Deciphering symbiotic interactions of 'Candidatus Aenigmarchaeota' with inferred horizontal gene transfers and co-occurrence networks

Yu-Xian Li
Sun Yat-Sen University
Yang-Zhi Rao
Sun Yat-Sen University
Yan-Ling Qi
Sun Yat-Sen University
Yan-Ni Qu
Sun Yat-Sen University
Ya-Ting Chen
Sun Yat-Sen University
Jian-Yu Jiao
Sun Yat-Sen University
Wen-Sheng Shu
South China Normal University
Hongchen Jiang
China University of Geosciences
Brian P. Hedlund
University of Nevada
Zheng-Shuang Hua
Sun Yat-Sen University
Wen-Jun Li (liwenjun3@mail.sysu.edu.cn)
Sun Yat-Sen University https://orcid.org/0000-0002-1233-736X

Research

Keywords: 'Ca. Aenigmarchaeota', DPANN, Symbiosis, Horizontal gene transfer, Co-evolution network

Posted Date: March 13th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-17151/v1

Version of Record: A version of this preprint was published at mSystems on July 27th, 2021. See the published version at https://doi.org/10.1128/mSystems.00606-21.

Abstract

Background: '*Ca.* Aenigmarchaeota' represents an evolutionary branch within the DPANN superphylum. However, their ecological roles and potential host-symbiont interactions are poorly understood.

Results: Here, we analyze eight metagenomic-assembled genomes from hot spring habitats and reveal their functional potentials. Although they have limited metabolic capacities, they harbor substantial carbohydrate metabolizing abilities. Further investigation suggests that horizontal gene transfer might be the main driver that endows these abilities to '*Ca.* Aenigmarchaeota', including enzymes involved in glycolysis. Additionally, members from the TACK superphylum and Euryarchaeota contribute substantially to the niche expansion of '*Ca.* Aenigmarchaeota', especially genes related to carbohydrate metabolism and stress responses. Based on co-occurrence network analysis, we conjecture that '*Ca.* Aenigmarchaeota' may be symbionts associated with TACK archaea and Euryarchaeota, though host-specificity might be wide and variable across different '*Ca.* Aenigmarchaeota' genomes.

Conclusion: This study provides significant insights into possible host-symbiont interactions and ecological roles of '*Ca.* Aenigmarchaeota'.

Background

With advances in sequencing technologies and bioinformatic approaches, insight into the "unseen majority" of prokaryotes has become possible, even when they inhabit complex microbial communities, leading a tremendous expansion of known archaeal diversity [1, 2, 3, 4, 5, 6, 7]. Among recently proposed major archaeal lineages, the DPANN superphylum has inspired considerable research attention, which has uncovered their surprisingly small genome sizes, lack of genes associated with core biosynthetic pathways [3, 8, 9], yet extensive phylogenetic and functional diversity [10, 11, 12]. 'Candidatus Aenigmarchaeota', representing the "A" of the DPANN superphylum, was first uncovered and named as "Deep Sea Euryarchaeotic Group (DSEG)" [13]. Later, based on single-amplified genomes (SAGs), this lineage was defined and proposed as a novel phylum [2]. Other studies integrating metagenomic and metatranscriptomic sequencing revealed that this phylum lacks many essential metabolic pathways and may possess fermentative and symbiotic lifestyles [3, 8]. However, our understanding of the metabolic characteristics, functional diversity, and potential host-symbiont interactions of 'Ca. Aenigmarchaeota' is far from complete.

Here, we apply comparative and evolutionary genomics analyses on eight new MAGs along with 15 publicly available genomes to fill these gaps. Our study reveals a symbiotic lifestyle for 'Ca. Aenigmarchaeota' based on the absence of a large proportion of genes involved in basic metabolism. Further analyses suggest that TACK archaea or Euryarchaeota might be the potential mutualistic hosts for 'Ca. Aenigmarchaeota'. Direct contact between the symbionts facilitates the occurrence of HGT, which further improves the competitive abilities of 'Ca. Aenigmarchaeota' by expanding their carbohydrate metabolic repertoire and stress responses. By integrating the HGT inference and co-occurrence network

construction, this study represents the first attempt to reveal a potential symbiotic relationship solely based on the genomic data.

Results And Discussion

Phylogeny and distribution of '*Ca*. Aenigmarchaeota'

Eight MAGs of '*Ca.* Aenigmarchaeota' were successfully reconstructed from five hot spring sediments collected in Tengchong county in Yunnan, China (Fig. 1a, Additional file 1: Table S1), including four from a single sample from Diretiyanqu-6 (DRTY-6) and one for each of the other springs Diretiyanqu-7 (DRTY-7), Gumingquan (GMQ), Qiaoquan (QQ), and Jinze (JZ-2) (Table 1) [14]. '*Ca*. Aenigmarchaeota' represents a rare group in hot spring habitats with relative abundances of all MAGs < 0.5% (Fig. 1b). Most MAGs are high-quality, with completeness > 90% and nearly no contamination, and detectable 16S rRNAs and tRNAs (>21) (Additional file 1: Fig. S1; Table 1; Additional file 1: Table S2) [15]. Compared to related MAGs from other studies, these have smaller genome sizes (0.64 vs. 0.86 Mbp; Mann-Whitney *U* test, P = 0.0003; Additional file 1: Fig. S2) and a lower range of GC content (average 31.74% vs. 38.59%; Mann-Whitney *U* test, P = 0.012) [2,3,8]. They harbor less genes (752 vs. 1070; Mann-Whitney *U* test, P = 0.002), shorter average gene length (771 vs. 683 bp; Mann-Whitney *U* test, P = 0.0005), remarkably high coding density (88-94.6%) and number of overlapping genes (~20.6%) (Table 1). This is consistent with previous studies suggesting that thermophiles harbor small genome sizes as a result of genomic streamlining due to high fitness costs of life at high temperature [16]. Both whole genome-based phylogenomic and 16S rRNA gene-based phylogenetic analyses revealed that the eight MAGs from this study are highly scattered within the phylum '*Ca.* Aenigmarchaeota' with high bootstrap confidences (Fig. 1c, Additional file 1: Fig. S3). Eight MAGs from Crystal Geyser share a high average AAI (>99 %), and the AAIs of the other 15 genomes range from 38% to 98.24% (Additional file 1: Fig. S4, Additional file 1: Table S3). Recruited 16S rRNA sequences representing this phylum demonstrate that '*Ca.* Aenigmarchaeota' represents an evolutionary diverse group that inhabits a broad range of habitats (Additional file 1: Fig. S5; Additional file 2), including freshwater (40.96%) and marine (27.71%) environments, hot springs (7.63%), hydrothermal vents (6.83%), and groundwater (7.63%). A minor portion of the 16S rRNA gene sequences were retrieved from hypersaline lakes and soils (<5%). PCA results based on KEGG and arCOG annotations show that genomes are clustered by phylogenetic position rather than habitat type (Additional file 1: Fig. S6).

Metabolic features of '*Ca*. Aenigmarchaeota'

Consistent with previous studies, all eight MAGs in this study have a limited metabolic capacity. Pathways including the tricarboxylic acid cycle (TCA), fatty acid metabolism, and dissimilatory/assimilatory sulfur and nitrogen metabolisms were missing [8,17,18]. Unlike most genomes in DPANN [3,8], '*Ca.* Aenigmarchaeota' MAGs possess a near-complete glycolytic pathway (Fig. 2). Phosphofructokinase (PFK) and glucokinase (GK) were solely found in DRTY-6_2 bin_201 and GMQ_1 bin_18-1, respectively. The absence of pyruvate kinase (PK) and pyruvate kinase isozymes R/L (PKLR) prohibit the conversion of phosphoenolpyruvate (PEP) to pyruvate during the last step of glycolysis.

However, a gene encoded by phosphoenolpyruvate synthase (*pps*), which has been shown to perform the same function in thermophiles [19,20], was detected in most MAGs in this study (Additional file 3). All MAGs lack gluconeogenesis pathways because they are missing glucose-6-phosphatase, a key enzyme that hydrolyzes glucose-6-phosphate to glucose [21]. Several MAGs are predicted to ferment by production of acetate, lactate, and ethanol based on the presence of acetate-CoA ligase, L-lactate dehydrogenase, and acetaldehyde dehydrogenase/alcohol dehydrogenase, and would therefore be predicted to promote syntrophy among community members [3]. The lack of electron transport chain (ETC), especially V/A type ATPase, suggests that '*Ca*. Aenigmarchaeota' cannot produce ATP via a membrane proton-motive force (PMF). Thus, glycolysis and other fermentation pathways could be the main sources of ATP for '*Ca.* Aenigmarchaeota' [22]. However, MAGs from groundwater lack a glycolytic pathway [3,8].

Two genes relevant to polysaccharide degradation including α-amylase (for starch and glycogen) and α-1,6-glucosidase (for starch and disaccharides) were identified in many of the hot spring MAGs (Additional file 1: Table S4). DRTY-6_1 bin 65 might also degrade and utilize pullulan (GH13_20) [23,24]. Aside from α-amylase, MAGs from groundwater harbored different and more glycoside hydrolases, including β-glucosidases (for disaccharides), endoglucanases (for cellulose) glucoamylases (for starch), β-1,2-mannosidases (for beta-1,2-mannotriose and beta-1,2-mannobiose) and α-1,4-galactosaminogalactan hydrolase (for galactosaminogalactan). This might reflect a greater abundance and heterogeneity of carbohydrates in groundwater. None of the known carbon fixation pathways were detected in these MAGs, though three MAGs contain archaeal ribulose-bisphosphate carboxylase (Rubisco). Phylogenetic analysis suggests that the six Rubisco genes recovered from '*Ca*. Aenigmarchaeota' belong to the form III group, of which five are from thermal environments (Additional file 1: Fig. S7). Four of them belong to group III-b and the remaining two could be classified as a novel lineage which clustered with Rubisco genes from Candidate Phyla Radiation (CPR) genes, which were previously suggested to have been passed by HGT from '*Ca*. Aenigmarchaeota' to CPR [25]. As previously described in other archaea, Rubisco genes may function in the CO_2-incorporating adenosine 5'-monophosphate (AMP) pathway, together with genes encoding for adenosine monophosphate phosphorylase and ribose-1,5-biphosphate isomerase [25,26]. We identified different types of hydrogenases in '*Ca*. Aenigmarchaeota'. Eight MAGs from thermal habitats harbor NiFe 3b-type hydrogenases and cluster into one clade, whereas NiFe 4-type has only been found in '*Ca*. Aenigmarchaeota' from non-thermal habitats (Additional file 1: Fig. S8).

Despite the possession of genes involved in glycolysis, fermentation, and the nucleotide salvage pathway, the absence of many pivotal pathways strongly suggests a symbiotic lifestyle for '*Ca*. Aenigmarchaeota'. Firstly, this archaeal phylum is devoid of *de novo* amino acid biosynthetic pathways. Consequently, they have to salvage amino acids from environmental sources by degrading extracellular proteins and peptides and transporting by-products into the cell by using a variety of extracellular peptidases, membrane peptidases, cytoplasmic peptidases, and proteases, along with amino acid transporters (Additional file 3) [3]. Secondly, *de novo* nucleotide biosynthetic pathways are absent in most

of the genomes of this phylum. Meanwhile, genes for purine and pyrimidine salvage pathways were detected in most of genomes, especially in MAGs from hot springs (Additional file 3). Thirdly, '*Ca*. Aenigmarchaeota' genomes are unable to synthesize cell membranes *de novo* due to the lack of genes for synthesis of sterol isoprenoids involved in the mevalonate pathway (MVA), although genes encoding mevalonate kinase, glycerol-1-phosphate dehydrogenase, and associated enzymes for phospholipid biosynthesis have been detected [27,28]. Nevertheless, genes encoding S-layers, archaella, and type-IV pili were identified in eight MAGs, endowing "*Ca.* Aenigmarchaeota" with protection, motility, and cell-to-cell attachment abilities, which could consequently facilitate host-symbiont interactions [29,30,31].

Stress responses employed by '*Ca*. Aenigmarchaeota'

Comparative genomics showed that "*Ca.* Aenigmarchaeota" inhabiting thermal habitats harbor more genes involved genetic information processing including transcription, translation, replication and repair, and folding, sorting and degradation (Additional file 1: Fig. S9a). Combine with the smaller genome sizes of these thermophiles, these hint the smaller number of accessory genes and genome reduction may drive the losses. In addition, genes involved in cell motility, posttranslational modification, protein turnover, and chaperones predominantly were enriched in the thermophiles (Additional file 1: Fig. S9b). This reflects the fact that cells in at high temperatures have to combat constant thermal denaturation of both macromolecules and monomers [32,33,34]. '*Ca.* Aenigmarchaeota' MAGs from thermal habitats have evolved multiple strategies to overcome this stress. Chaperonin GroEL, associated with repair of DNA and protein damage caused by high temperature, was present all '*Ca.* Aenigmarchaeota' genomes (Fig. 3) [8,35]. The prevalence of DNA repair protein RadA could be used for homologous recombination and an alternative strategy for DNA repair [36], indicating the prevalence of genome streamlining among these genomes. Type I (IA and IB type) topoisomerases and reverse gyrase, the latter considered a hallmark of hyperthermophily, were solely detected in MAGs from thermal habitats, which could stabilize DNA and modulate DNA topology to maintain the structure and integrity of chromosome [34,37,38]. HSP70 (DnaK), DnaJ and GrpE were consistently absent in MAGs from thermal habitats, but commonly detected in groundwater MAGs (Fig. 3), consistent with the potential role of this system in the adaptation to mesophily [39].

The presence of cell defense systems provides protection from virus infection (Additional file 1: Fig. S10). One MAG DRTY-7_1 bin_34 was detected to contain a CRISPR-Cas system (Class III-A, Additional file 1: Fig. S10a). All three types of restriction modification (RM) systems were found in '*Ca*. Aenigmarchaeota' [40] (Additional file 1: Fig. S10b). The type II RM system was detected in five of the MAGs. Additionally, the recently discovered Hachiman system was detected in DRTY-6 bin_65 and Gabija system in DRTY-6 bin_215, providing broad protection against viruses [41] (Additional file 1: Fig. S10c). The broad distribution of defense systems in thermophilic '*Ca*. Aenigmarchaeota' suggests viruses could be an important threat for the survival of these and other microbes in hot spring habitats [42,43,44]. Viruses with different morphologies have been detected in hot springs and have been shown to be highly active *in situ* [14,41]. Hence, it seems plausible that '*Ca.* Aenigmarchaeota' may confer their hosts with the immunity to viruses by serving as 'viral decoys' [45]. Meanwhile, the attracted virus could be degraded,

and the DNA can be recycled as a nucleotide source [45]. Therefore, the host-symbiont interaction between '*Ca.* Aenigmarchaeota' and its potential hosts appears to be mutually beneficial.

Horizontal gene transfer (HGT) in '*Ca*. Aenigmarchaeota'

Frequent horizontal gene transfer (HGT) between symbionts has been observed, owing to the physical proximity of genetic material [16,46,47]. As a consequence, organisms in stable symbioses can pass genes between each other via HGT, facilitating the co-evolution of partners during the long-term development of symbionts. Therefore, the detection of HGTs among '*Ca.* Aenigmarchaeota' genomes may provide a clue to infer potential hosts. Surprisingly, results uncovered a lower proportion of HGT-derived genes in '*Ca.* Aenigmarchaeota' in comparison to thermophilic, free-living Archaea; e.g., Aigarchaeota (mean 7.5 vs. 22.9%; Mann-Whitney *U* test, P = 0.00029) [7]. This result might reflect extensive genomic streamlining in symbiotic '*Ca.* Aenigmarchaeota'. Additionally, it seems plausible that restricted contact with non-hosts may limit the diversity of genes available for transfer to '*Ca.* Aenigmarchaeota'. Intriguingly, most MAGs within '*Ca.* Aenigmarchaeota' possessed comparable percentages of HGT-derived genes (mean 7.5%; Additional file 1: Table S5). A significant positive correlation (Pearson's R^2 = 0.74, P value: 6.185e-6) was observed between detected HGTs and predicted gene totals in corresponding genomes regardless of habitat and genome sizes (Fig. 4a and b). MAGs from hot springs showed relatively larger variance in HGT rate, and thermophiles had both the highest and lowest HGT frequencies.

As is typical, very few of the detected HGTs in '*Ca*. Aenigmarchaeota' are predicted to be involved in information processing (Additional file 1: Fig. S11c and d). HGT was less prevalent in the genomes of thermophiles; however, genes involved in carbohydrate metabolism and protein folding, sorting, and degradation were more frequent among HGTs in MAGs from thermal habitats. Thus, we hypothesize that HGT-derived genes for carbohydrate metabolic genes may have expanded their niche and/or facilitated their survival under stressful conditions. By looking into the potential donors, we found that members from DPANN including, phyla '*Ca*. Micrarchaeota' and '*Ca*. Woesearchaeota', contributed the most to the genetic innovations of MAGs from groundwater (Fig. 4c). However, the reported small genome size and potential symbiotic lifestyle of these members suggest they are not the hosts of '*Ca*. Aenigmarchaeota'. Unlike the MAGs from groundwater, less HGTs were observed between these two phyla and '*Ca*. Aenigmarchaeota' in hot springs (Fig. 4c). Other HGT-derived genes could be traced to Euryarchaeota and TACK superphylum archaea, indicating a potential host-symbiont relationship, consistent with previous studies reporting symbioses between DPANN and TACK archaea and Euryarchaeota [10,16,48,49,50].

To infer the potential hosts for MAGs inhabiting hot springs and groundwater, we compared the acquired genes between these groups. Results indicated very little overlap between HGT-derived genes among these two habitat groups, only 39 KOs and 54 arCOGs (14.6 and 17.4% among all HGTs which could be assigned into a KO or arCOG number) (Additional file 1: Fig. S11a and b), suggesting they have distinct hosts, despite the broad phylogenetic distribution of these two habitat groups. MAGs from hot springs tend to form symbionts with members from TACK archaea or Euryarchaeota. This conjecture can be

exemplified by the case study of the PFK gene, which was derived via HGT and likely drives niche expansion due to an expanded carbohydrate metabolism in '*Ca*. Aenigmarchaeota' (Additional file 1: Fig. S11c; Additional file 4). A BLAST search against the NCBI-nr database showed that the best hit to the only detected PFK gene in DRTY-6_2 bin_201 is from '*Ca*. Bathyarchaeota'. Intriguingly, potential hosts may confer abilities to adapt harsh conditions. The detected reverse gyrase in GMQ bin_18-1 suggested that TACK-derived genes ('*Ca*. Bathyarchaeota') may have improved the fitness of this organism to invade high-temperature niches (Additional file 4). Also, several genes encoding superoxide dismutase (SOD2) and 8-oxo-dGTP diphosphatase (*mutT*) were identified as HGTs (Additional file 4 and 5), which could be employed to resist oxidative damage. BLAST searches of *mutT* genes confirmed that TACK members including Thaumarchaeota and '*Ca*. Bathyarchaeota' may have served as donors. In QQ bin_128, a gene encoding an alkyl hydroperoxide reductase (*ahpF*) is situated near the *mutT* gene, which might be involved in responding to oxidative stress. Strikingly, the SOD2 gene in GMQ bin_18-1 even shows 89.1% identity and 100% of coverage to that from *Pyrobaculum* sp. WP30, indicating a rather recent HGT event. BLAST searches against NCBI-nr database suggest that all genes except this one in GMQ bin_18-1 show identity < 80%, even for those vertical inherited genes, revealing a novel lineage of this MAG with high divergence to the extant genomes. Therefore, it is reasonable to infer that *Pyrobaculum* has high possibility to be the host of GMQ bin_18-1 and the genetic interaction might be ongoing. In DRTY-6 bin_65, DRTY-6 bin_202 and DRTY-7 bin_34 MAGs, BLAST searches suggest that less than ten genes of each show identity > 80% to the existing genes. Three of them including transcription initiation factor IIB (TFIIB), phage integrase and translation elongation factor EF-1 (EEF1A) are identified to belong to phyla outside of DPANN, with former two show high identity to Euryarchaeota and the third one to Planctomycetes. However, none of them are identified as HGTs. In DRTY-6 bin_65, all three genes are in the same scaffold. The taxonomic information of nearby genes of TFIIB and EEF1A were mostly affiliated to '*Ca*. Aenigmarchaeota', consolidating the inference of vertical inheritance of these two genes. The phage integrase may be transferred horizontally from Euryarchaeota since six and three of the ten downstream genes are close relatives to Theionarchaea (51.3-77.4%) and Thermoplasmata (48.3-61.1%). Among them, three genes exhibit homologies to methyltransferase, DNA binding protein and restriction endonucleases associated with type II restriction-modification (RM) systems. This suggests that the integrase-mediated HGT may confer '*Ca*. Aenigmarchaeota' the special niche to resist the infection of virus. Thus, we conjecture that these three MAGs tend to form symbionts with Euryarchaeota and result potential bidirectional gene flows between symbionts and hosts (Additional file 1: Fig. S12). Overall, the above observations further support the hypothesis that TACK archaea and Euryarchaeota may serve as possible hosts. Though several bacterial phyla contribute a substantial of HGTs to '*Ca*. Aenigmarchaeota', such as Firmicutes, we rule out the possibilities of them as hosts since most of the transferred genes represent ancient events with high divergence (~ 42% of identity for all best hits belonging to Firmicutes), which cannot reflect the current symbiotic state.

Co-occurrence pattern of '*Ca*. Aenigmarchaeota' with community members

To further explore the potential host-symbiont relationships between '*Ca*. Aenigmarchaeota' and the TACK superphylum and Euryarchaeota, a network interface was constructed, aiming to reveal microbial co-

occurrence patterns and possible ecological interactions. The network encompasses 891 nodes and 11 of them are identified as '*Ca*. Aenigmarchaeota' OTUs. Modularity analysis showed that four modules contained OTUs from '*Ca*. Aenigmarchaeota' (Fig. 5). After extracting '*Ca*. Aenigmarchaeota'-containing sub-networks from each module, we observed tight connections between this phylum and OTUs from '*Ca*. Bathyarchaeota' (module 3 and 5), Crenarchaeota (module 5), Thaumarchaeota (module 3), and Euryarchaeota (module 2, 3, and 5). Thus, both HGT analysis and network analysis consistently suggest that TACK superphylum archaea and Euryarchaeota are likely to be the hosts of '*Ca*. Aenigmarchaeota'. Notably, we observed significant positive correlations between '*Ca*. Aenigmarchaeota' and several bacterial OTUs from Chloroflexi, Firmicutes, Nitrospirae, and Proteobacteria. Previous studies reported that these lineages are present at high abundance in their respective communities [51,52,53], and they contain several lineages of chemolithoautotrophs (lineages of Nitrospira, Firmicutes and Proteobacteria) or photoautotrophs (lineages of Chloroflexi) [54,55,56,57,58,59] and subsequently the primary producer of the community. The observed correlations could be a result of the nutrient dependency of '*Ca*. Aenigmarchaeota' hosts or the symbionts themselves to these primary producers (Additional file 1: Fig. S12).

Here, we propose a co-evolutionary scenario among archaea living in hot spring habitats. Generally, all members in the community form a complex network (Fig. 5). The heterotrophic hosts and symbionts would stay in proximity and form a mutual beneficial relationship with each other. By direct contact, they can fuel themselves by extracting these requisite nutrients directly into the cell to sustain their growth. In return, provide the host with virus defenses by using themselves as decoy and degrade viral DNA by multiple systems. Subsequently use them as a potential DNA and protein sources. Spontaneously, horizontal gene transfer event could occur with the direct contact which plays a pivotal role in expanding the niche diversity of symbionts '*Ca*. Aenigmarchaeota' (Additional file 1: Fig. S12).

Conclusions

The enigmatic 'Ca. Aenigmarchaeota', is still underexplored due to the insufficient cultured representatives or assembled genomes available. Here, we expanded the phylogenetic diversity of 'Ca. Aenigmarchaeota' in hot spring environment and showed that 'Ca. Aenigmarchaeota' are ubiquitous on earth. They harbor limited metabolic capacities by missing several pivotal biosynthetic pathways, such as nucleotides, amino acids and cell membrane biosynthesis, suggesting that such molecules need to be obtained from environment or from host as symbionts. Comparative genomics analysis reveals that genomes from thermal habitat possess smaller genome size but stronger capacity in metabolizing carbohydrates. HGT seems to have significantly contributed in these abilities, such as glycolysis in DRTY-6_2 bin_201. Besides, HGT identifies a salient number of gene flows from TACK and Euryarchaeota to 'Ca. Aenigmarchaeota', indicating the potential symbioses relationship between them. Further co-occurrence network confirms the inference by observing the close relationship between them. The tight correlation between bacterial member and 'Ca. Aenigmarchaeota' reflects the complicated relationship among different members. Overall, this study enables us to better understand the metabolic potentials and

possible interactions between ‘Ca. Aenigmarchaeota’ and their hosts, shedding light on the understanding of co-evolution and community assembly mechanisms.

Materials And Methods

Site description, sampling, DNA extraction and sequencing

All five hot spring sediment samples were collected from Tengchong County, Yunnan, China (N 24.95°, E 98.44°) in January 2017 and May 2018. DRTY-6, DRTY-7, GMQS and QQ samples are collected from DiReTiYanQu (DRTY), GuMingQuan (GMQ) and QiaoQuan (QQ) hot springs in Rehai National Park. JZ-2 (JinZe-2) sample were collected from JinZe Hot Spring Resort. See Additional file 1: Table S1 and Supplementary Note 1 to find the descriptions and geographical and physical information in detail of each sample site. The top 1 cm of sediment of each site was collected with a sterile iron spoon and transferred to a 50 mL centrifuge tube. All sediment samples were then stored in liquid nitrogen before transporting to the lab and stored at -20 ˚C in the lab until DNA extraction.

Community DNA was extracted from approximately 20 g of sediment for each sample using the PowerSoil DNA Isolation kit (MoBio). Libraries with insert size of 350 bp were constructed by using M220 Focused-ultrasonicator NEBNext and Ultra II DNA Library Prep Kit. The concentration of genomic DNA was measured with a Qubit fluorometer. The total genomic DNA was sequenced using Illumina Hiseq 4000 instrument at Beijing Novogene Bioinformatics Technology Co., Ltd (Beijing, China). On average 30 Gbp (2× 150 bp) of raw sequencing data for each sample was generated.

Metagenomic assembly and genome binning

Raw sequencing data were preprocessed to eliminate replicated reads and trim bases with low qualities, following workflows that were described previously [54]. All quality reads of each dataset were *de novo* assembled using SPAdes (version 3.9.0) [60] with the following parameters: -k 33,55,77,99,111 –meta. BBMap (version 38.85; http://sourceforge.net/projects/bbmap/) was used to compute the coverage information with the minimal identity cutoff of 90%. Genome binning was performed based on the calculated sequence depth and tetranucleotide frequency (TNF) of each scaffold using MetaBAT [61]. CheckM (version 1.0.5) [62] was used to identify all marker genes in each genome bin and then used to estimate the genome quality including genome completeness, contamination, and strain heterogeneity. Marker genes occurred more than once in each bin were treated as contaminations and associated contigs were removed manually. Subsequently, quality reads were recruited by mapping onto all optimized genome bins using BBMap and reassembled using SPAdes under the “–careful” mode. Genome bins were also visualized by fragmenting each scaffold into sequences with the length ranging from 5 to 10 kb using ESOM [63] for further curation. Contaminations and strain heterogeneity were estimated by CheckM [62], while genome completeness was estimated by 54 conserved archaeal single-copy genes (SCGs) (Additional file 1: Table S2) [3]. Finally, eight MAGs identified as ‘*Ca.* Aenigmarchaeota’ were kept for the later analysis.

Functional annotation of genome bins

The eight MAGs were submitted to the IMG-M (https://img.jgi.doe.gov/cgi-bin/m/main.cgi) database for gene prediction and functional annotation. For comparative genomics analysis, the annotation pipeline was also conducted locally. In brief, putative protein coding sequences (CDS) of all MAGs were determined using Prodigal [64] under "-p single" model. Functional annotations were performed by comparing predicted CDSs against KEGG [65], eggNOG [66], PFam [67] and arCOG [68] databases using DIAMOND [69] with the cutoff of E-value < $1e^{-5}$. Ribosomal RNA-coding regions were identified using RNAmmer [70]. All MAGs were uploaded to tRNAscan-SE On-Line [71] to identify the tRNA. The dbCAN webserver [72] was used to identify carbohydrate-active enzymes based on the Carbohydrate-Active enZYmes (CAZy) database.

Phylogenetic and phylogenomic analysis

Sixteen ribosomal protein sequences (L2, L3, L4, L5, L6, L14, L15, L16, L18, L22, L24, S3, S8, S10, S17 and S19) were selected to reconstruct the phylogenomic tree [73]. These sequences were identified by AMPHORA2 [74] and aligned using MUSCLE by iterations of 100 times [75]. The poorly aligned regions were eliminated using TrimAl with the parameters as: -gt 0.95 -cons 50 [76]. Then multiple alignments were concatenated and applied to reconstruct a maximal-likelihood phylogenetic tree using IQ-TREE (v1.6.11) with the following parameters: iqtree -s a -alrt 1000 -bb 1000 -nt AUTO [77,78].

16S rRNA genes were predicted for each MAG using RNAmmer. MAGs without 16S rRNA genes were further searched against RDP database [79] using the BLASTn program, and sequences with length > 300 bp were selected and combined with those retrieved by RNAmmer. All 16S rRNA sequences were uploaded to the SILVA web interface [80] and aligned using SINA alignment algorithm [81]. Columns contained more than 95% gaps were removed, after which a maximum -likelihood phylogenetic tree was constructed using IQ-TREE (v1.6.11) with the same parameters as the 16 ribosomal protein concatenated tree.

Detection of horizontally transferred genes

Twenty genomes with completeness > 75% and contamination < 5% evaluated by CheckM were taken into consideration for the inferences of horizontal gene transfers (HGTs). Putative HGTs were identified for each genome using HGTector v2 [82]. Due to lack of representative archaeal genomes in the pre-built default database, especially the genomes from DPANN and TACK superphyla, 3358 genomes were downloaded from RefSeq database on May 14, 2019. Genome quality was evaluated using CheckM except microbes from DPANN superphylum. Genome quality of DPANN superphylum was performed using the same procedure as mentioned above. Genomes with completeness < 80% and contamination > 10% were discarded. The remaining high-quality genomes were dereplicated at the phylum level using dRep (v2.3.2) [83]. Finally, 1133 genomes were picked out and 689 of those genomes complementary to the default database were appended. The combined sequences were compiled into a database using DIAMOND v0.8.22.84 [69] and the relevant taxonomy files were changed correspondingly. Then, the

“search” step was performed for the 20 high-quality ‘*Ca*. Aenigmarchaeota’ genomes with default parameters. During the “analyze” step, genomes belonging to unclassified Aenigmarchaeota (TaxID: 1046937) were treated as “closeTax”, but different species were used as “selfTax”. Specifically, Aenigmarchaeum subterraneum SCGC AAA011-O16 (TaxID: 743730) was set as the “selfTax” for all MAGs for DRTY-6_2 bin_215 and JZ-2_2 bin_245; Aenigmarchaeota archaeon ex4484_224 (TaxID: 2012503) was used as “selfTax” for GMQ_1 bin_18-1 and QQ_2 bin_128; Aenigmarchaeota archaeon ex4484_14 (TaxID: 2012502) was used as “selfTax” for DRTY-6_2 bin_201; Aenigmarchaeota archaeon JGI 0000106-F11 (TaxID: 1130284) was used for DRTY-6_1 bin_65, DRTY-7_1 bin_34 and DRTY-6_2 bin_202. The identified HGTs for each genome were visualized using SankeyMATIC (http://sankeymatic.com/).

Network-based co-occurrence analysis

A total of 88 hot spring samples across the time and spatial scale were used to reveal the co-occurrence pattern of “*Ca.* Aenigmarchaeota” with other community members. Metagenomic sequencing was conducted for all samples. Detailed quality control and assembly steps were done as described above. The generated scaffolds for each dataset were filtered with length < 2.5 kbp and *rpS3* gene were called using AMPHORA2 [74]. Coverage information was calculated by mapping the corresponding clean reads to the *rpS3* gene sequences. Specifically, the corresponding clean reads were mapped to *rpS3* gene sequences from the same sample using BBmap with the same parameters settings as described above. The generated .bam files were sorted using samtools v.1.3.1 [84]. Coverage information was subsequently calculated using “jgi_summarize_bam_contig_depths” program in MetaBAT. All predicted *rpS3* gene sequences from different datasets were combined and clustered into OTUs using USEARCH 9.2.84 with parameters as following: -cluster_smallmem -id 0.95. Taxonomy of the representative sequences was assigned by BLASTing to NCBI-nr database. The coverage information from each *rpS3* gene sequence was used to build the OTU table. OTUs occurred in less than three samples were filtered out to reduce the complexity and 891 OTUs were kept for the subsequent network construction. The co-occurrence network was constructed based on the Spearman correlation matrix using igraph v.1.2.4 [85]. Those significant (P value < 0.01) and robust correlations (rho > 0.6) between pairwise OTUs were used to infer a reliable network. Network visualization and relevant parameter calculations regarding to modularity, betweenness, closeness, average clustering coefficient, average weighted degree and average shortest path length were conducted in Gephi v.0.9.2 [86].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

All genomes in our study are available at JGI IMG-MER under the Study ID Gs0127627 and WGS accessions Ga0181641 (Unclassified Aenigmarchaeota DRTY7 bin_34), Ga0181640 (Unclassified Aenigmarchaeota DRTY6 bin_65), Ga0181639 (Unclassified Aenigmarchaeota GMQ bin_18-1), Ga0227293 (Unclassified Aenigmarchaeota JZ-2 bin_245), Ga0227294 (Unclassified Aenigmarchaeota DRTY-6 bin_215), Ga0261588 (Unclassified Aenigmarchaeota DRTY-6 bin_201), Ga0261590 (Unclassified Aenigmarchaeota DRTY-6 bin_202), Ga0261591 (Unclassified Aenigmarchaeota QQ bin_128).

Competing interests

The authors declare no competing interests.

Funding

This work was financially supported by National Natural Science Foundation of China (Nos. 91951205 and 91751206), Natural Science Foundation of Guangdong Province, China (No. 2016A030312003), Guangzhou Municipal People's Livelihood Science and Technology Plan (No. 201803030030).

Authors' contributions

Y.X.L., Y.Z.R., Y.L.Q., Y.N.Q., Z.S.H., and W.J.L. conceived the study. Y.X.L., Y.Z.R., Y.L.Q., H.C. J., and Y.N.Q., performed the measurement of physiochemical parameters and DNA extraction. Y.X.L., Y.Z.R., Z.S.H., Y.L.Q., Y.N.Q., Y.T.C., J.Y. J. and B.P.H. performed the metagenomic analysis, genome binning, functional annotation and evolutionary analysis. Y.X.L., Y.Z.R., Z.S.H., Y.L.Q., Y.N.Q., W.J.L., W.S.S. and B.P.H. wrote the manuscript. All authors discussed the results and commented on the manuscript.

Acknowledgements

We thank Guangdong Magigene Biotechnology Co., Ltd. China for the assistance in data analysis and the entire staffs from Yunnan Tengchong Volcano and Spa Tourist Attraction Development Corporation for strong support.

References

1. Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: The unseen majority. Proc. Natl Acad. Sci. USA. 1998;95:6578-83.
2. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature. 2013;499:431-37.
3. Castelle CJ, Wrighton KC, Thomas BC, Hug LA, Brown CT, Wilkins MJ, et al. Genomic expansion of domain Archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. Curr. Biol. 2015;25:690-701.

4. Adam PS, Borrel G, Brochier-Armanet C, Gribaldo S. The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. ISME J. 2017;11:2407-25.
5. Spang A, Ettema T. Microbial diversity: the tree of life comes of age. Nat. Microbiol. 2016;1:1-2.
6. Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, et al. Asgard Archaea illuminate the origin of eukaryotic cellular complexity. Nature. 2017;541:353-58.
7. Hua ZS, Qu YN, Zhu Q, Zhou EM, Qi YL, Yin YR, et al. Genomic inference of the metabolism and evolution of the archaeal phylum Aigarchaeota. Nat. Commun. 2018;9:1-11.
8. Castelle CJ, Brown CT, Anantharaman K, Probst AJ, Huang RH, Banfield JF. Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. Nat. Rev. Microbiol. 2018;499:1-17.
9. Kellner S, Spang A, Offre P, Szöllosi GJ, Petitjean C, Williams TA. Genome size evolution in the Archaea. Emerg. Top. Life Sci. 2018;6:1-11.
10. Wurch L, Giannone RJ, Belisle BS, Swift C, Utturkar S, Hettich RL, et al. Genomics-informed isolation and characterization of a symbiotic Nanoarchaeota system from a terrestrial geothermal environment. Nat. Commun. 2016;7:12115.
11. Ortiz-Alvarez R, Casamayor EO. High occurrence of Pacearchaeota and Woesearchaeota (Archaea superphylum DPANN) in the surface waters of oligotrophic high-altitude lakes. Environ. Microbiol. Rep. 2016;8:210-7.
12. Youssef NH, Rinke C, Stepanauskas R, Farag I, Woyke T, Elshahed MS. Insights into the metabolism, lifestyle and putative evolutionary history of the novel archaeal phylum 'Diapherotrites'. ISME J. 2015;9:447-60.
13. Takai K, Moser DP, DeFlaun M, Onstott TC, Fredrickson JK. Archaeal diversity in waters from deep south African gold mines. Appl. Environ. Microbiol. 2001;67:5750-60.
14. Hedlund BP, Cole JK, Williams AJ, Hou W, Zhou E, Li W, et al. A review of the microbiology of the Rehai geothermal field in Tengchong, Yunnan Province, China. Geoscience Frontiers. 2012;3:273-88.
15. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TB, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat. Biotech. 2017;35:725-31.
16. Sabath N, Ferrada E, Barve A, Wagner A. Growth temperature and genome size in Bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation. Genome Biol. Evol. 2013;5:966-77.
17. Golyshina OV, Toshchakov SV, Makarova KS, Gavrilov SN, Korzhenkov AA, La Cono V, et al. ‘ARMAN’archaea depend on association with euryarchaeal host in culture and in situ. Nat. Commun. 2017;8:1-12.
18. Baker BJ, Comolli LR, Dick GJ, Hauser LJ, Hyatt D, Dill BD, et al. Enigmatic, ultrasmall, uncultivated Archaea. Proc. Natl Acad. Sci. USA. 2010;107:8806-11.

19. Sapra R, Bagramyan K, Adams MW. A simple energy-conserving system: proton reduction coupled to proton translocation. Proc. Natl Acad. Sci. USA. 2003;100:7545-50.
20. Imanaka H, Yamatsu A, Fukui T, Atomi H, Imanaka T. Phosphoenolpyruvate synthase plays an essential role for glycolysis in the modified Embden-Meyerhof pathway in Thermococcus kodakarensis. Mol. Microbiol. 2006;61:898-909.
21. Arion WJ, Lange AJ, Walls HE, Ballas LM. Evidence for the participation of independent translocation for phosphate and glucose 6-phosphate in the microsomal glucose-6-phosphatase system. Interactions of the system with orthophosphate, inorganic pyrophosphate, and carbamyl phosphate. J. Biol. Chem. 1980;255:10396-406.
22. Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, et al. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. Science. 2012;337:1661-5.
23. Blesák K, Janeček Š. Sequence fingerprints of enzyme specificities from the glycoside hydrolase family GH57. Extremophiles. 2012;16:497-506.
24. Møller MS, Henriksen A, Svensson B. Structure and function of α-glucan debranching enzymes. Cell. Mol. Life Sci. 2016;73:2619-41.
25. Jaffe AL, Castelle CJ, Dupont CL, Banfield JF. Lateral gene transfer shapes the distribution of RuBisCO among candidate phyla radiation Bacteria and DPANN Archaea. Mol. Biol. Evol. 2018;36:435-46.
26. Sato T, Atomi H, Imanaka T. Archaeal type III RuBisCOs function in a pathway for AMP metabolism. Science. 2007;315:1003-6.
27. Villanueva L, von Meijenfeldt FB, Westbye AB, Hopmans EC, Dutilh BE, Damste JS. Bridging the divide: bacteria synthesizing archaeal membrane lipids. bioRxiv. 2018:448035.
28. Coleman GA, Pancost RD, Williams TA. Extensive transfer of membrane lipid biosynthetic genes between Archaea and Bacteria. bioRxiv. 2018:312991.
29. Albers SV, Meyer BH. The archaeal cell envelope. Nat. Rev. Microbiol. 2011;9:414-27.
30. Pohlschroder M, Pfeiffer F, Schulze S, Halim MF. Archaeal cell surface biogenesis. FEMS Microbiol. Rev. 2018;42:694-717.
31. Albers SV, Jarrell KF. The archaellum: how Archaea swim. Front. Microbiol. 2015;6:23.
32. Hedlund BP, Thomas SC, Dodsworth JA, Zhang CL. Life in High-Temperature Environments. Manual of Environmental Microbiology. 2016:4-3.
33. Bains W, Xiao Y, Yu C. Prediction of the maximum temperature for life based on the stability of metabolites to decomposition in water. Life. 2015;5:1054-100.
34. López-García P, Zivanovic Y, Deschamps P, Moreira D. Bacterial gene import and mesophilic adaptation in archaea. Nat. Rev. Microbiol. 2015;13:447-56.
35. Zeilstra-Ryalls J, Fayet O, Georgopoulos C. The universally conserved GroE (Hsp60) chaperonins. Annu. Rev. Microbiol. 1991;45:301-25.

36. Nakai R, Abe T, Takeyama H, Naganuma T. Metagenomic analysis of 0.2-μm-passable microorganisms in deep-sea hydrothermal fluid. Mar. biotech. 201;13:900-8.
37. Forterre P. A hot story from comparative genomics: reverse gyrase is the only hyperthermophile-specific protein. Trends Genet. 2002;18:236-7.
38. Kampmann M, Stock D. Reverse gyrase has heat-protective DNA chaperone activity independent of supercoiling. Nucleic Acids Res. 2004;32:3537-45.
39. Petitjean C, Moreira D, López-García P, Brochier-Armanet C. Horizontal gene transfer of a chloroplast DnaJ-Fer protein to Thaumarchaeota and the evolutionary history of the DnaK chaperone system in Archaea. BMC Evol. Biol. 2012;12:226.
40. Dupuis MÈ, Villion M, Magadán AH, Moineau S. CRISPR-Cas and restriction–modification systems are compatible and increase phage resistance. Nat. Commun. 2013;4:2087.
41. Doron S, Melamed S, Ofir G, Leavitt A, Lopatina A, Keren M, et al. Systematic discovery of antiphage defense systems in the microbial pangenome. Science. 2018;359:eaar4120.
42. Rachel R, Bettstetter M, Hedlund BP, Häring M, Kessler A, Stetter KO, et al. Remarkable morphological diversity of viruses and virus-like particles in hot terrestrial environments. Arch. Virol. 2002;147:2419-29.
43. Breitbart M, Wegley L, Leeds S, Schoenfeld T, Rohwer F. Phage community dynamics in hot springs. Appl. Environ. Microbiol. 2004;70:1633-40.
44. Ortmann AC, Wiedenheft B, Douglas T, Young M. Hot crenarchaeal viruses reveal deep evolutionary connections. Nat. Rev. Microbiol. 2006;4:520-8.
45. Burstein D, Sun CL, Brown CT, Sharon I, Anantharaman K, Probst AJ, et al. Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. Nat. Commun. 2016;7:10613.
46. Jain R, Rivera MC, Moore JE, Lake JA. Horizontal gene transfer accelerates genome innovation and evolution. Mol. Biol. Evol. 2003;20:1598-602.
47. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature. 2004;428:37-43.
48. La Cono V, Messina E, Rohde M, Arcadi E, Ciordia S, Crisafi F, et al. Differential polysaccharide utilization is the basis for a nanohaloarchaeon: haloarchaeon symbiosis. bioRxiv. 2019:794461.
49. Giannone RJ, Wurch LL, Heimerl T, Martin S, Yang Z, Huber H, et al. Life on the edge: functional genomic response of Ignicoccus hospitalis to the presence of Nanoarchaeum equitans. ISME J. 2015;9:101-14.
50. Hamm JN, Erdmann S, Eloe-Fadrosh EA, Angeloni A, Zhong L, Brownlee C, et al. Unexpected host dependency of Antarctic Nanohaloarchaeota. Proc. Natl Acad. Sci. 2019;116:14661-70.
51. Bohorquez LC, Delgado-Serrano L, López G, Osorio-Forero C, Klepac-Ceraj V, Kolter R, et al. In-depth characterization via complementing culture-independent approaches of the microbial community in an acidic hot spring of the Colombian Andes. Microb. Ecol. 2012;63:103-15.

52. Hou W, Wang S, Dong H, Jiang H, Briggs BR, Peacock JP, et al. A comprehensive census of microbial diversity in hot springs of Tengchong, Yunnan Province China using 16S rRNA gene pyrosequencing. PloS one. 2013;8.
53. Miller SR, Strong AL, Jones KL, Ungerer MC. Bar-coded pyrosequencing reveals shared bacterial community properties along the temperature gradients of two alkaline hot springs in Yellowstone National Park. Appl. Environ. Microbiol. 2009;75:4565-72.
54. Hua ZS, Han YJ, Chen LX, Liu J, Hu M, Li SJ, et al. Ecological roles of dominant and rare prokaryotes in acid mine drainage revealed by metagenomics and metatranscriptomics. ISME J. 2015;9:1280-94.
55. Kulp TR, Hoeft SE, Asao M, Madigan MT, Hollibaugh JT, Fisher JC, et al. Arsenic (III) fuels anoxygenic photosynthesis in hot spring biofilms from Mono Lake, California. Science. 2008;321:967-70.
56. Liu Z, Klatt CG, Wood JM, Rusch DB, Ludwig M, Wittekindt N, et al. Metatranscriptomic analyses of chlorophototrophs of a hot-spring microbial mat. ISME J. 2011;5:1279-90.
57. Miroshnichenko ML, Gongadze GM, Rainey FA, Kostyukova AS, Lysenko AM, Chernyh NA, et al. Thermococcus gorgonarius sp. nov. and Thermococcus pacificus sp. nov.: heterotrophic extremely thermophilic archaea from New Zealand submarine hot vents. Int. J. Syst. Evol. Microbiol. 1998;48:23-9.
58. Miroshnichenko ML, Tourova TP, Kolganova TV, Kostrikina NA, Chernych N, Bonch-Osmolovskaya EA. Ammonifex thiophilus sp. nov., a hyperthermophilic anaerobic bacterium from a Kamchatka hot spring. Int. J. Syst. Evol. Microbiol. 2008;58:2935-8.
59. Scott KM, Sievert SM, Abril FN, Ball LA, Barrett CJ, Blake RA, et al. The genome of deep-sea vent chemolithoautotroph Thiomicrospira crunogena XCL-2. PLoS Biol. 2006;4.
60. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J Comp. Biol. 2012;19:455-77.
61. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. Peer J. 2015;3:e1165.
62. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043-55.
63. Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP, Banfield JF. Community-wide analysis of microbial genome sequence signatures. Genome Biol. 2009;10:R85.
64. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
65. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45:D353-61.
66. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 2016;44:D286-93.

67. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44:D279-85.
68. Makarova K, Wolf Y, Koonin EV. Archaeal clusters of orthologous genes (arCOGs): An update and application for analysis of shared features between Thermococcales, Methanococcales, and Methanobacteriales. Life. 2015;5:818-40.
69. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat. Methods. 2015;12:59-60.
70. Lagesen K, Hallin P, Rødland EA, Stærfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 2007;35:3100-8.
71. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25:955-64.
72. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 2014;42:D490-5.
73. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. Nat. Microbiol. 2016;1:16048.
74. Wu M, Scott AJ. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics. 2012;28:1033-4.
75. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792-97.
76. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics.2009;25:1972-3.
77. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 2015;32:268-74.
78. Kalyaanamoorthy S, Minh BQ, Wong TK, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods. 2017;14:587.
79. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. 2014;42:D633-42.
80. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. 2007;35:7188-96.
81. Pruesse E, Peplies J, Glöckner FO. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics. 2012;28:1823-9.
82. Zhu Q, Kosoy M, Dittmar K. HGTector: an automated method facilitating genome-wide discovery of putative horizontal gene transfers. BMC Genomics. 2014;15:717.
83. Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. 2017;11:2864-8.

84. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078-9.

85. Csardi G, Nepusz T. The igraph software package for complex network research. Inter Journal Complex Systems. 2006;1695:1-9.

86. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media. 2009.

Table

Table 1 | General genomic features of the '*Ca.* Aenigmarchaeota' MAGs reconstructed from hot spring metagenome sequencing

Bins	DRTY-6_1 bin 65	DRTY-6_2 bin 201	DRTY-6_2 bin 202	DRTY-6_2 bin 215	DRTY-7_1 bin 34	GMQ_1 bin 18-1	JZ-2_2 bin 245	QQ_2 bin 128
Genome size (Mbp)	0.75	0.71	0.69	0.55	0.82	0.55	0.52	0.64
GC content (%)	31.6	30.9	28.6	25.7	29.1	25.5	47.9	34.2
N50	394,136	119,657	15,097	31,689	45,374	62,116	24,348	157,870
No. of scaffolds	19	9	69	25	27	19	25	8
Completeness (%)[†]	96.3	92.6	98.1	94.4	90.7	94.4	96.3	94.4
Contamination[‡]	0	0	0	0	0	0	0	0.93
Strain heterogeneity[‡]	0	0	0	0	0	0	0	0
No. of RNAs	42	20	35	22	43	47	28	34
5S rRNAs	2	1	1	0	1	1	1	0
16S rRNAs	0	1	1	1	1	1	1	1
23S rRNAs	1	1	0	0	1	1	1	1
tRNAs	39	17	33	21	40	44	25	32
No. of protein coding genes	865	729	840	662	923	662	637	757
Average length (bp)	817	857	754	789	820	790	747	747
Coding density (%)	93.9	88.0	92.1	94.5	92.8	94.4	92.3	88.0
Overlapped genes	223 (25.8%)	50 (6.9%)	193 (23%)	113 (17.1%)	229 (24.8%)	202 (30.5%)	105 (16.5%)	154 (20.3%)
No. of genes annotated by COG[*]	392 (43.2%)	422 (56.3%)	382 (43.7%)	367 (53.6%)	451 (46.6%)	336 (47.3%)	350 (52.6%)	383 (48.4%)
No. of genes annotated by KO[*]	313 (34.5%)	348 (46.4%)	324 (37%)	321 (46.9%)	353 (36.5%)	279 (39.3%)	287 (43.2%)	321 (40.5%)

[†]Genome completeness was calculated as the percentage of detected marker genes among 54 conserved single copy genes as listed in Table S2.

[‡]Genome quality including contamination and heterogeneity were estimated by CheckM[16].

[*]Functional annotation was conducted by uploading genomes to IMG database.

Figures

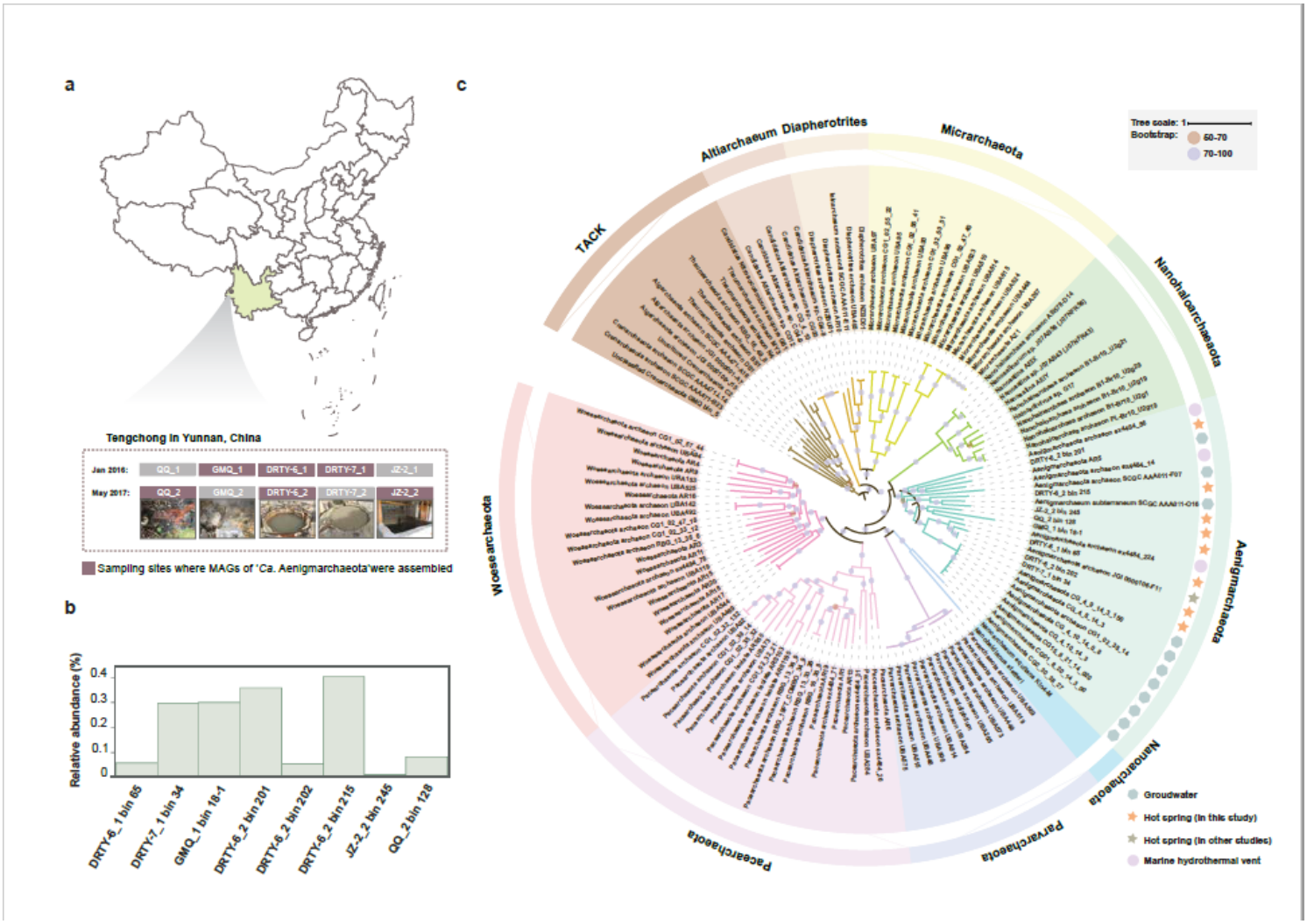

Figure 1

Phylogenetic analysis of reconstructed genomes of ‘Ca. Aenigmarchaeota’ sampled from hot spring sediments in Yunnan, China. (a) Geographic locations of sampling sites and relative abundance of ‘Ca. Aenigmarchaeota’ MAGs’ in each sample sites. Hot spring sediments of total five sites were collected in Tengchong, Yunnan, China, in Jan. 2016 and May. 2017. (b) Relative abundance of ‘Ca. Aenigmarchaeota’ MAGs’ was calculated by the coverage of the scaffolds of each MAG over the coverage of the all scaffolds of the corresponding metagenome. (c) Phylogenetic placement of the reconstructed MAGs. Maximum-likelihood phylogeny of eight Ca. Aenigmarchaeota MAGs in this study and reference genomes from TACK and DPANN superphyla. Phylogeny was constructed based on a concatenated set of 16 ribosomal proteins with 1000 bootstrap iterations. Bootstrap values ≥ 50 were presented in the middle of the branches in the phylogenetic tree. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

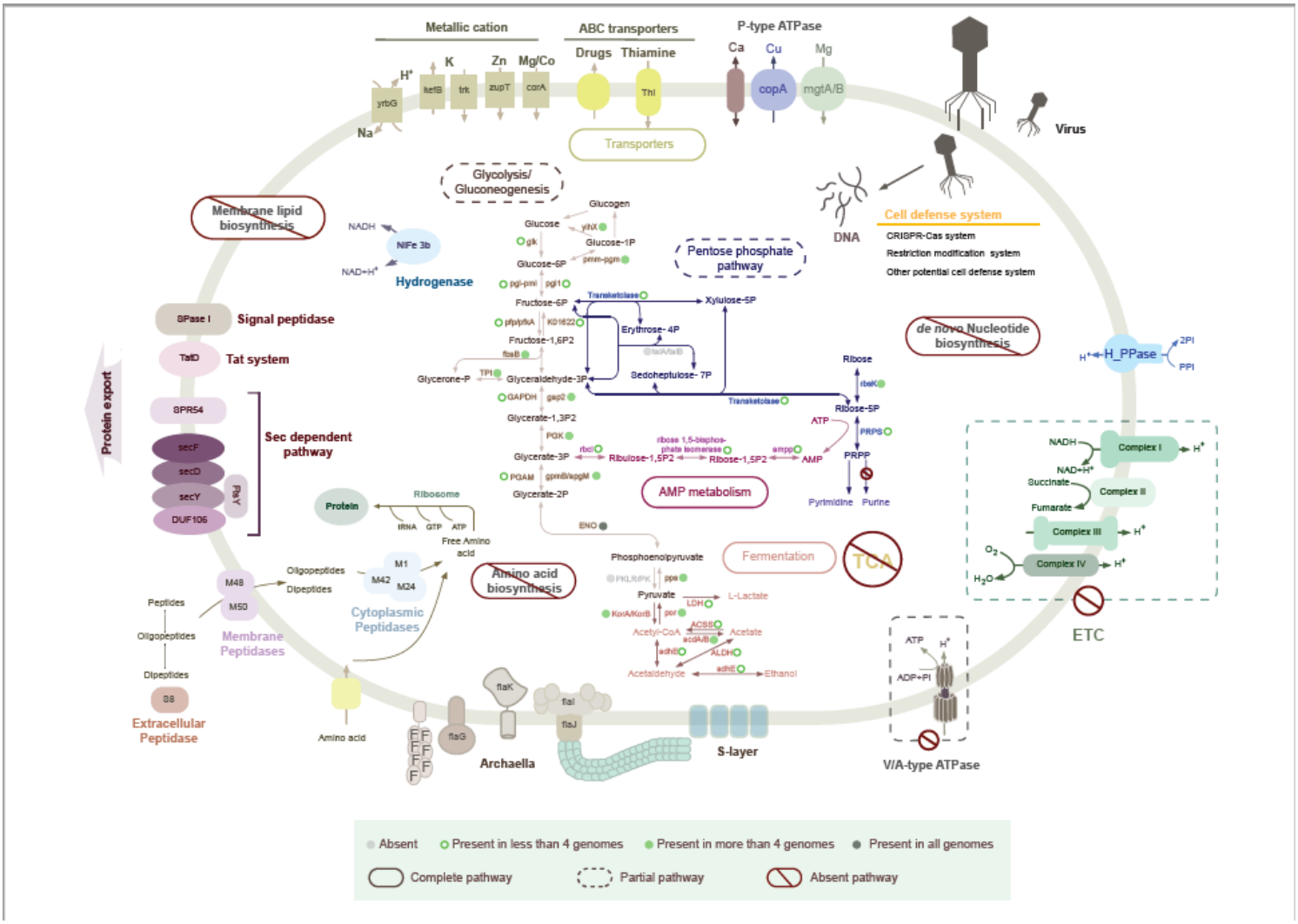

Figure 2

Reconstructed metabolic pathway of ‘Ca. Aenigmarchaeota’. Key genes involved in glycolysis, gluconeogenesis, galactose metabolism, pentose phosphate pathway, pyruvate metabolism, fermentation, AMP metabolism, protein biosynthesis and exportation, transporters, and flagellum are shown.

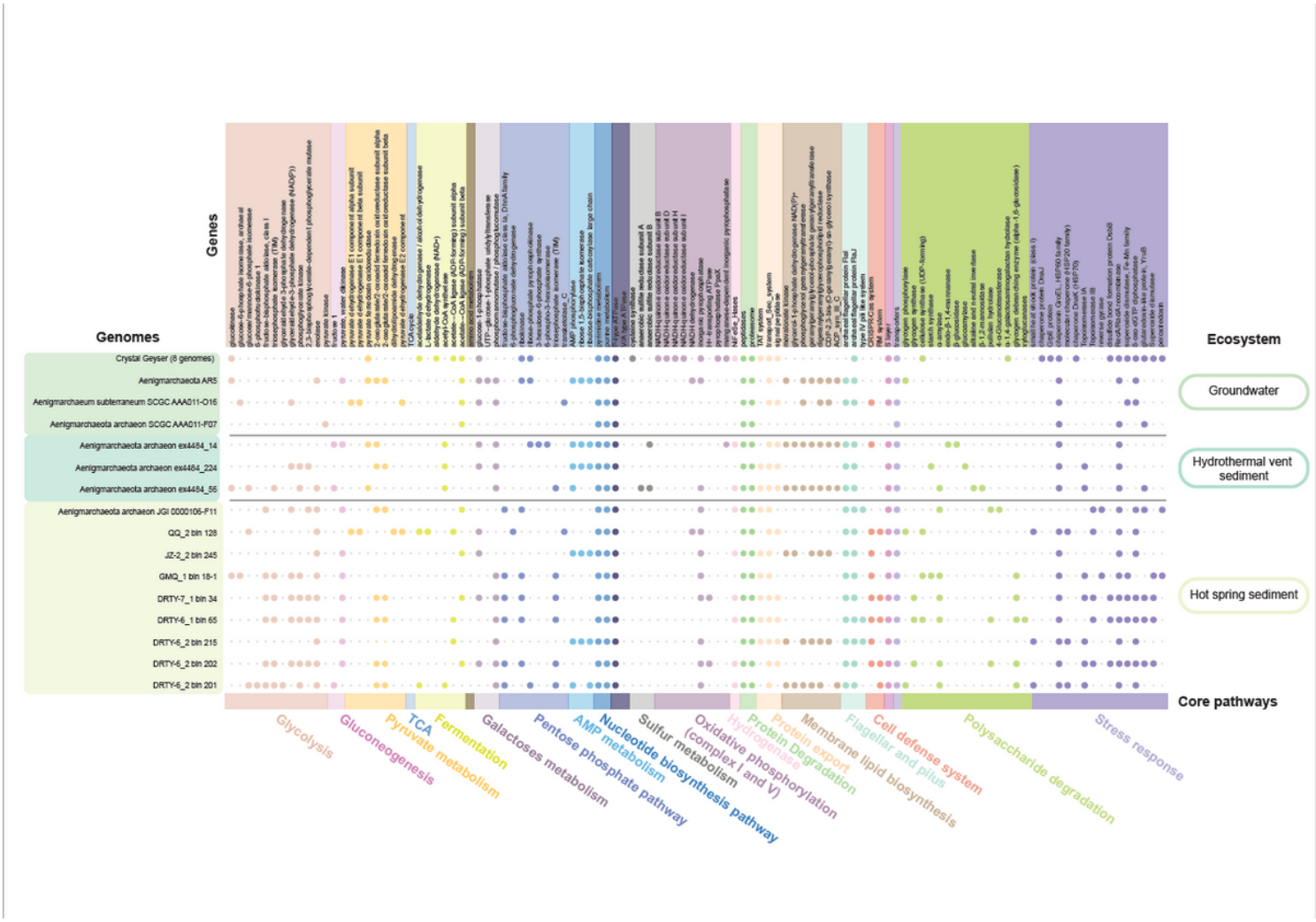

Figure 3

The core metabolic pathways with the presence/absence of genes of all ‘Ca. Aenigmarchaeota’ genomes. The metabolic potential (columns) is mainly generated based on KEGG Orthologues (KOs), clusters of orthologous groups (COGs) and archaeal clusters of orthologous groups (arCOGs). Genomes (rows) of ‘Ca. Aenigmarchaeota’ were clustered by habitats. Eight genomes obtained from groundwater in Crestal Geyser were shown as one row, because of the high metabolic similarity of the 8 genomes. Small grey dots represented genes or metabolic pathways were absent, and big dots represented genes or metabolic pathways were present. Distinct metabolic pathways were distinguished by different color dots.

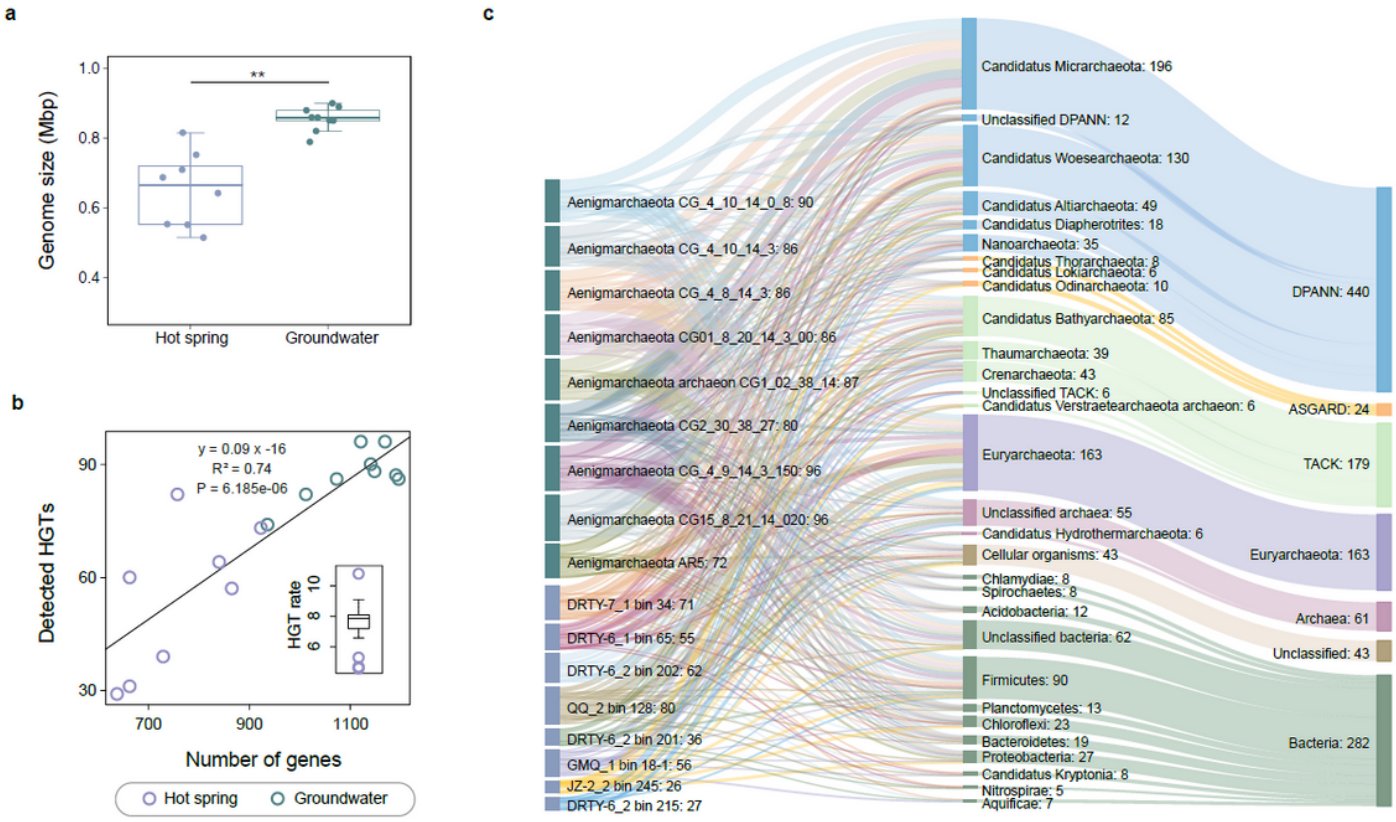

Figure 4

The identified horizontally gene transfers (HGTs) in genomes from this study and groundwater. (a) Comparisons of genomes sizes between these two groups. (b) Linear regression analysis was conducted to model the relationship between their genomes size and detected HGTs. The fitted formula was shown on the plot with R2 = 0.74 and P = 6.185e-06. (c) The Sankey plot shows the detected HGTs and potential donors for each genome. HGTs were detected using HGTector as described in Methods.

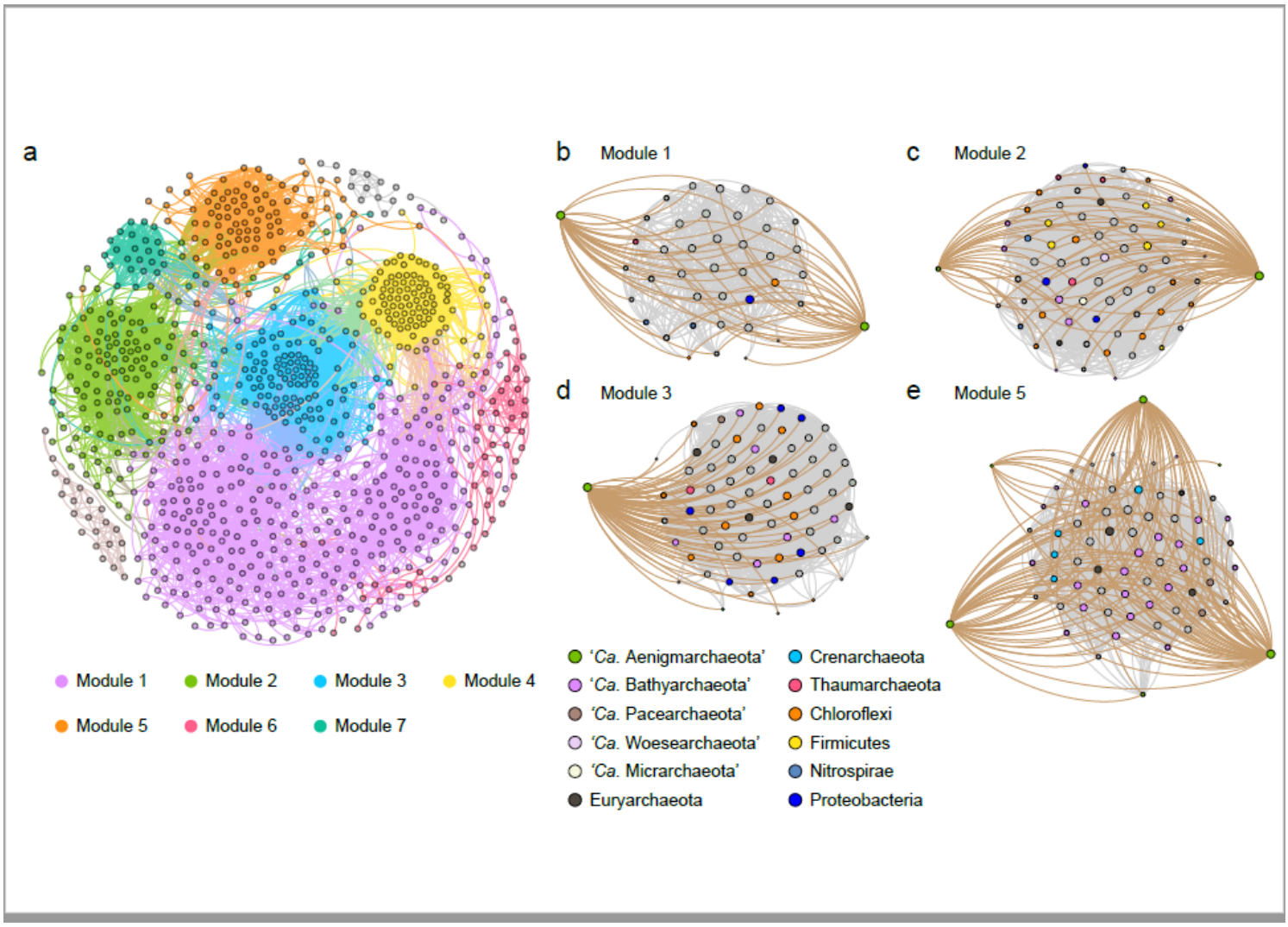

Figure 5

The co-occurrence network between 'Ca. Aenigmarchaeota' and other community members. (a) The co-occurrence network based on all OTUs in all samples. The nodes represent OTUs at 95%-cutoff and edges connecting these nodes denote significant correlations (Spearman correlation; rho > 0.75 and P < 0.05) between OTUs. OTUs are colored by modularity classes. Nodes have the same size and edges have the same thickness. (b), (c), (d) and (e) are sub-networks that contain 'Ca. Aenigmarchaeota' rpS3 genes. Only nodes and edges have connections with 'Ca. Aenigmarchaeota' in the corresponding modules are shown. Modules 1 (b), 2 (c), 3 (d), and 5 (e) in (a) were detected to contain 'Ca. Aenigmarchaeota' rpS3 genes. The size of each node is proportional to the number of connections (i.e. degree). OTUs are colored by the phylum-level taxonomy. The thickness of edges denotes the Spearman rank correlation coefficients (rho values). Edges in brown show the connections between 'Ca. Aenigmarchaeota' (green circles deviated from core networks) and other members.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx
- Additionalfile5.xlsx
- Additionalfile3.xlsx
- Additionalfile2.xlsx
- Additionalfile4.xlsx