The ISME Journal (2015) 9, 447–460

www.nature.com/ismej

ORIGINAL ARTICLE

Insights into the metabolism, lifestyle and putative evolutionary history of the novel archaeal phylum 'Diapherotrites'

Noha H Youssef[1], Christian Rinke[2], Ramunas Stepanauskas[3], Ibrahim Farag[1], Tanja Woyke[2] and Mostafa S Elshahed[1]

[1]*Department of Microbiology and Molecular Genetics, Oklahoma State University, Stillwater, OK, USA;* [2]*DOE Joint Genome Institute, Walnut Creek, CA, USA and* [3]*Bigelow Laboratory for Ocean Sciences, East Boothbay, ME, USA*

The archaeal phylum 'Diapherotrites' was recently proposed based on phylogenomic analysis of genomes recovered from an underground water seep in an abandoned gold mine (Homestake mine in Lead, SD, USA). Here we present a detailed analysis of the metabolic capabilities and genomic features of three single amplified genomes (SAGs) belonging to the 'Diapherotrites'. The most complete of the SAGs, *Candidatus* 'lainarchaeum andersonii' (Cand. IA), had a small genome (~1.24 Mb), short average gene length (822 bp), one ribosomal RNA operon, high coding density (~90.4%), high percentage of overlapping genes (27.6%) and low incidence of gene duplication (2.16%). Cand. IA genome possesses limited catabolic capacities that, nevertheless, could theoretically support a free-living lifestyle by channeling a narrow range of substrates such as ribose, polyhydroxybutyrate and several amino acids to acetyl-coenzyme A. On the other hand, Cand. IA possesses relatively well-developed anabolic capabilities, although it remains auxotrophic for several amino acids and cofactors. Phylogenetic analysis suggests that the majority of Cand. IA anabolic genes were acquired from bacterial donors via horizontal gene transfer. We thus propose that members of the 'Diapherotrites' have evolved from an obligate symbiotic ancestor by acquiring anabolic genes from bacteria that enabled independent biosynthesis of biological molecules previously acquired from symbiotic hosts. 'Diapherotrites' 16S rRNA genes exhibit multiple mismatches with the majority of archaeal 16S rRNA primers, a fact that could be responsible for their observed rarity in amplicon-generated data sets. The limited substrate range, complex growth requirements and slow growth rate predicted could be responsible for its refraction to isolation.

The ISME Journal (2015) **9,** 447–460; doi:10.1038/ismej.2014.141; published online 1 August 2014

Introduction

Our view of the phylogenetic diversity and ecological distribution of members of the domain Archaea is rapidly evolving. Historically, Archaea were regarded as a collection of extremophiles that could only thrive in extreme habitats, for example, high temperature (Huber *et al.,* 1991; Jannasch *et al.,* 1992; Antoine *et al.,* 1995; Whitaker *et al.,* 2003), high salinity (Oren *et al.,* 1990; Cui *et al.,* 2010; Goh *et al.,* 2011; Inoue *et al.,* 2011), low pH (Edwards *et al.,* 2000; Dopson *et al.,* 2004), strict anaerobic conditions (Mikucki *et al.,* 2003; Sakai *et al.,* 2007) or combinations thereof (Huber *et al.,* 1989; Itoh *et al.*, 1999; Minegishi *et al.,* 2008, 2013). This notion was subsequently challenged by culture-independent diversity surveys that documented the occurrence of the Archaea in both extreme (Bond *et al.,* 2000; Orphan *et al.,* 2000; Benlloch *et al.,* 2001, 2002; Baker and Banfield, 2003) and temperate habitats, for example, soil (Bintrim *et al.,* 1997; Walsh *et al.,* 2005; Bates *et al.,* 2011; Tripathi *et al.,* 2013), marine environments (DeLong, 1992; Fuhrman *et al.*, 1992; Massana *et al.,* 1997; Karner *et al.,* 2001) and freshwater ecosystems (Lin *et al.,* 2012; Yergeau *et al.,* 2012; Berdjeb *et al.,* 2013; Bricheux *et al.,* 2013; Silveira *et al.,* 2013; Vila-Costa *et al.,* 2013). In addition to establishing the ubiquity of the Archaea on a global scale, these studies have also significantly expanded the scope of phylogenetic diversity within this domain, as many of the sequences identified represented novel, high-rank phylogenetic lineages (DeLong, 1992; Vetriani *et al.,* 1999; Takai *et al.,* 2001; Hallam *et al.,* 2004; Elkins *et al.,* 2008; Hu *et al.,* 2011; Nunoura *et al.,* 2011).

Correspondence: NH Youssef, Department of Microbiology and Molecular Genetics, Oklahoma State University, 1110S Innovation Way Drive, Stillwater, OK 74074, USA.
E-mail: noha@okstate.edu
Received 1 May 2014; revised 22 June 2014; accepted 1 July 2014; published online 1 August 2014

Environmental genomic approaches and dedicated isolation efforts have yielded valuable insights into the metabolic capabilities and ecological roles of many of these novel lineages (Hallam *et al.,* 2004; Konneke *et al.,* 2005; Elkins *et al.,* 2008; Walker *et al.,* 2010; Lloyd *et al.,* 2013). Recently, the pace of discovery and characterization of archaeal lineages has significantly accelerated, driven by recent methodological and computational advances in single cell sorting and amplification, sequencing methodologies and novel binning and assembly approaches that enable efficient genomic reconstruction from metagenomic sequence data (Dick *et al.,* 2009; Wrighton *et al.,* 2012; Rinke *et al.,* 2013; Swan *et al.,* 2013). These advances led to the identification and genomic characterization of multiple novel high-rank archaeal lineages that have previously escaped detection in 16S ribosomal RNA (rRNA) gene-based diversity surveys because of their extremely low relative abundance, limited distribution or mismatches to archaeal 16S rRNA gene primer sequences (Baker *et al.,* 2010; Nunoura *et al.,* 2011; Narasingarao *et al.,* 2012). These discoveries necessitated phylogenetic and phylogenomic-based reassessment of the taxonomic structure of the domain *Archaea* (Lake *et al.,* 1984; Brochier-Armanet *et al.,* 2008; Elkins *et al.,* 2008; Ghai *et al.,* 2011; Guy and Ettema, 2011; Williams *et al.,* 2012). The most recent and comprehensive phylogenomics-based assessment of the domain Archaea (Rinke *et al.,* 2013) combined data from 35 novel archaeal single amplified genomes (SAGs) with previously published archaeal genomes to propose a three archaeal superphyla scheme. These superphyla are the Euryarchaeota, the TACK superphylum (encompassing the Thaumarchaeota, 'Aigarchaeota', Crenarchaeota and Korachaeota as previously suggested; Guy and Ettema, 2011; Williams *et al.,* 2012) and the newly proposed DPANN superphylum.

The DPANN superphylum encompasses the 'Nanoarchaeota' (Waters *et al.,* 2003; Podar *et al.,* 2013), the only DPANN phylum with cultured representatives, as well as the candidate phyla 'Nanohaloarchaeota' (defined from metagenomic assembly (Narasingarao *et al.,* 2012) and SAGs (Ghai *et al.,* 2011) from hypersaline environments); 'Parvarchaeota' (defined from a metagenomic assembly from an acid mine drainage; Baker *et al.,* 2010), 'Aenigmarchaeota' (defined from three SAGs from Homestake mine groundwater seep (Lead, SD, USA) and the Great Boiling Spring sediments; Rinke *et al.,* 2013) and 'Diapherotrites' (defined from SAGs from Homestake mine groundwater seep; Rinke *et al.,* 2013). As such, the DPANN superphylum represents an intriguing collection of phyla with disparate physiological preferences and environmental distribution, ranging from the obligatory symbiotic and thermophilic species within the 'Nanoarchaeota', to the acidophilic candidate phylum 'Parvarachaeota' and to the non-extremophilic candidate phyla 'Aenigmarchaeota' and 'Diapherotrites'.

Here, we present a detailed analysis of the genomic, metabolic and ecological features of three SAGs belonging to the 'Diapherotrites'. We present evidence of genomic streamlining as well as limited metabolic capacities within the analyzed genomes. We also demonstrate the prevalence of cross-kingdom horizontal gene transfer (HGT) events, and argue that HGT process represents an important evolutionary mechanism that contributes to the observed genomic features, ecological distribution and proposed trophic lifestyle of members of this phylum.

Materials and methods

Origin of Diapherotrites SAGs

Candidate phylum 'Diapherotrites' (CP-'Diapherotrites') SAGs analyzed in this study were all obtained from a groundwater seep from the ceiling of Homestake Mine (Lead, SD, USA) drift at a depth of 100 m as described previously (Rinke *et al.,* 2013). Cell sorting and lysis, single-cell whole genome amplification, 16S rRNA amplicon sequencing as well as SAG sequencing and assemblies have been detailed before (Rinke *et al.,* 2013). Three CP-'Diapherotrites' SAGs were deposited under Genbank assembly IDs GCA_000402355.1, GCA_000404545.1 and GCA_000404525.1, and Integrated Microbial Genomes (IMG) SAG names SCGC_AAA011-E11, SCGC_AAA011-K09 and SC GC_AAA011-N19. The candidatus-type species for CP-'Diapherotrites' is the SAG SCGC_AAA011-E11, for which the name *Candidatus* 'Iainarchaeaum andersonii' was proposed (Rinke *et al.,* 2013). Our analysis was mainly conducted on the *Candidatus* Iainarchaeaum andersonii genome (henceforth referred to as Cand. IA), as it had the highest estimated genome completion (88.5%) among the three CP-'Diapherotrites' genomes (Supplementary Table S1), with the other two partial SAGs mainly used for confirmatory purposes.

Genome annotation, general genomic features and metabolic reconstruction

Genome functional annotation was performed using the IMG platform (http://img.jgi.doe.gov) as previously described (Rinke *et al.,* 2013). Metabolic reconstruction was conducted using both KEGG (Kyoto Encyclopedia of Genes and Genomes) and Metacyc databases (Kanehisa, 2002; Karp *et al.,* 2002). Proteases, peptidases and protease inhibitors were predicted using Blastp against the Merops database (Rawlings *et al.,* 2014). Transporters were identified by querying the genome against the TCDB (transporter classification database) (Saier *et al.*, 2014) using Blastp (Altschul *et al.,* 1990). Gene duplications were identified by running local Blastp using the proteins as both the subject and the query, where non-self hits with a similarity cutoff of $>90\%$ are considered duplicates. Overlapping genes were

identified by comparing the coordinates of the start and stop codons for all genes on the same contig. Protein subcellular localizations were identified online using the PSORTb V 3.0.2 (Yu *et al.,* 2010). Protein COG (Cluster of orthologous groups; Tatusov *et al.,* 2000) family distributions for Cand. IA and reference genomes were either obtained from the corresponding IMG genome page or, in case the genome was not available in the IMG database, were identified using the web batch conserved domain (Marchler-Bauer *et al.,* 2010) search tool (available at http://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi) against the COG database with an *E*-value threshold of 0.001.

HGT in Cand. IA genome assembly
A two-tier approach was utilized to evaluate the putative frequency of occurrence of HGT in Cand. IA genome assembly. Blastp (Altschul *et al.,* 1990) was conducted for all genes in Cand. IA as well as other representatives of the DPANN superphylum. Genes were classified according to the Blastp first hit. For in-depth evaluation of HGT within Cand. IA, 21 HGT candidate genes with bacterial Blastp first hit were chosen for further analysis according to the following two criteria: (1) protein length is longer than 100 amino acids and (2) placement on a contig with other genes involved in the same pathway but of apparent archaeal origin. In cases where all genes involved in a pathway had bacterial Blastp first hit, priority was for genes present on longer contigs with other genes of apparent archaeal origin. Homologs of genes of interest were obtained from Genbank by BLASTP of the target gene. The obtained homolog data sets were comparatively aligned using COBALT (Papadopoulos and Agarwala, 2007). A guide tree was used to select the closest relatives of the target gene. In cases where all the closest relatives belonged to bacterial phyla, archaeal counterparts were obtained from Genbank and included in the downstream analysis. A eukaryotic outgroup was also included. For phylogenetic tree construction, target genes and their homologs were aligned using ClustalW (Larkin *et al.,* 2007). Maximum likelihood trees were calculated from the alignment using the BLOSUM62 model and GAMMA approximation implemented in RaxML (Stamatakis, 2014).

Principal component analysis (PCA)
PCA was conducted to identify salient differences in genomic features and COG distributions between DPANN and other model archaeal genomes. A list of genomes, as well as features used in this analysis, is presented in Supplementary Table S2. Initially, all genomic features previously implied as determinants of genome streamlining (Giovannoni *et al.,* 2005; Lauro *et al.,* 2009; Swan *et al.,* 2013) were included in the analysis. These include: genome size, number of genes, GC content, noncoding density, number of ribosomal RNA operons, percentage of transporter proteins per genome size, COG categories distribution (that is, percentage of proteins belonging to each COG category), protein sublocalization (percentage of proteins destined to the cytoplasm, cytoplasmic membrane, cell wall and extracellular milieu) and encoded amino acid frequency. Data were then implemented in the PCA using the prcomp function in the labdsv package (Roberts, 2012) of R (R Development Core Team, 2008). A biplot was constructed using the biplot function in R, where genomes are represented as points and variables are represented as arrows pointing in the direction of maximal abundance (R Development Core Team, 2008). To simplify the biplots, all variables that showed minimal effect on the PCA biplots (for example, variables whose arrows clustered at the origin) were removed and the analysis was repeated.

Mismatches in CP-'Diapherotrites' 16S rRNA genes to archaeal primers and secondary structure prediction
16S rRNA genes from the three CP-'Diapherotrites' SAGs were aligned to reference small subunit rRNA sequences, and sites of universal archaeal primers previously utilized in culture-independent surveys were examined to identify putative mismatches and/or indel occurrences. Secondary structure prediction was achieved using the Mfold web server (Zuker, 2003), with the minimum energy structure predicted compared with the conserved secondary structure of both the *Escherichia coli* and *Methanospirilum hungatei* 16S rRNA molecules.

Mining public databases to elucidate the distribution of the 'Diapherotrites' in nature
We attempted to identify the occurrence of members of the CP-'Diapherotrites' in various ecosystems by mining metagenomic data sets in the IMG database ($n = 893$, accessed in December 2013), Sanger-generated 16S rRNA gene sequences in the nr database ($n = 53\,65\,062$ sequences, accessed in January 2014) and partial, high-throughput (pyrosequencing and Illumina)-generated archaeal 16S rRNA gene sequences in MG-RAST (Meyer *et al.*, 2008) and SRA archive (Leinonen *et al.*, 2011) ($n = 31\,972\,882$ sequences in 775 data sets generated using archaeal primers). Identification of CP-'Diapherotrites' in metagenomic data sets was conducted using the three 'Diapherotrites' SAG assemblies for anchoring metagenomic reads as previously described (Rinke *et al.,* 2013). To identify the distribution using data in Sanger- and high-throughput-generated data sets, we followed the protocols previously described in Farag *et al.* (2014).

Results

General features of 'Diapherotrites' genomes
The general features of Cand. IA partial genome compared with sequenced representatives of the

Table 1 Genomic features of *Candidatus* 'Iainarchaeum andersonii' genome compared with other DPANN superphylum members

Phylum *Genome*	*Diapherotrites* *Candidatus* *'Iainarchaeum andersonii'*	*Nanoarchaeota* *Nanoarchaeum equitans* *Kin-4*	*Parvarchaeota* *Candidatus* *'Micrarchaeum acidiphilum'* *ARMAN-2*	*Nanohaloarchaeota* *Candidatus* *'Nanosalina' sp.* *J07AB43*	*Aenigmarchaeota* *Candidatus* *'Aenigmarchaeum subterraneum'*
Genome size, Mb	1.1	0.49	~1	1.23	0.43
Genome completeness (%)	88.5	100	~99	~100	47.3
Estimated size (Mb)	1.24	0.49	1	1.23	0.91
GC %	47	31.6	47.2	44	38.6
% Noncoding	9.56	7.6	9.3	12.4	5.24
% Overlapping genes	27.63	44.1	18-19	24.9	19.42
Gene duplication[a] (%)	2.16	ND	ND	ND	ND
Average gene length, bp	822	827	892	639	766
RNA genes					
5S rRNA count	0	1	0	1	2
16S rRNA count	1	1	1	1	1
23S rRNA count	1	1	1	0	0
tRNA count	39	38	35	59	36
Number of CDS	1202	556	1033	1678	525
With function prediction	782	549	668	773	308
Without function prediction	420	7	365	905	217
Number of CDS with inteins	5	1	1	1	0
Number of CDS with IS elements	24	0	10	9	3
Number of paralog clusters	156	53	ND	198	220
% Of proteins sublocalized to					
Cytoplasm	79.9	63.3	65.6	68.9	61.5
Cytoplasmic membrane	14.6	12.3	18.9	12.8	19.23
Extracellular milieu	1.4	0.8	1.2	2.3	1.7

Abbreviations: CDS, coding sequence; IS elements, insertion sequence elements; ND, not determined; rRNA, ribosomal RNA; tRNA, transfer RNA.
[a]Percentage of genes having duplicates with protein sequence similarity $\geqslant$90%.

DPANN superphylum are shown in Table 1. Cand. IA exhibits general genomic features previously observed in DPANN genomes (Baker and Banfield, 2003; Waters *et al.,* 2003; Narasingarao *et al.,* 2012; Rinke *et al.,* 2013). It has a relatively small genome (estimated size ~1.24 Mb), short average gene length (822 bp), single ribosomal RNA operon, high coding density (~90.4%), high percentage of overlapping genes (27.6%, mean overlap 4 bp, range 1–12 bp) and very low incidence of gene duplication (2.16%).

Metabolic features of Cand. IA

Catabolic capacities. Analysis of Cand. IA genome suggests that it possesses a relatively limited catabolic potential. The genome lacks evidence for a complete glycolysis, tricarboxylic acid cycle or short-chain fatty acid or alcohol degradation. Nevertheless, the genome suggests the capability to channel few distinct substrates (valine, alanine, aspartate, ribose and polyhydroxybutyrate) into acetyl- (or acyl-) coenzyme A (CoA), and subsequently generate adenosine triphosphate (ATP) via acetyl-CoA synthase (Figure 1). Details of the pathways involved in these processes are provided in the Supplementary Text. In brief, for amino acid degradation, the genome harbors multiple transaminases that allow for the conversion of alanine, valine and aspartate to the corresponding 2-oxoacids. Oxidative decarboxylation of the 2-oxoacids to acyl-CoA can occur via pyruvate/2-oxoacid:ferredoxin oxidoreductases, or oxaloacetate-decarboxylating malate dehydrogenase (EC: 1.1.1.38) followed by conversion into the corresponding free acid via acyl-CoA synthetase with the concurrent production of 1 ATP. An incomplete tricarboxylic acid cycle could potentially replenish 2-ketoglutarate for the transamination reactions.

Cand. IA also appears to possess the capability for ribose degradation (Figure 1), where ribose could potentially be activated by ribokinase and ribose-5-phosphate pyrophosphokinase. The presence of type III ribulose-1,5-bisphosphate carboxylase (Rubisco) in Cand. IA genome suggests the employment of the novel adenosine monophosphate (AMP) metabolism pathway suggested by Aono *et al.* (2012). The combination of ribose activation and AMP salvage enzymes eventually lead to the production of 3-phosphoglycerate that feeds into the lower arm of glycolysis and subsequently leads to ATP production as described above. The process results in a net production of 2 ATP/ribose (Supplementary Text).

Another potential carbon and energy source for Cand. IA is polyhydroxybutyrate (PHB), the storage compounds produced by many bacteria under

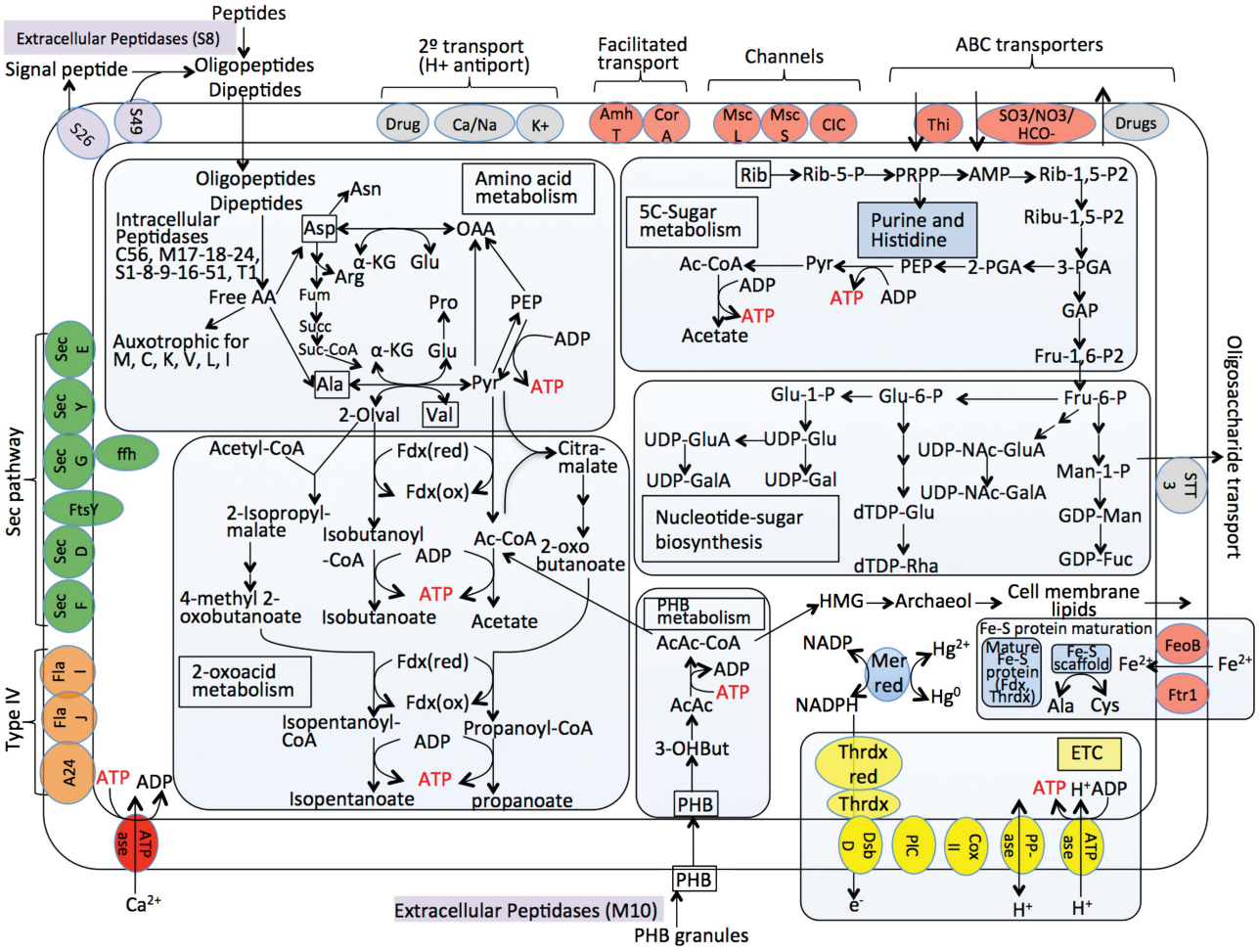

Figure 1 Metabolic reconstruction for *Candidatus* Iainarchaeum andersonii genome. Double lines surrounding the cell depict cell membrane. Possible catabolic ATP-producing pathways are shown in light blue boxes, and sites of ATP production are shown in red. Potential substrates are shown in boxes, and include the amino acids alanine, valine and aspartate, the 5C-sugar ribose and polyhydroxybutyrate. AcAc, acetoacetate; AcAc-CoA, acetoacetyl-CoA; Fdx, ferredoxin; 3-OHBut, 3-hydroxybutyrate. Electron transport chain components are shown in yellow. ATPase, V-type ATPase pump; CoxII, cytochrome oxidase subunit II; DsbD, disulfide bond oxidoreductase D; PlC, plastocyanin; PPase, inorganic pyrophosphatase; Thrdx, thioredoxin; Thrdx red, thioredoxin reductase. NADPH could potentially act as the electron donor to the ETC. Possible sites of NADPH production in the cell include mercuric reductase enzyme (Mer Red). All predicted transporters with known functions are shown. (1) Secretory pathways: components of the Sec pathway are shown in green, whereas components of type IV pili assembly are shown in orange. (2) Transporters: exporters are shown in gray including secondary antiporters, ABC transporters, STT3 the oligosaccharide exporter, whereas importers are shown in red including the facilitated transporters AmhT and CorA, the channels of MscS, MscL and ClC families, ABC transport systems for thiamine and sulfonate/nitrate/bicarbonate as well as ferrous iron transporters FeoB and Ftr1 for Fe-S assembly. AmhT, for ammonium; ClC, chloride channel families; CorA, for cobalt; MscL, large mechanosensitive; MscS, small mechanosensitive. Peptidases are shown in purple.

conditions of excess carbon. The genome encodes a PHB depolymerase downstream of, and overlapping with, a transmembrane matrixin-coding gene (peptidase family M10) (Visse and Nagase, 2003; Rawlings *et al.,* 2014), known to hydrolyze extracellular matrix components. Therefore, it appears that matrixin is used by Cand. IA to break down the protein shell of PHB granules, and the released PHB is subsequently depolymerized into β-hydroxybutyrate. The produced β-hydroxybutyrate could potentially be oxidized to acetoacetate (using a NAD-dependent dehydrogenase belonging to the GFO/IDH/MOCA family (pfam 01408)), followed by activation of acetoacetate to acetoacetyl-CoA using acyl-CoA synthase. The concerted action of acetyl-CoA C-acetyltransferase and acetyl-CoA synthetase converts acetoacetyl-CoA to two molecules of acetate with the concomitant production of ATP.

Finally, it is plausible that Cand. IA might also employ a modified electron transport chain, similar to that recently suggested for candidate division TM6 (McLean *et al.,* 2013). The chain will involve disulfide bond oxidoreductase D (DsbD), a thioredoxin-disulfide (NADPH) reductase (EC. 1.8.1.9) and thioredoxins as initial substrate (NADPH) oxidoreductases, plastocyanins as potential cytochrome equivalents and cytochrome *c* oxidase as a potential terminal oxidase. An inorganic pyrophosphatase and all subunits of V-type ATP synthase

could potentially employ the proton motive force generated across the membrane for ATP synthesis (Figure 1 and Supplementary Text).

Anabolic potential. Although the catabolic potential of Cand. IA appears to be rather limited, the genome suggests a fairly well-developed anabolic machinery. Cand. IA genome suggests the capacity for biosynthesis of multiple carbohydrates, amino acids, lipids, nucleotides as well as several cofactors (Figure 1, Table 2 and Supplementary Table S3). Cand. IA genome encodes a partial gluconeogenic pathway up to the level of fructose-6-phosphate. The genome shows evidence for synthesizing most amino acids with the exception of lysine, cysteine, methionine and branched chain amino acids (Figure 1 and Table 2). In addition, all enzymes essential for archaeol, the archaeal membrane lipid, biosynthesis from acetyl-CoA via the mevalonate pathway are encoded in the genome. Cand. IA genome shows evidence for biosynthesis of some cofactors (riboflavin, pyridoxal phosphate, nicotinate and nicotinamide, coenzyme A and ferredoxin) and nucleotides (complete evidence for *de novo* biosynthesis of pyrimidine nucleotides, partial evidence for purines). Finally, all necessary enzymes for assembly of Fe-S and maturation of Fe-S proteins are also encoded by the genome.

Transporters and extracellular peptidases. Although the number of transporters encoded by Cand. IA genome are fairly limited ($n = 54$), they appear to be able to uptake several cations and anions (Supplementary Text). Several transporters belonging to transporter families with unidentified substrates are annotated as membrane proteins with unknown function and hence might be involved in the transport of amino acids, sugars or cofactors for which Cand. IA appears to be auxotrophic. The genome also encodes for various extracellular and membrane-associated peptidases (see Supplementary Text for more details) that presumably act to cleave extracellular peptides, with the resulting oligopeptides and amino acids subsequently transported into the cell and used to supplement auxotrophies and/or ATP production.

HGT in Cand. IA

The metabolic analysis described above clearly indicates that Cand. IA genome possesses relatively limited metabolic capabilities as compared with free-living archaeal copiotrophs and oligotrophs. Nevertheless, unlike the model obligate archaeal symbionts within the 'Nanoarchaeota' ('*Nanoarchaeum equitans*' and 'Nanoarchaeota' strain Nst1; Podar *et al.,* 2013), it appears to possess pathways allowing for the independent production of ATP and the biosynthesis of multiple key metabolites. To examine whether the observed metabolic capacities within Cand. IA are because of a reductive evolutionary process of gene loss from a metabolically versatile ancestor, or to gene acquisition by a metabolically limited ancestor, we analyzed the occurrence and prevalence of HGT events within Cand. IA genome. Because of the limited genomes available representing the DPANN superphylum, identification of incidents of HGT from an archaeal donor was not feasible. Therefore, our analysis was limited to the identification of genes having non-archaeal origins within Cand. IA genome.

Cand. IA genome showed a higher percentage of proteins (25.4%, $n = 305$) with apparent bacterial Blastp first hits compared with all other archaeal genomes examined (those ranged from 1.44% in '*Nanoarchaeum equitans*' to 16.9% in *Candidatus* Micrarchaeum acidiphilum ARMAN2; Figure 2a). A disproportionately large percentage (116 out of 196) of metabolism-related genes in Cand. IA genome were of apparent bacterial origin, with the majority of these metabolic genes (86 out of 116) involved in biosynthetic pathways (Figure 2b). Maximum likelihood analysis of the phylogenetic affiliation of 21 'Diapherotrites' anabolic proteins confirmed their putative bacterial origin (Supplementary Figure S1). Detailed analysis of the impact of HGT on every metabolic pathway examined is summarized in Table 2 and Supplementary Table S3. Out of 26 metabolic pathways examined, 21 showed evidence of HGT with at least one gene in the pathway with bacterial Blastp first hit. In some cases, an entire metabolic pathway was completely bacterial in origin, for example, biosynthesis of threonine, histidine, arginine and proline. Within other pathways, a fraction of the genes were bacterial in origin; but these genes appear to mediate the critical/defining steps of the pathway, for example, PHB depolymerase in PHB degradation pathway and phosphoenolpyruvate synthase in gluconeogenesis. It is notable that many of the metabolic pathways affected by HGT are absent from '*Nanoarchaeum equitans*', the model Nanoarchaeal obligate endosymbiont (Table 2 and Supplementary Table S3). Phylogenetic analysis suggests multiple bacterial phyla as potential gene donors to the Cand. IA genome (Table 2), although within HGT acquired genes, overlapping genes as well as genes with multiple copies usually appear to have the same bacterial donor.

Collectively, the acquired genes enabled Cand. IA to: (1) synthesize alanine, threonine, arginine, proline, histidine and aromatic amino acids, purine and pyrimidine nucleotides and the cofactor pyridoxal phosphate, (2) to convert pterin to folate, (3) to uptake and activate thiamine, (4) to break down proteins extracellularly and potentially uptake the resulting oligopeptides and amino acids and (5) to utilize alanine and valine as possible C and energy sources, and to depolymerize PHB. These results demonstrate that cross-kingdom HGT events represent an important mechanism that contributes to enhancing the metabolic capacities of Cand. IA.

Table 2 Metabolic features and potential origins of metabolic genes in *Cand*. IA genome

Pathway	*No. of HGT/total*[a]	*Potential bacterial phyla donor(s)*	*Nanoarchaea*[b]
I. Biosynthetic pathways			
A. Pathways entirely bacterial in origin			
1. Amino acids			
a. Gly from Ser	1/1	Deferribacter	0/1
b. Ala from pyruvate	1/1	Thermodesulfobacteria	0/1
c. Glu from Ala and α-KG	1/1	Thermodesulfobacteria	0/1
d. Thr from Asp (missing 1)	4/4	Bacteroidetes, Acidobacteria, Spirochaetes	0/4
e. His from PRPP	8/8	Acidobacteria, Synergistetes, Proteobacteria, Firmicutes, Aminicenantes, and Parcubacteria	2/9
f. Arg from carbamoyl-P and Asp	6/6	Thermotoga, Firmicutes, Spirochaetes, Bacteroidetes and Cyanobacteria	1/6
g. Pro from Glu	3/3	Firmicutes and Aquificae	0/3
2. Cofactors			
a. B6	2/2	Armatimonadetes, and CD JS-1	0/2
b. Folate and 1C pool	7/7	Firmicutes, Thermotoga, Proteobacteria, and Deferribacter	0/7
B. Pathways partially bacterial in origin			
1. Amino acids			
a. Asn from Asp	2/5	Firmicutes and Aquificae	3/5
b. Aromatic amino acids from 7-phospho-2-dehydro-3-deoxy-D-arabino-heptonate			
i.Trp	8/12	Proteobacteria, Therodesulfobacteria, Planctomycetes,	2/12
ii. Phe	7/9	Nitrospira, Firmicutes and Hydrogenedentes	3/9
iii. Tyr			
2. Cofactors			
a. Thiamine	3/4	Firmicutes and Proteobacteria	1/4
b. Riboflavin	2/6	Firmicutes	0/6
c. Nicotinate and nicotinamide	2/6	Firmicutes and Proteobacteria	0/6
d. Pantothenate and CoA	1/4	Aminicenantes (CD-OP8)	0/4
3. Nucleotides			
a. Purine from ribose-1-P	10/12	Planctomycetes, Proteobacteria, Firmicutes, Lentisphaera, and Omnitrophica	1/12
b. Pyrimidine from glutamine and PRPP	11/18	Bacteroidetes, Proteobacteria, Cyanobacteria, Firmicutes, Planctomycetes, and Omnitrophica	2/18
4. Carbohydrates			
Gluconeogenesis to fructose-6-P	6/13	Firmicutes, Bacteroidetes, Proteobacteria, and Parcubacteria	9/13
C. Pathways entirely archaeal in origin			
a. Ser from 3-P glycerate	0/3	NA	2/3
b. Gln from Glu	0/6	NA	1/6
c. Asp from oxaloacetate	0/1	NA	0/1
II. Catabolic pathways (partially bacterial in origin)			
1. PHB degradation	2/6	Actinobacteria and Proteobacteria	2/6
2. Ribose degradation	5/9	Thermodesulfobacteria, Acidobacteria, Firmicutes, Bacteroidetes, and Parcubacteria	4/9
3. Peptidases	3/5	Actinobacteria and Firmicutes	1/5
4. Potential ETS	2/12	Proteobacteria and Firmicutes	5/12

Abbreviations: *Cand*. IA, *Candidatus* 'Iainarchaeum andersonii'; HGT, horizontal gene transfer; PHB, polyhydroxybutyrate; PRPP, 5-phospho-α-D-ribosyl 1-pyrophosphate.
[a]Number of HGT candidates/total number of genes in the pathway.
[b]Number of genes/total number of genes in the pathway with homologs in the published Nanoarchaeal genomes.

Genome architecture and COG distribution patterns in Cand. IA genome

Various genomic features and COG distribution patterns in microbial genomes have successfully been correlated to putative trophic lifestyles, for example, oligotrophy, copiotrophy (Giovannoni *et al.,* 2005; Lauro *et al.,* 2009; Swan *et al.,* 2013) as well as obligate symbiosis (Moran and Wernegreen, 2000; Shigenobu *et al.,* 2000; Akman *et al.,* 2002; Tamas *et al.,* 2002; Moran *et al.,* 2003;

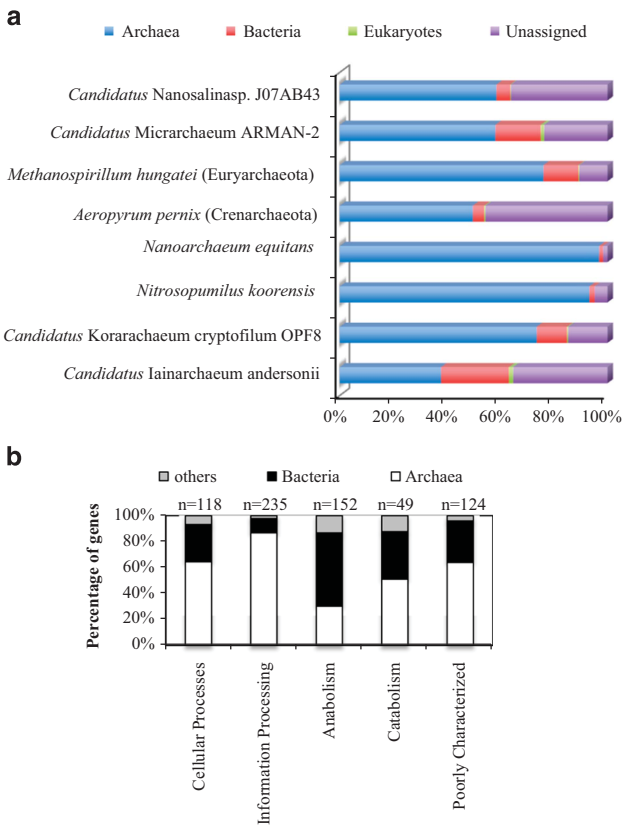

Figure 2 (**a**) Phylogenetic distribution of non-self Blastp first hits of *Candidatus* Iainarchaeum andersonii proteins compared with other archaeal phyla representatives. (**b**) Phylogenetic distribution at the domain level of non-self Blastp first hits of *Candidatus* Iainarchaeum andersonii proteins classified by metabolic category in the X axis. Total number of proteins belonging to each category are shown above each column.

Waters *et al.,* 2003; Moya *et al.,* 2008; Nikoh *et al.,* 2011; Hendry *et al.*, 2013; Podar *et al.,* 2013). In an effort to decipher the putative trophic lifestyle of Cand. IA, we used PCA to compare Cand. IA genome with those of 19 other archaeal genomes (Supplementary Table S2) encompassing obligate oligotrophs, obligate archaeal symbionts, fast-growing copiotrophs as well as the slow-growing archaeal copiotrophs (marine groups MCG and Thermoplasmatales MBG-D group thriving in C-rich ocean sediments). PCA biplot (Figure 3) showed that, in general, genomes clustered according to their trophic lifestyle into four major groups: fast-growing copiotrophs, slow-growing copitrophs, obligate oligotrophs and obligate symbionts, with the later being highly divergent and clustering away from other genomes. This was expected as archaeal symbionts exhibit a dramatic reduction in metabolism-related and an expansion in information processing-related gene families (Waters *et al.,* 2003; Podar *et al.,* 2013).

Cand. IA genome did not cluster with any of the above groups in the PCA biplot, but rather showed a distinct position between the obligate nanoarchaeal symbionts and the archaeal oligotrophs. This position is a reflection of the overrepresentation of replication-related, post-translational modification-related and nucleotide metabolism-related proteins compared with other genomes, as well as the overrepresentation of translation-related proteins compared with all other genomes with the exception of 'Nanoarchaeota'. The position of Cand. IA in the same quadrant with archaeal symbionts and oligotrophs is a reflection of the shared streamlining genomic characteristics, for example, smaller genome size and significantly lower percentage of cell membrane proteins, when compared with copiotrophs. However, salient differences exist between Cand. IA and typical oligotrophs and obligate symbionts. Compared with '*Nitrosopumilis maritimus*' genome (a model archaeal oligotroph), Cand. IA genome has an overrepresentation of replication, translation, transcription, nucleotide metabolism, intracellular trafficking and secretion and cell wall biogenesis COGs as well as proteins destined to the cell wall, and an underrepresentation of signal transduction, defense mechanisms, energy, amino acid, coenzyme, inorganic ion and secondary metabolism COGs as well as transporters and extracellular proteins (Figure 3 and Supplementary Table S2). Similarly, compared with '*Nanoarchaeum equitans*' genome (a model archaeal obligate symbiont), Cand. IA genome has a larger genome, an overrepresentation of energy, amino acid, nucleotide, carbohydrate, coenzyme, lipid and inorganic ion metabolism, cell wall biogenesis and signal transduction COGs, and an underrepresentation of translation, replication and post-translational modification COGs. Interestingly, many of these defining features between archaeal symbionts and Cand. IA could be mediated by the acquisition of genes via HGT as described above. The disparate position of all DPANN genomes analyzed is striking, and underscores the high level of diversity in genomic architecture, metabolic potential and trophic lifestyle within the DPANN superphylum (Rinke *et al.,* 2013).

Global distribution of the 'Diapherotrites'

The 16S rRNA genes obtained from the three available 'Diapherotrites' SAG genome assemblies were shorter (average length is 1312 bp) than other archaeal counterparts, mainly because of the absence of bases corresponding to 1–20 and 1381–1540 in *Methanobacterium formimicum* (Gutell *et al.,* 1985). Therefore, universal 16S rRNA gene primers targeting these regions (for example, A1F, U1406R and U1510R) (Baker *et al.,* 2003) would theoretically fail to identify 'Diapherotrites' members. Furthermore, within these shortened 16S rRNA genes, mismatches to almost all known archaeal-specific or universal 16S rRNA gene primers were identified in all three available 'Diapherotrites' SAG genome assemblies, with the exception of 109F, 515F and 534R (Table 3).

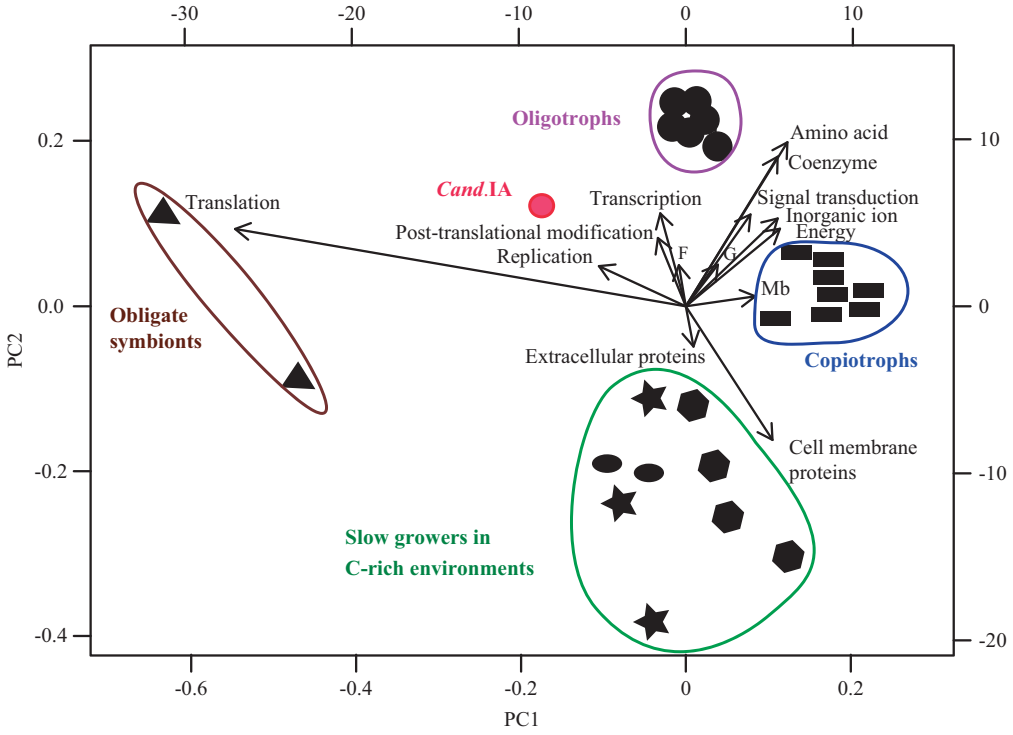

Figure 3 PCA biplot of the genomic features and COG category distribution in the genomes compared. Genomes are represented by symbols according to their trophic lifestyle. A list of genomes, as well as features used in this analysis, and trophic lifestyles are presented in Supplementary Table S2. Arrows represent genomic features or COG categories used for comparison. The arrow directions follow the maximal abundance, and their lengths are proportional to the maximal rate of change between genomes. The first two components explained 75% of variation. Obligate oligotrophs of the Thaumarchaeota (depicted by circles) clustered together because of abundance of amino acid, nucleotide and coenzyme metabolism-related as well as post-translational modification and transcription-related proteins, the copiotrophs (depicted by rectangles) clustered together mainly because of their large genome sizes and a higher percentage of membrane proteins relative to other groups, and the slow-growing copiotrophs of the Thermoplasmatales (depicted by hexagons) clustered together because of the expansion of membrane as well as extracellular proteins consistent with previous reports of a higher percentage of membrane transporters and extracellular peptidases in those genomes (Lloyd *et al.,* 2013). Parvarchaeota genomes (depicted by stars) and Nanohaloarchaeota genomes (depicted by ovals) also clustered close to the slow-growing copiotrophs. Finally, the two obligate symbionts of Nanoarchaeota (depicted by triangles) clustered together away from all other genomes mainly because of expansion of translation-related proteins. *Candidatus* Iainarchaeum andersonii genome is represented by a red circle and has an intermediate position in the plot. F, COG category nucleotide metabolism; G, COG category carbohydrate metabolism; Mb, genome size.

With mismatches to known primers, 'Diapherotrites' sequences would theoretically be missed in cultivation-independent PCR-based surveys. Indeed, comparison of 'Diapherotrites' to Sanger-generated, near-full-length 16S rRNA genes deposited in GenBank nr database failed to identify any 'Diapherotrites'-related 16S rRNA gene sequences. In addition, within a 16S rRNA data set generated using primer pair 926wF and 1392R from the same sample from which the three 'Diapherotrites' SAGs were obtained, no 'Diapherotrites' 16S rRNA gene sequences were amplified (Supplementary Table S4). Furthermore, within a collection of 31 972 882 archaeal next-generation sequences spanning 58 habitats and 775 data sets, only 66 sequences were confidently assigned to the 'Diapherotrites' phylum from 3 different studies, where they were identified in a paddy soil (Feng *et al.,* 2013), three distinct soils (Portillo *et al.,* 2013) and wastewater treatment plant (Vishnivetskaya *et al.,* 2013).

Finally, we used phylogenetic anchoring to identify the presence of members of the 'Diapherotrites' in publicly available metagenomic data sets ($n = 893$). The 'Diapherotrites' were only identified in a handful of environments (11 out of 893 analyzed). These include Amazon forest soil, mangrove sediment on Isabella Island, Sakinaw lake, Etoliko lagoon sediment and Kolumbo Volcano red mat (Supplementary Figure S2). Within these studies, the 'Diapherotrites' were always identified as an extremely minor fraction of the community ($<0.006\%$ of anchored metagenomic reads).

Discussion

In this study, we present a detailed analysis of the metabolic capabilities and genomic features of three SAGs belonging to the recently proposed archaeal phylum 'Diapherotrites', as well as a survey of the putative distribution of members of this phylum using database-mining approaches. Our detailed genomic analysis of 'Diapherotrites' SAG Cand. IA uncovers evidence for genome streamlining:

Table 3 Mismatches of 'Diapherotrites' 16S rRNA gene to universal archaeal primers[a]

Primer name	*Primer sequence (5′–3′)*	*Sequence in Diapherotrites (5′–3′)*
A109F	ACKGCTCAGTAACACGT	ACKGCTCAGTAACACGT
A333F	TCCAGGCCCTACGGG	**C**CC**TA**GCCCTA**M**GGG
AB341F	CCTACGGGRSGCAGCAG	CCTA**M**GGGR**T**GCAGCAG
A344F	ACGGGGTGCAGCAGGCGCGA	A**M**GGGGTGCAGCAG**KYRS**GA
U515F	GTGCCAGCMGCCGCGGTAA	GTGCCAGCMGCCGCGGTAA
U519F	CAGCMGCCGCGGTAATWC	CAGCMGCCGCGGTAATWC
UA571F	GCYTAAAGSRICCGTAGC	GCYTAAAGSRI**YY**GTAGC
UA751F	CCGACGGTGAGRGRYGAA	C**T**GACGGTGAGR**R**RYGAA
AB779F	GCRAASSGGATTAGATACCC	GCRA**C**SSGGATTAGATACCC
EB787F	ATTAGATACCCTGGTA	ATTAGATACCC**GGT**TA
AB787F	ATTAGATACCCGGGTA	ATTAGATACCCGG**T**TA
Ab789F	TAGATACCCSSGTAGTCC	TAGATACCCSS**T**TAGTCC
AB906F	GAAACTTAAAKGAATTG	GAAACTTAA**W**KGAATTG
A1040F	GAGAGGWGGTGCATGGCC	GAGAGGWG**Y**TGT**T**ATGG**Y**C
U1053F	GCATGGCYGYCGTCAG	G**T**ATGGCY**A**Y**YKK**CAG
A1098F	GGCAACGAGCGMGACCC	GGCAACG**G**GCGMGACC**Y**
AB127R	CCACGTGTTACTSAGC	C**B**ACGTGTTACTSAGC
A348R	CCCCGTAGGGCCYGG	CCCC**K**TAGGGC**TA**GG
U534R	GWATTACCGCGGCKGCTG	GTATTACCGCGGCGGCTG
U529R	ACCGCGGCKGCTGGC	ACCGCGGCGGCTGGC
AB909R	TTTCAGYCTTGCGRCCGTAC	TTTCA**C**YCTTGCGR**GYR**TAC
A927R	CCCGCCAATTCCTTTAAGTTTC	CCCGCCAATTCC**W**TTAAGTTTC
A926R	CCGTCAATTCCTTTRAGTTT	CCG**C**CAATTCC**W**TTRAGTTT
A934R	GTGCTCCCCCGCCAATTCCT	GTG**MGCY**CCCGCCAATTCC**W**
A976R	YCCGGCGTTGAMTCCAATT	YC**Y**GGCGTTG**TR**TCC**R**ATT
A1115R	GGGTCTCGCTCGTTG	**R**GGTCTCGC**Y**CGTTG
UA1204R	TTMGGGGCATRCIKACCT	TTMGGG**CY**ATRCIKA**Y**CT

Mismatches in 'Diapherotrites' sequences are shown in bold.
[a]Using the 16S rRNA gene sequences of the three Diapherotrites single amplified genomes (SAGs).

prevalence of HGT events, especially in metabolism-related genes; limited catabolic capabilities with only few substrates that could potentially be utilized for ATP production; and a limited representation of members of this phylum in amplicon-generated and metagenomic data sets.

Many of the genomic streamlining features observed in Cand. IA genome, such as small genome size, small intergenic regions, low incidences of gene duplication and low number of rRNA operons, have been associated with specific trophic lifestyles, mainly oligotrophy and obligate symbiosis (Giovannoni *et al.,* 2005; Lauro *et al.,* 2009; Walker *et al.,* 2010; Grote *et al.,* 2012; Swan *et al.,* 2013), where they appear to be a reflection of the accessibility of nutrients, as well as the occurrence of genetic drift in obligate symbionts (Mira *et al.,* 2001; Wernegreen, 2002; Giovannoni *et al.,* 2005; Oakeson *et al.,* 2014). However, detailed comparative analysis of the metabolism and genomic features of Cand. IA revealed salient differences when compared with the genomes of model archaeal obligate symbionts and oligotrophs.

Compared with the genome of the model archaeal obligate symbiont '*Nanoarchaeum equitans*', Cand. IA genome possesses multiple catabolic abilities that allow for the production of ATP from few substrates (valine, alanine, aspartate, ribose and PHB) (Figure 1, Table 2, and Supplementary Text). Such capabilities are completely absent from '*N. equitans*' genome (Waters *et al.,* 2003). More importantly, Cand. IA possesses an anabolic machinery that allows for the biosynthesis of multiple amino acids, nucleotides and cofactors (Figure 1, Table 2 and Supplementary Text); a feature that is absent in '*N. equitans*' because of its dependence on its host for supplying such metabolites (Waters *et al.,* 2003).

In contrast to archaeal oligotrophs that have numerous transport capabilities, a well-developed essential biosynthetic machinery as well as complete central metabolic pathways (Walker *et al.,* 2010; Nunoura *et al.,* 2011), Cand. IA exhibits lower transport capabilities (Supplementary Table S2), higher level of auxotrophy and incomplete and/or less developed pathways (for example, respiratory chain, tricarboxylic acid cycle and pentose phosphate pathway). Furthermore, the catabolic capabilities of Cand. IA appear to be geared toward utilizing substrates that are more common in non-oligotrophic habitats. For example, ribose, being a lysis product of RNA, is presumably more available in non-oligotrophic environments characterized by higher rates of cell turnover and lysis. Indeed, both the transporters and the carbohydrate utilization patterns of the SAR11 clade comprising the oligotrophic ocean bacteria suggest the inability to take-up and utilize carbohydrates (including ribose) as a carbon and energy source (Jiao and Zheng, 2011). Similarly, PHBs are storage molecules produced by several bacterial species in response to either an excess of carbon or a limitation of another nutrient,

for example, nitrogen or phosphorous (Jendrossek and Handrick, 2002), and hence would be expected to exist in C-rich environments (Jendrossek and Handrick, 2002).

Analysis of the global ecological distribution of members of the 'Diapherotrites' identified its presence in only a few environments (Supplementary Figure S2). However, this observed pattern of paucity of 'Diapherotrites' sequences in either amplicon-generated or metagenomic data sets could be influenced by the mismatches identified to the most commonly utilized 16S rRNA gene primers (Table 3), or the limited number of 'Diapherotrites' reference sequences available to serve as substrates for phylogenetic anchoring analysis, respectively. Nevertheless, examination of the origin and trophic status of habitats where the 'Diapherotrites' were identified suggest their preference to non-oligotrophic environments (for example, microbial mats, high productivity forests, wastewater treatment plants (Vishnivetskaya *et al.,* 2013) and soils (Feng *et al.,* 2013; Portillo *et al.,* 2013)).

We argue that the observed metabolic capacities, genomic features and ecological distribution as well as the observed high proportion of genes involved in cross-kingdom HGT (Table 2 and Supplementary Table S3) could be explained by a conceptual model where gene acquisition plays an important role in shaping the evolutionary history of Cand. IA. Specifically, we argue that this acquisition process is mediating the putative transition of Cand. IA from a symbiotic ancestor with a streamlined genome and extremely limited metabolic capabilities to a free-living microorganism, capable of ATP production (although from a limited number of substrates), as well as biosynthesis of multiple amino acids, nucleotides and cofactors, although it remains auxotrophic to other several cellular building blocks. Indeed, most of the key differences in genome architecture between '*N. equitans*' and Cand. IA (Figure 3) could be brought about by the presence of additional genes of apparent bacterial origin in the genome assembly. Theoretically, removal of genes of bacterial origin from the Cand. IA genome would produce a genome assembly with features and metabolic capacities very similar to the genome '*N. equitans*' (Supplementary Figure S3).

Although the role of HGT in conferring specific capabilities to recipient prokaryotic species, for example, antibiotic resistance or heavy metal resistance (Andam *et al.,* 2011; Navarro *et al.,* 2013), has long been recognized, the impact of HGT on prokaryotic evolutionary history and its potential role in organismal transition to new habitats and lifestyles has received less attention. The acquisition of bacterial genes was recently proposed as a driver of Halobacteriales evolution from a methanogenic ancestor (Nelson-Sathi *et al.,* 2012). Within the eukaryotes, HGT has been shown to be important in the development of thermoacidophily and subsequent adaptation of the red algae *Galdieria sulphuraria* to hot acidic habitats (Qiu *et al.,* 2013; Schönknecht *et al.,* 2013), as well as the adaptation of gut fungi (Neocallimastigomycota) to the strict anaerobic, eutrophic and plant biomass-rich habitat in the herbivorous gut (Youssef *et al.,* 2013). Additional research to provide a more detailed understanding of the impact of such processes on microbial (especially prokaryotic) evolution is certainly warranted.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

This work was supported by the National Science Foundation Microbial Observatories Program (Grant EF0801858).

References

Akman L, Yamashita A, Watanabe H, Oshima K, Shiba T, Hattori M *et al.* (2002). Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. *Nat Genet* **32**: 402–407.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Andam CP, Fournier GP, Gogarten JP. (2011). Multilevel populations and the evolution of antibiotic resistance through horizontal gene transfer. *FEMS Microbiol Rev* **35**: 756–767.

Antoine E, Guezennec J, Meunier JR, Lesongeur F, Barbier G. (1995). Isolation and characterization of extremely thermophilic archaebacteria related to the genus *Thermococcus* from deep-sea hydrothermal guaymas basin. *Curr Microbiol* **31**: 186–192.

Aono R, Sato T, Yano A, Yoshida S, Nishitani Y, Miki K *et al.* (2012). Enzymatic characterization of AMP phosphorylase and ribose-1,5-bisphosphate isomerase functioning in an Archaeal AMP metabolic pathway. *J Bacteriol* **194**: 6847–6855.

Baker BJ, Banfield JF. (2003). Microbial communities in acid mine drainage. *FEMS Microbiol Ecol* **44**: 139–152.

Baker BJ, Comolli LR, Dick GJ, Hauser LJ, Hyatt D, Dill BD *et al.* (2010). Enigmatic, ultrasmall, uncultivated Archaea. *Proc Natl Acad Sci USA* **107**: 8806–8811.

Baker GC, Smith JJ, Cowan DA. (2003). Review and re-analysis of domain-specific 16S primers. *J Microbiol Meth* **55**: 541–555.

Bates ST, Berg-Lyons D, Caporaso JG, Walters WA, Knight R, Fierer N. (2011). Examining the global distribution of dominant archaeal populations in soil. *ISME J* **5**: 908–917.

Benlloch S, Acinas SG, Antón J, López-López A, Luz SP, Rodríguez-Valera F. (2001). Archaeal biodiversity in crystallizer ponds from a solar Saltern: culture versus PCR. *Microb Ecol* **41**: 12–19.

Benlloch S, Lopez-Lopez A, Casamayor EO, Ovreas L, Goddard V, Daae FL *et al.* (2002). Prokaryotic genetic diversity throughout the salinity gradient of a coastal solar saltern. *Environ Microbiol* **4**: 349–360.

Berdjeb L, Pollet T, Chardon C, Jacquet S. (2013). Spatio-temporal changes in the structure of archaeal communities in two deep freshwater lakes. *FEMS Microbiol Ecol* **86**: 215–230.

Bintrim SB, Donohue TJ, Handelsman J, Roberts GP, Goodman RM. (1997). Molecular phylogeny of Archaea from soil. *Proc Natl Acad Sci USA* **94**: 277–282.

Bond PL, Smriga SP, Banfield JF. (2000). Phylogeny of microorganisms populating a thick, subaerial, predominantly lithotrophic biofilm at an extreme acid mine drainage site. *Appl Environ Microbiol* **66**: 3842–3849.

Bricheux G, Morin L, Le Moal G, Coffe G, Balestrino D, Charbonnel N *et al.* (2013). Pyrosequencing assessment of prokaryotic and eukaryotic diversity in biofilm communities from a French river. *Microbiol Open* **2**: 402–414.

Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P. (2008). Mesophilic crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nat Rev Micro* **6**: 245–252.

Cui H-L, Gao X, Sun F-F, Dong Y, Xu X-W, Zhou Y-G *et al.* (2010). *Halogranum rubrum* gen. nov., sp. nov., a halophilic archaeon isolated from a marine solar saltern. *Int J Syst Evol Microbiol* **60**: 1366–1371.

DeLong EF. (1992). Archaea in coastal marine environments. *Proc Natl Acad Sci USA* **89**: 5685–5689.

Dick G, Andersson A, Baker B, Simmons S, Thomas B, Yelton AP *et al.* (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biol* **10**: R85.

Dopson M, Baker-Austin C, Hind A, Bowman JP, Bond PL. (2004). Characterization of *Ferroplasma* Isolates and *Ferroplasma acidarmanus* sp. nov., extreme acidophiles from acid mine drainage and industrial bioleaching environments. *Appl Environ Microbiol* **70**: 2079–2088.

Edwards KJ, Bond PL, Gihring TM, Banfield JF. (2000). An Archaeal iron-oxidizing extreme acidophile important in acid mine drainage. *Science* **287**: 1796–1799.

Elkins JG, Podar M, Graham DE, Makarova KS, Wolf Y, Randau L *et al.* (2008). A Korarchaeal genome reveals insights into the evolution of the Archaea. *Proc Natl Acad Sci USA* **105**: 8102–8107.

Farag IF, Davis JP, Youssef NH, Elshahed MS. (2014). Global patterns of abundance, diversity and community structure of the Aminicenantes (candidate phylum OP8). *PLoS One* **9**: e92139.

Feng Y, Lin X, Yu Y, Zhang H, Chu H, Zhu J. (2013). Elevated ground-level O_3 negatively influences paddy methanogenic archaeal community. *Sci Rep* **3**: 3193.

Fuhrman JA, McCallum K, Davis AA. (1992). Novel major archaebacterial group from marine plankton. *Nature* **356**: 148–149.

Ghai R, Pašić L, Fernández AB, Martin-Cuadrado A-B, Mizuno CM, McMahon KD *et al.* (2011). New abundant microbial groups in aquatic hypersaline environments. *Sci Rep* **1**: 135.

Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D *et al.* (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**: 1242–1245.

Goh F, Jeon YJ, Barrow K, Neilan BA, Burns BP. (2011). Osmoadaptive strategies of the archaeon *Halococcus hamelinensis* Isolated from a hypersaline stromatolite environment. *Astrobiology* **11**: 529–536.

Grote J, Thrash JC, Huggett MJ, Landry ZC, Carini P, Giovannoni SJ *et al.* (2012). Streamlining and core genome conservation among highly divergent members of the SAR11 clade. *mBio* **3**: e00252–12.

Gutell RR, Weiser B, Woese CR, Noller HF. (1985). Comparative anatomy of 16-S-like ribosomal RNA. *Prog Nucleic Acid Res Mol Biol* **32**: 155–216.

Guy L, Ettema TJ. (2011). The archaeal 'TACK' superphylum and the origin of eukaryotes. *Trends Microbiol* **19**: 580–587.

Hallam SJ, Putnam N, Preston CM, Detter JC, Rokhsar D, Richardson PM *et al.* (2004). Reverse methanogenesis: testing the hypothesis with environmental genomics. *Science* **305**: 1457–1462.

Hendry TA, de Wet JR, Dunlap PV. (2013). Genomic signatures of obligate host dependence in the luminous bacterial symbiont of a vertebrate. *Environ Microbiol*; e-pub ahead of print 10 October 2013; doi:10.1111/1462-2920.12302.

Hu A, Jiao N, Zhang R, Yang Z. (2011). Niche partitioning of marine group I Crenarchaeota in the euphotic and upper mesopelagic zones of the East China Sea. *Appl Environ Microbiol* **77**: 7469–7478.

Huber G, Spinnler C, Gambacorta A, Stetter KO. (1989). *Metallosphaera sedula* gen, and sp. nov. Represents a new genus of aerobic, metal-mobilizing, thermoacidophilic Archaebacteria. *Syst Appl Microbiol* **12**: 38–47.

Huber G, Huber R, Jones BE, Lauerer G, Neuner A, Segerer A *et al.* (1991). Hyperthermophilic archaea and bacteria occurring within Indonesian hydrothermal areas. *Syst Appl Microbiol* **14**: 397–404.

Inoue K, Itoh T, Ohkuma M, Kogure K. (2011). *Halomarina oriensis* gen. nov., sp. nov., a halophilic archaeon isolated from a seawater aquarium. *Int J Syst Evol Microbiol* **61**: 942–846.

Itoh T, Suzuki K-i, Sanchez PC, Nakase T. (1999). *Caldivirga maquilingensis* gen. nov., sp. nov., a new genus of rod-shaped crenarchaeote isolated from a hot spring in the Philippines. *Int J Syst Bacteriol* **49**: 1157–1163.

Jannasch HW, Wirsen CO, Molyneaux SJ, Langworthy TA. (1992). Comparative physiological studies on hyperthermophilic Archaea isolated from deep-sea hot vents with emphasis on *Pyrococcus* strain GB-D. *Appl Environ Microbiol* **58**: 3472–3481.

Jendrossek D, Handrick R. (2002). Microbial degradation of polyhydroxyalkanoates. *Annu Rev Microbiol* **56**: 403–432.

Jiao N, Zheng Q. (2011). The microbial carbon pump: from genes to ecosystems. *Appl Environ Microbiol* **77**: 7439–7444.

Kanehisa M. (2002). The KEGG database. *Novartis Found Symp* **247**: 91–101. discussion 101-103, 119-128, 244-152.

Karner MB, DeLong EF, Karl DM. (2001). Archaeal dominance in the mesopelagic zone of the Pacific Ocean. *Nature* **409**: 507–510.

Karp PD, Riley M, Paley SM, Pellegrini-Toole A. (2002). The MetaCyc Database. *Nucleic Acids Res* **30**: 59–61.

Konneke M, Bernhard AE, de la Torre JR, Walker CB, Waterbury JB, Stahl DA. (2005). Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* **437**: 543–546.

Lake JA, Henderson E, Oakes M, Clark MW. (1984). Eocytes: a new ribosome structure indicates a

kingdom with a close relationship to eukaryotes. *Proc Natl Acad Sci USA* **81**: 3786–3790.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H *et al.* (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947–2948.

Lauro FM, McDougald D, Thomas T, Williams TJ, Egan S, Rice S *et al.* (2009). The genomic basis of trophic strategy in marine bacteria. *Proc Natl Acad Sci USA* **106**: 15527–15533.

Leinonen R, Sugawara H, Shumway M. (2011). The sequence read archive. *Nucleic Acids Res* **39**: D19–D21.

Lin X, McKinley J, Resch CT, Kaluzny R, Lauber CL, Fredrickson J *et al.* (2012). Spatial and temporal dynamics of the microbial community in the Hanford unconfined aquifer. *ISME J* **6**: 1665–1676.

Lloyd KG, Schreiber L, Petersen DG, Kjeldsen KU, Lever MA, Steen AD *et al.* (2013). Predominant archaea in marine sediments degrade detrital proteins. *Nature* **496**: 215–218.

Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C *et al.* (2010). CDD: a conserved domain database for the functional annotation of proteins. *Nucleic Acids Res* **39**: D225–D229.

Massana R, Murray AE, Preston CM, DeLong EF. (1997). Vertical distribution and phylogenetic characterization of marine planktonic Archaea in the Santa Barbara Channel. *Appl Environ Microbiol* **63**: 50–56.

McLean JS, Lombardo M-J, Badger JH, Edlund A, Novotny M, Yee-Greenbaum J *et al.* (2013). Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. *Proc Natl Acad Sci USA* **110**: E2390–E2399.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M *et al.* (2008). The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinfomatics* **9**: 386.

Mikucki JA, Liu Y, Delwiche M, Colwell FS, Boone DR. (2003). Isolation of a methanogen from deep marine sediments that contain methane hydrates, and description of *Methanoculleus submarinus* sp. nov. *Appl Environ Microbiol* **69**: 3311–3316.

Minegishi H, Mizuki T, Echigo A, Fukushima T, Kamekura M, Usami R. (2008). Acidophilic haloarchaeal strains are isolated from various solar salts. *Saline Systems* **4**: 16.

Minegishi H, Yamauchi Y, Echigo A, Shimane Y, Kamekura M, Itoh T *et al.* (2013). *Halarchaeum nitratireducens* sp. nov., a moderately acidophilic haloarchaeon isolated from commercial sea salt. *Int J Syst Evol Microbiol* **63**: 4202–4206.

Mira A, Ochman H, Moran NA. (2001). Deletional bias and the evolution of bacterial genomes. *Trends Genet* **17**: 589–596.

Moran NA, Wernegreen JJ. (2000). Lifestyle evolution in symbiotic bacteria: insights from genomics. *Trends Ecol Evol* **15**: 321–326.

Moran NA, Plague GR, Sandstrom JP, Wilcox JL. (2003). A genomic perspective on nutrient provisioning by bacterial symbionts of insects. *Proc Natl Acad Sci USA* **100**: 14543–14548.

Moya A, Pereto J, Gil R, Latorre A. (2008). Learning how to live together: genomic insights into prokaryote-animal symbioses. *Nat Rev Genet* **9**: 218–229.

Narasingarao P, Podell S, Ugalde JA, Brochier-Armanet C, Emerson JB, Brocks JJ *et al.* (2012). De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. *ISME J* **6**: 81–93.

Navarro CA, von Bernath D, Jerez CA. (2013). Heavy metal resistance strategies of acidophilic bacteria and their acquisition: importance for biomining and bioremediation. *Biol Res* **46**: 363–371.

Nelson-Sathi S, Dagan T, Landan G, Janssen A, Steel M, McInerney JO *et al.* (2012). Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proc Natl Acad Sci USA* **109**: 20537–20542.

Nikoh N, Hosokawa T, Oshima K, Hattori M, Fukatsu T. (2011). Reductive evolution of bacterial genome in insect gut environment. *Genome Biol Evol* **3**: 702–714.

Nunoura T, Takaki Y, Kakuta J, Nishi S, Sugahara J, Kazama H *et al.* (2011). Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. *Nucleic Acids Res* **39**: 3204–3223.

Oakeson KF, Gil R, Clayton AL, Dunn DM, von Niederhausern AC, Hamil C *et al.* (2014). Genome degeneration and adaptation in a nascent stage of symbiosis. *Genome Biol Evol* **6**: 76–93.

Oren A, Ginzburg M, Ginzburg BZ, Hochstein LI, Volcani BE. (1990). *Haloarcula marismortui* (Volcani) sp. nov., nom. rev., an extremely halophilic bacterium from the Dead sea. *Int J Syst Bacteriol* **40**: 209–210.

Orphan VJ, Taylor LT, Hafenbradl D, Delong EF. (2000). Culture-dependent and culture-independent characterization of microbial assemblages associated with high-temperature petroleum reservoirs. *Appl Environ Microbiol* **66**: 700–711.

Papadopoulos JS, Agarwala R. (2007). COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics* **23**: 1073–1079.

Podar M, Makarova K, Graham D, Wolf Y, Koonin E, Reysenbach A-L. (2013). Insights into archaeal evolution and symbiosis from the genomes of a nanoarchaeon and its inferred crenarchaeal host from Obsidian Pool, Yellowstone National Park. *Biol Direct* **8**: 9.

Portillo MC, Leff JW, Lauber CL, Fierer N. (2013). Cell size distributions of soil bacterial and archaeal taxa. *Appl Environ Microbiol* **79**: 7610–7617.

Qiu H, Price DC, Weber APM, Reeb V, Chan Yang E, Lee JM *et al.* (2013). Adaptation through horizontal gene transfer in the cryptoendolithic red alga Galdieria phlegrea. *Curr Biol* **23**: R865–R866.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria.

Rawlings ND, Barrett AJ, Bateman A. (2014). MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res* **42**: D503–D509.

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F *et al.* (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**: 431–437.

Roberts DW. (2012). labdsv: Ordination and Multivariate Analysis for Ecology, R package version 1.5-0. http://cran.r-project.org/web/packages/labdsv/.

Saier MH, Reddy VS, Tamang DG, Västermark Å. (2014). The transporter classification database. *Nucleic Acids Res* **42**: D251–D258.

Sakai S, Imachi H, Sekiguchi Y, Ohashi A, Harada H, Kamagata Y. (2007). Isolation of key methanogens for global methane emission from rice paddy fields: a novel isolate affiliated with the clone cluster rice cluster I. *Appl Environ Microbiol* **73**: 4326–4331.

Schönknecht G, Chen W-H, Ternes CM, Barbier GG, Shrestha RP, Stanke M *et al.* (2013). Gene transfer from Bacteria and Archaea facilitated evolution of an extremophilic eukaryote. *Science* **339**: 1207–1210.

Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H. (2000). Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature* **407**: 81–86.

Silveira CB, Cardoso AM, Coutinho FH, Lima JL, Pinto LH, Albano RM *et al.* (2013). Tropical aquatic archaea show environment-specific community composition. *PLoS One* **8**: e76321.

Stamatakis A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.

Swan BK, Tupper B, Sczyrba A, Lauro FM, Martinez-Garcia M, González JM *et al.* (2013). Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc Natl Acad Sci USA* **110**: 11463–11468.

Takai K, Moser DP, DeFlaun M, Onstott TC, Fredrickson JK. (2001). Archaeal diversity in waters from deep South African gold mines. *Appl Environ Microbiol* **67**: 5750–5760.

Tamas I, Klasson L, Canbäck B, Näslund AK, Eriksson A-S, Wernegreen JJ *et al.* (2002). 50 million years of genomic stasis in endosymbiotic bacteria. *Science* **296**: 2376–2379.

Tatusov RL, Galperin MY, Natale DA, Koonin EV. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**: 33–36.

Tripathi BM, Kim M, Lai-Hoe A, Shukor NAA, Rahim RA, Go R *et al.* (2013). pH dominates variation in tropical soil archaeal diversity and community structure. *FEMS Microbiol Ecol* **86**: 303–311.

Vetriani C, Jannasch HW, MacGregor BJ, Stahl DA, Reysenbach A-L. (1999). Population structure and phylogenetic characterization of marine benthic Archaea in deep-sea sediments. *Appl Environ Microbiol* **65**: 4375–4384.

Vila-Costa M, Barberan A, Auguet JC, Sharma S, Moran MA, Casamayor EO. (2013). Bacterial and archaeal community structure in the surface microlayer of high mountain lakes examined under two atmospheric aerosol loading scenarios. *FEMS Microbiol Ecol* **84**: 387–397.

Vishnivetskaya TA, Fisher LS, Brodie GA, Phelps TJ. (2013). Microbial communities involved in biological ammonium removal from coal combustion wastewaters. *Microb Ecol* **66**: 49–59.

Visse R, Nagase H. (2003). Matrix metalloproteinases and tissue inhibitors of metalloproteinases: structure, function, and biochemistry. *Circ Res* **92**: 827–839.

Walker CB, de la Torre JR, Klotz MG, Urakawa H, Pinel N, Arp DJ *et al.* (2010). Nitrosopumilus maritimus genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. *Proc Natl Acad Sci USA* **107**: 8818–8823.

Walsh DA, Papke RT, Doolittle WF. (2005). Archaeal diversity along a soil salinity gradient prone to disturbance. *Environ Microbiol* **7**: 1655–1666.

Waters E, Hohn MJ, Ahel I, Graham DE, Adams MD, Barnstead M *et al.* (2003). The genome of Nanoarchaeum equitans: Insights into early archaeal evolution and derived parasitism. *Proc Natl Acad Sci USA* **100**: 12984–12988.

Wernegreen JJ. (2002). Genome evolution in bacterial endosymbionts of insects. *Nat Rev Genet* **3**: 850–861.

Whitaker RJ, Grogan DW, Taylor JW. (2003). Geographic barriers isolate endemic populations of hyperthermophilic archaea. *Science* **301**: 976–978.

Williams TA, Foster PG, Nye TM, Cox CJ, Embley TM. (2012). A congruent phylogenomic signal places eukaryotes within the Archaea. *Proc Biol Sci* **279**: 4870–4879.

Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC *et al.* (2012). Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* **337**: 1661–1665.

Yergeau E, Lawrence JR, Sanschagrin S, Waiser MJ, Korber DR, Greer CW. (2012). Next-generation sequencing of microbial communities in the Athabasca River and its tributaries in relation to oil sands mining activities. *Appl Environ Microbiol* **78**: 7626–7637.

Youssef NH, Couger MB, Struchtemeyer CG, Liggenst offer AS, Prade RA, Najar FZ *et al.* (2013). Genome of the anaerobic fungus Orpinomyces sp. C1A reveals the unique evolutionary history of a remarkable plant biomass degrader. *Appl Environ Microbiol* **79**: 4620–4634.

Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R *et al.* (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* **26**: 1608–1615.

Zuker M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**: 3406–3415.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)