subgroups, suggesting a relatively greater biosynthetic capacity. Interestingly, the MEP pathway, prevalent in bacteria that synthesize isoprenoid precursors, appears to be specific to subgroup J. These differences highlight the notion that Woesearchaeota are diversified organisms, as also confirmed by the genomic size variation, and phylogenetic analysis based on both, the 16S rRNA gene and single-copy orthologs. More surprisingly, two nearly complete MAGs from subgroup J (Yap20000.bnd4.8 and YT1_182) encode the complete glycolysis pathway and an amyloase (GH57), suggesting their full potential to metabolize starch. In the glycolytic pathway, two NADH and four reduced ferredoxins are generated respectively in the conversion of glyceraldehyde 3-phosphate to 3-phospho-D-glycerol phosphate via glyceraldehyde 3-phosphate dehydrogenase, and in the citric acid decarboxylation by pyruvate/2-oxoacid-ferredoxin oxidoreductase. These two reducing agents could be synergistically utilized by [FeFe] hydrogenase driving the evolution of $\mathbb{H}_2$. Reduced ferredoxins could also fuel the translocation of sodium ion or proton across the membrane by Rnf complex, generating potential gradient and NADH. In EtAB/Bcd complex, NADH donates electrons to reduce ferredoxin. Therefore, it is likely that the balance of the reducing pool is maintained by the glycolytic pathway, [FeFe] hydrogenase, Rnf complex, and EtAB/Bcd complex in these Woesearchaeota. These results indicate that some subgroup J members, if not all, may be capable of anaerobic heterotrophy with fermentative metabolism. To some degree, their lifestyle may have a resemblance with some fermentative bacterial organisms like Thermotoga maritima\(^{31,32}\) and Clostridium thermocellum\(^{33}\). According to previous studies\(^{16,17}\), most CAZymes of DPANN archaea are extracellular. It is also possible that subgroup J members secrete some CAZymes and peptidases into the surrounding medium to decompose starch or other organic matter derived from dead cells or exuded by living cells. Therefore, subgroup J might be able to associate themselves with particles. However, fluorescent in situ hybridization, isotopic labeling, or pure culture experiments are essential to confirm the actual state or lifestyle of these organisms\(^{34,35}\). [FeFe] hydrogenases are capable of catalyzing $\mathbb{H}_2$ formation with efficiency hitherto unreported\(^{22}\). Ecologically, members of subgroup J might be important in anoxic environments like hydrothermal vents and marine sediments where they occur more frequently (Supplementary Fig. 5f) and could benefit other microbes like $\mathbb{H}_2$-utilizing methanogenic archaea\(^{18}\) by producing $\mathbb{H}_2$. Therefore, the above observations suggest that Woesearchaeota may impact the carbon and hydrogen cycle\(^{6}\).

Whether DPANN archaea form a clan is still debated\(^{36,37}\), as some DPANN lineages exhibit high rates of sequence evolution, making them vulnerable to long-branch attraction. Some recent reports support the clanhood of DPANN archaea\(^{36,37}\). Further, inference of the ancestral metabolism of DPANN archaea revealed incomplete glycolysis pathway and TCA cycle in their common ancestor\(^{37}\). Woesearchaeota are placed together with Nanoarchaeota, Nanohaloarchaeota, Parvarchaeota, Parchearchaeota, Aenigmarchaeota, and Huberarchaeota in most reports\(^{6,27,37}\). Owing to their reduced genomes, DPANN archaea are thought to be symbionts or parasites of other prokaryotes\(^{6,7}\). Indeed, the host dependence of some of these organisms has been substantiated. For example, *N. equitans*, the first DPANN archaean to be characterized, is obligately dependent on *L. hospitalis*\(^8\), and Candidatus Nanoaloarchaeum antarcticus requires *Halorubrum lacusprofundi* as a host for growth\(^{10}\). Suggesting that symbiotic or parasitic lifestyle may be a common feature of these DPANN lineages. Nevertheless, the hypothetical lifestyle of DPANN has been recently challenged in a study that proposed that most DPANN archaea do not lead a symbiotic lifestyle in subsurface environments\(^{11}\).

The presence of a complete glycolysis pathway and extensive repertoires of genes for amino acid and nucleotide biosynthesis as well as hydrogenase in subgroup J highlights the metabolic flexibility of these microbes. Subgroup J archaea first experienced gains of genes related to the transport and metabolism of amino acids and nucleotides and then genes of energy production and conversion. These two steps perhaps enable a reduced dependence on a host or shift to another life strategy that does not require cell–cell association to obtain exogenous cellular components, unlike many other autotrophic microorganisms\(^8\). The acquisition of these genes has putatively enhanced the biosynthetic capacity and energy production of these organisms. However, currently, the underpinning drivers are unclear. The acquisition of genes related to $\mathbb{H}_2$ production could be induced by symbiotic switching or an alternative resource acquisition strategy\(^39\). Bacterial organisms, most likely, play major roles in the origination of additional metabolic traits in subgroup J. The acquisition of many genes related to isoprenoid and amino acid biosynthesis is predicted as origination events and likely from bacterial donors. Likewise, genes encoding protein complex vital for energy production like Rnf complex are also projected as origins and possibly transferred from bacteria. Although, *pfk*, *pta*, and *ackA*, important genes in carbohydrate metabolism, are not inferred as gain through originations at the nodes of subgroup J, phylogenetic analyses indicated that they are likely transferred from bacteria as well (Supplementary Figs. 39–41). For example, *pfk* of subgroup J are phylogenetically closely related to those of *Candidatus* Abyssubacteria (Supplementary Fig. 39). This is consistent with previous analysis which showed genes related to carbohydrate metabolism are prone to lateral gene transfer\(^37\).

**Methods**

**Processing of 16S rRNA gene amplicon data.** Raw sequence data were retrieved from sediment samples (78 samples)\(^{39}\) collected along the southeast coast of China and from the EMP\(^{41}\), based on in silico binding analysis of primers 515F and 806R\(^{41}\). These data were then combined and trimmed using Sickle (v1.33; https://github.com/jclove/sickle) and the “–q 25” setting to control sequence quality. Sample metadata were subsequently retrieved from a reference paper\(^{40}\) and the EMP website (https://earthmicrobiome.org/protocols-and-standards/metadata-guide/). To remove chimeric sequences, data that passed quality control were searched using USEARCH\(^{42}\) (v2.13.3) against the Greengenes database (v13.8)\(^{43}\). The reads were clustered into OTUs at the threshold of 97% sequence identity. Taxonomy was assigned using the "assign_taxonomy.py" function in QIME\(^{44}\) (v2018.11) against the custom database composed of SILVA SSU 132 database (https://www.arb-silva.de/documentation/release-132) and an in-house Woesearchaeota archaea database. Finally, 2163 16S rRNA gene libraries were selected, using the following criteria: minimum sequencing depth of 30,000 and minimum relative abundance of Woesearchaeota in the libraries. OTUs were aligned with "align_seqs.py" using the default setting. The OTU table was rarefied to 30,000. Alpha and beta diversity were calculated using "alpha_diversity.py" and "beta_diversity.py" respectively. Metadata of libraries used in the final analysis could be found in Supplementary Data 2.

**Sample collection and sequencing.** Six YT sediment samples were collected on November 15, 2018 from a seagrass bed and a neighboring site not covered by seagrass at the Rongcheng Swan Nature Reserve (Rongcheng) as described in Liu et al.\(^{45}\). Samples were taken at depths of 0–2, 21–26, and 41–46 cm, sealed in sterile plastic bags, and transported to the laboratory in a pre-cooled container. Total DNA was extracted from 10 g of each sample using PowerSoil DNA Isolation kit (QIAGEN, Hilden, Germany), according to the manufacturer’s protocol. The extracted DNA was then sequenced with Illumina HiSeq2500 (San Diego, Calif., U.S.) PE150 at Novogene.

Five YT sediment samples from Futian Nature Reserve (Shenzhen) were collected on April 17, 2017 as described for YT samples at depths of 0–2, 6–8,
Binning processes were performed with MetaBAT249 12 times, with 12 combinations of speciﬁcities and sensitivity parameters (–m maxP 60 or 95 AND –minS 60 or 95 AND –maxEdges 200 or 500) with MetaBAT256 (v2.12.1) at a minimum sequence length of 2000 bp. Finally, the bins were screened using the Das-Tool39 (v1.0) program to obtain high-quality and high-completeness bins.

**JLR and Yap metagenomes.** For JLR metagenomes, raw sequencing reads were dereplicated at 100% identity and trimmed with Sickle (v1.33, https://github.com/najoshi/sickle) using default settings. The trimmed reads were assembled using IDBA-UD35 (v1.1.1) with the parameters: –minK 65, –maxK 145, –step 10. The binning processes were performed with MetaBAT256 12 times, with 12 combinations of speciﬁcities and sensitivity parameters (–m maxP 60 or 95 AND –minS 60 or 95 AND –maxEdges 200 or 500) with MetaBAT256 (v2.12.1) at a minimum sequence length of 2000 bp. Finally, the bins were screened using the Das-Tool39 (v1.0) program to obtain high-quality and high-completeness bins.

**Gene calls and functional annotations.** Proteins were predicted using prodigal (v2.6.3) embedded in prokka (v1.13) and the Kingdom Archaean database as well as the GC/A%-biased options75,78. All predicted proteins were subjected to a uniform annotation protocol. Speciﬁcally, InterProScan (v5.38-76.0, client version)79 was used to classify protein functions, with applications including CDD80, Pfam81, SMART82, and TIGRFAM83 enabled. Protein functions were also annotated using eggNOG-mapper (v0.13) and a KO hit and cluster of orthologous groups (COG) hit, based on eggNOG (v5.0) clusters64. Additionally, hmmscan (v3.1b2; settings: –e 1e–4)65 and the Carbohydrate-active enzymes (CAZymes) database were used to identify carbohydrate-active enzymes (downloaded from dbCAN in July 201984 in Woesearchaeota genomes). Peptidases were also annotated using DIAMOND BLASTP (v0.9.24; settings: –k 1 –e 1e–10 query-cover 80–50) against the MEROPS database (release 12.0)85. Transporters were also annotated using DIAMOND BLASTP (v0.9.24; settings: –e 1e–4) against the Transporter Classiﬁcation Database (downloaded in November 2020)86. Hydrogenases were classiﬁed by using HydBk (https://services.birc.au.dk/hydbk). The annotations were then manually inspected, and are summarized in Supplementary Data 3.

**Selection of orthologs for phylogenetic analysis of Woesearchaeota.** To enable accurate classiﬁcation of Woesearchaeota genomes based on phylogeny, we ﬁrst identiﬁed a curated set of pepF (OG 20604), cnaM (OG 20640), and rplF (OG 20672) genes (79% complete and <5% contaminated) by running the OMA standalone algorithm (v2.4.1)73. This set of proteins were annotated with arcOGs hhm models from eggNOG (v5.0)84 and TIGRFAM (v15.0) database with hmmscan (v3.1b2)64 and the best hit for each protein was selected based on the lowest e-value and the highest bit score (see Supplementary Data 6). Based on the annotations, we excluded 12 OGs that are affected by lateral gene transfer (LGT) or have complex gene history according to previous publications47,72. For the remaining OGs, whenever available, we extracted HMM proﬁles from arcOG and TIGRFAM databases and if not present in the databases, proﬁles were generated with HMMBuild87. To minimize the risk of detecting distant paralogs, the reference gene history according to previous publications36,72. For the remaining OGs, whenever available, we extracted HMM proﬁles from arcOG and TIGRFAM databases and if not present in the databases, proﬁles were generated with HMMBuild87. To minimize the risk of detecting distant paralogs, the reference gene history according to previous publications36,72. For the remaining OGs, whenever available, we extracted HMM proﬁles from arcOG and TIGRFAM databases and if not present in the databases, proﬁles were generated with HMMBuild87. To minimize the risk of detecting distant paralogs, the reference gene history according to previous publications36,72. For the remaining OGs, whenever available, we extracted HMM proﬁles from arcOG and TIGRFAM databases and if not present in the databases, proﬁles were generated with HMMBuild87. To minimize the risk of detecting distant paralogs, the reference gene history according to previous publications36,72. For the remaining OGs, whenever available, we extracted HMM proﬁles from arcOG and TIGRFAM databases and if not present in the databases, proﬁles were generated with HMMBuild87. To minimize the risk of detecting distant paralogs, the reference gene history according to previous publications36,72. For the remaining OGs, whenever available, we extracted HMM proﬁles from arcOG and TIGRFAM databases and if not present in the databases, proﬁles were generated with HMMBuild87. To minimize the risk of detecting distant paralogs, the reference gene history according to previous publications36,72.
Phylogenetic analysis of [FeFe] hydrogenase. Sequences for classifying [FeFe] hydrogenase of Woesearchaeota were retrieved from HydDB. To reduce sampling size, sequences affiliated with [FeFe] hydrogenase groups A, B, and C were dereplicated using CD-hit (v4.8.1)94 with a sequence identity of 0.9, 0.85, and 0.65 respectively prior to the selection of [FeFe] hydrogenase catalytic subunits of Woesearchaeota. Catalytic domains of [FeFe] hydrogenases were identified according to Pfam (v32.0)95 annotations. Amino acid sequences of catalytic domains were aligned with MAFFT-LINSI74 and trimmed with ClipKIT (v0.1; settings: -m medium)84. Using the trimmed alignments, a maximum-likelihood tree was inferred based on LG+C+G+ model using IQ-Tree (v1.6.8)37 with 1000 ultrafast bootstraps88. Note that group A [FeFe] hydrogenase could not be reliably subdivided by phylogeny and are classified into subtypes based on their genetic organizations20. The genetic organizations of all [FeFe] hydrogenase in Woesearchaeota MAGs were visualized with DNA Features Viewer (v3.0.3) in Prokaryotic Relative Hydrogenase Catalytic Subunit, Woesearchaeota sequences were queried against Genebank non-redundant database (nr, downloaded March, 2020)53 with DIAMOND BLASTP subunit, Woesearchaeota sequences were aligned with MAFFT-LINSI74 and trimmed with BMGE (v1.12; default settings)75. A tree was generated using IQ-Tree (v1.6.8; settings: -m LG+G+Gubb-1000)77. Phylogenetic markers were used in ETE Toolkit (v3.1.2)86.

Phylogenetic analysis of [NiFe] hydrogenase. Backbone sequences for classifying [NiFe] hydrogenase of Woesearchaeota were retrieved from HydDB. A total of 218 sequences [NiFe] hydrogenase was randomly selected from each subtype and combined with Woesearchaeota sequences. They were aligned with MAFFT-LINSI74 and trimmed with BMGE (v1.12; default settings)77. A tree was generated using IQ-Tree (v1.6.8; settings: -m LG+G+Gubb-1000)77.

Phylogenetic analysis of 16S rRNA gene. The 16S rRNA gene sequences of Woesearchaeota and Pacearchaeota were first retrieved from the SILVA SSU 132 database25, and then combined with sequences from a recent study18. The 16S rRNA gene sequences from Woesearchaeota and Pacearchaeota MAGs were identified using barーン v0.93 (https://github.com/marriott/barrn). A total of 816 16S rRNA gene sequences (from Woesearchaeota MAGs, 28 from Pacearchaeota MAGs, and three from unclassified MAGs; 823 from SILVA SSU 132 database or Liu et al.18) were dereplicated and were used as an outgroup for the phylogenetic analysis. These sequences were aligned using the SINA Align Service44 (https://www.arb-silva.de/aligner/). Full alignments were trimmed with trimAl95 (v1.4.rev15; settings: -gappyout). A maximum-likelihood tree was then inferred using IQ-Tree (v2.0.7; settings: -m LG+C+G+F +F +R +bb 1000 -alrt 1000) based on the concatenation of trimmed 15 ribosomal proteins78,99.

Phylogenetic analysis of other individual proteins. Phylogenetic analysis of other individual protein trees was conducted as described for searching closest relatives of [FeFe] hydrogenase catalytic subunit for Woesearchaeota sequences in the nr database35. Detailed models used for each protein could be found in Supplementary Figs. 10–22, 24–41.

Statistical analysis. Statistical analysis was done in R version 3.6.097 (R development Core Team, Vienna, Austria). One-way analysis of variance (ANOVA) and post-hoc test was conducted to compare the relative abundance of Woesearchaeota across different biotopes using “aov” and “TukeyHSD” function. The t-distributed stochastic neighbor embedding (t-SNE) analysis and permutation multivariate analysis of variance (PERMANOVA) based on the unweighted UniFrac metric were completed to test whether Woesearchaeota community shifted among different salinity and biotopes using “Rtsete” in Rssete package (v0.15) and “adonis” function in vegan package (v2.5–7) respectively. Significance of associations of physicochemical parameters with Woesearchaeota community was assessed using Mantel tests with Pearson’s correlation and 999 permutations.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
The Woesearchaeota MAGs have been deposited in eMSG (an elibrary of Microbial Systematics and Genomics, https://www.biosinfo.org/elmng/index) and are also available from the NCBI under the BioProject identifier PRJNA460835. The accession number for the MAGs are available in Supplementary Data 1. DNA sequencing data are deposited in the NCBI SRA under the BioProject identifier PRJNA460835 and PRJNA460840. Dataset generated in this study (i.e., protein files for the MAGs, alignments, and tree files) are available through https://doi.org/10.6084/m9.figshare.14459535. Public database used in this study included SILVA SSU 132 database [https://www.arb-silva.de/documentation/release-132], Greengenes (v13.8) [https://greengenes.microbio.me/greengenes_release/ gg_13_8.otus/], CDD (v3.17) [ftp://ftp.ncbi.nih.gov/pub/ncbi-cdd/cdd/], Pfam (v32.0) [ftp://ftp.ncbi.nih.gov/pub/ncbi-cdd/cdd/Pfam/v32.0/]. Custom scripts for analyzing phylogenetic trees and ecological analysis have been deposited at https://doi.org/10.6084/m9.figshare.14459535. Additionally, we used the published custom code ALI helper scripts [https://github.com/Tancatha/phytree/master/ALI] to parse ALE outputs.

Received: 10 September 2020; Accepted: 12 August 2021; Published online: 06 September 2021

References
1. Brochier-Armant, C., Bousau, B., Gribaldo, S. & Forterre, P. Mesophilic archaea illuminate the origin of eukaryotic cellular complexity. Nat. Rev. Microbiol. 6, 245–252 (2008).
2. Zhou, Z., Pan, J., Wang, F., Gu, J.-B. & Li, M. Bathyarchaeota: globally distributed metabolic generalists in anoxic environments. FEMS Microbiol. Rev. 42, 639–655 (2018).
3. Vanwonterghem, I. et al. Methylo trophic methanogenesis discovered in the archaean phyllum Ver straelarchaeota. Nat. Microbiol. 1, 1–9 (2016).
4. Zaremba-Niedzwiedzka, K. et al. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541, 353–358 (2017).
5. Zhou, Z., Liu, Y., Li, M. & Gu, J.-D. Two or three domains: a new view of tree of life in the genomics era. Appl. Microbiol. Biotechnol. 102, 3049–3058 (2018).
6. Castelle, C. J. et al. Genomic expansion of domain archaia highlights roles for organisms from new phyla in anaerobic carbon cycling. Curr. Biol. 25, 690–701 (2015).

NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-021-25565-9 | www.nature.com/naturecommunications
7. Rinke, C. et al. Insights into the physiological and coding potential of microbial dark matter. Nature 499, 431–437 (2013).
8. Jaeger, U. et al. Improved 16S rRNA gene amplification from Ignicoccus hospitalis: new insights into a unique, intimate association of two archaea. J. Bacteriol. 190, 1743–1750 (2008).
9. Golyshina, O. V. et al. ‘ARMAN’ archaea depend on association with euryarchaeal host in culture and in situ. Nat. Commun. 8, 60 (2017).
10. Hansen, J. N. et al. Unexpected host dependency of Antarctic Nanohaloarchaeota. Proc. Natl Acad. Sci. USA 116, 14661–14670 (2019).
11. Beam, J. P. et al. Ancestral absence of electron transport chains in the hyperthermophilic bacterium Thermotoga maritima. Proc. Natl Acad. Sci. USA 114, E6402–E6411 (2017).
12. Wang, H. et al. Different microbial distributions in the Yellow River delta. Curr. Opin. Biotechnol. 84, 1–11 (2021).
13. Lipsewers, Y. A., Hopmans, E. C., Sinninghe Damsté, J. S. & Villanueva, L. 16S rRNA gene evolution: a case study of the enigmatic firmicute genus Thaumarchaeum. ISME J. 11, 174–181 (2017).
14. Buckel, W. & Thauer, R. K. Flavin-based electron bifurcation, ferredoxin, and alternative electron transport in the hydrogen economy. Fems Microbiol. Rev. 34, 380–406 (2013).
15. Chen, Y., Liu, Y. & Wang, X. Spatiotemporal variation of bacterial and archaeal community structures in the Black Sea. ISME J. 11, 184–197 (2017).
16. Castelle, C. J. & Banfield, J. F. Microbial communities in the oceanic water column: a comprehensive analysis. ISME J. 11, 383–397 (2017).
17. Castelle, C. J. et al. Synthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. Nat. Rev. Microbiol. 16, 629–645 (2018).
18. Liu, X. et al. Insights into the ecology, evolution, and metabolism of the widespread Woesearchaeal lineages. Microbiome 6, 102 (2018).
19. Ortiz-Alvarez, R. & Casamayor, E. O. High occurrence of Pacearchaeota and Archaea that lack rRNA sequences in microbial communities suggests an underappreciated diversity of novel ammonia-oxidizing archaea from the Jiujiang River Estuary. Appl. Environ. Microbiol. 86, 1 (2020).
20. Uritski, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome 6, 138 (2018).
21. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHT: an ultra-fast and comprehensive metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31, 1674–1676 (2015).
22. Sieber, C. M. K. et al. Recovery of genomes from metagenomes via a dereplication, aggregation, and scoring strategy. Nat. Microbiol. 3, 836–843 (2018).
23. Peng, Y., Leung, H. C. M., Liu, Z. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28, 1420–1428 (2012).
24. Bolger, A. M., Lohse, M. & Usadel, B. Trimmmomatic: a flexible trimmer for illumina sequence data. Bioinformatics 30, 2114–2120 (2014).
25. Benson, D. A. et al. GenBank. Nucleic Acids Res. 28, 15–18 (2000).
26. Markowitz, V. M. et al. The integrated microbial genomes (IMG) system. Nucleic Acids Res. 34, D34–D348 (2006).
27. Parks, D. H. et al. A complete domain-to-species taxonomy for bacteria and archaea using automated annotation on single cells, and metagenomes. Genome Res. 25, 1043 (2015).
28. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinform. 11, 119 (2010).
29. Seemann, T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068–2069 (2014).
30. Quevillon, E. et al. InterProScan: protein domains identifier. Nucleic Acids Res. 33, W116–W120 (2005).
31. Marchler-Bauer, A. et al. CDD: NCBI’s conserved domain database. Nucleic Acids Res. 43, D222–D226 (2015).
32. Finn, R. D. et al. Pfam: the protein families database. Nucleic Acids Res. 42, D222–D230 (2014).
33. Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P. & Bork, P. SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Res. 28, 231–234 (2000).
34. Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. Nucleic Acids Res. 31, 371 (2003).
35. Eddy, S. R. A new generation of homology search tools based on probabilistic inference. Genome Inf. Int. Conf. Genome Inf. 23, 205–211 (2009).
36. Zhang, H. et al. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 46, W95–W101 (2018).
37. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12, 59–60 (2015).
38. Rumble, P. N. et al. The IFAAGP database of proteolytic enzymes, their substrates, and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. Nucleic Acids Res. 46, D264–D268 (2018).
69. Saier, M. H., Reddy, V. S., Tamang, D. G. & Västermark, A. The transporter classification database. Nucleic Acids Res. 42, D251–D258 (2014).

70. Søndergaard, D., Pedersen, C. N. S. & Greening, C. HyDB: a web tool for hydrogenase classification and analysis. Sci. Rep. 6, 1–8 (2016).

71. Altenhoff, A. M. et al. OMA stand-alone: orthology inference among public and custom genomes and transcriptomes. Genome Res. 29, 1152–1163 (2019).

72. Wolfe, J. M. & Fournier, G. P. Horizontal gene transfer constrains the timing of microbial evolution. Mol. Ecol. Evol. 2, 897–903 (2018).

73. Dombrowski, N. et al. Undinarchaeota illuminate DPANN phylogeny and the impact of gene transfer on archaeal evolution. Zenodo https://doi.org/10.5281/zenodo.3672835 (2020).

74. Kataki, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3099–3110 (2002).

75. Criscuolo, A. & Gribaldo, S. BMGE (block mapping and gathering with entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol. Biol. 10, 210 (2010).

76. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One 5, e9490 (2010).

77. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37, 1530–1534 (2020).

78. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and developments. Nucleic Acids Res. 47, W256–W259 (2019).

79. Minh, B. Q. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol. 36, 996–1004 (2018).

80. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20, 238 (2019).

81. Aramaki, T. et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics 36, 2251–2252 (2020).

82. Steenwyk, J. L., Iii, T. J. B., Li, Y., Shen, X.-X. & Rokas, A. ClipKIT: a multiple sequence alignment trimming software for accurate phylogenetic inference. PLoS Biol. 18, e3001007 (2020).

83. Sheridan, P. O. et al. Gene duplication drives genome expansion in a major lineage of Thaumarchaeota. Nat. Commun. 11, 5494 (2020).

84. Parks, D. H. et al. A unified approach for comprehensive microbial community profiling and analysis. Microbiome 6, 165 (2018).

85. Sheridan, P. O. et al. Gene duplication drives genome expansion in a major lineage of Thaumarchaeota. Nat. Commun. 11, 5494 (2020).

86. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Bioinformatics 36, 3230–3232 (2020).

87. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973 (2009).

88. Boyd, J. A., Woodcroft, B. J. & Tyson, G. W. GraffM: a tool for scalable, phylogenetically informed classification of genomes within metagenomes. Nucleic Acids Res. 46, e59 (2018).

89. Zulkower, V. & Rosser, S. DNA features viewer: a sequence annotation formatting and plotting library for Python. Bioinformatics 28, 1823–1829 (2012).

90. Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python.

91. Pruesse, E., Peplies, J. & Glöckner, F. O. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics 28, 1823–1829 (2012).