The new archaeal order Lutiacidiplasmatales reveals convergent evolution in Thermoplasmatota

Paul O. Sheridan^{1,2}, Yiyu Meng¹, Tom A. Williams², Cécile Gubry-Rangin^{1,#}

Author affiliations:

¹School of Biological Sciences, University of Aberdeen, UK.

²School of Biological Sciences, University of Bristol, UK.

[#]Corresponding author: Cécile Gubry-Rangin. c.rangin@abdn.ac.uk

Keywords

horizontal gene transfer, archaea, acid tolerance, terrestrial, adaptation

Abstract

The archaeal Terrestrial Miscellaneous Euryarchaeota Group (TMEG) has been identified in various environments, and the single genome investigated thus far suggests that these organisms are anaerobic sulfite reducers. We assembled 35 new TMEG genomes that, based on genome analysis, appear to possess aerobic and facultative anaerobic lifestyles and may oxidise rather than reduce sulfite. We propose naming this order (representing 16 genera) "Lutiacidiplasmatales" due to their occurrence in various acidic environments and placement within the phylum Thermoplasmatota. A phylum-level analysis revealed that Thermoplasmatota evolution had been punctuated by several periods of high levels of novel gene family acquisition. Several essential metabolisms, such as aerobic respiration and acid tolerance, were likely acquired independently by divergent lineages through convergent evolution rather than inherited from a common ancestor. Ultimately, this study describes the terrestrially prevalent Lutiacidiplasmatales and indicates convergent evolution as an important evolutionary driving force in archaeal lineages with complex histories.

Introduction

The role of archaea in soils is understudied, with many novel lineages performing previously unknown metabolic functions within the myriad of ecosystems they inhabit^{1,2}. The Terrestrial Miscellaneous Euryarchaeota Group (TMEG) represent one of these enigmatic lineages. TMEG are members of the phylum Thermoplasmatota³ (previously the superclass Diaforarchaea⁴), whose members occupy a wide variety of environments and ecological

niches⁵. TMEG were initially described based on metagenomic clones from South African gold mines⁶ and subsequently found to represent a distinct taxonomic order⁷. They inhabit numerous environments such as soil⁸, acid mine drainage⁷ and marine subsurface sediments⁹. Functional TMEG metabolism predictions are currently based on a single studied genome, TMEG-bg1, which was reassembled from a metagenome of deep anoxic peat layers⁸. Based on this analysis, these organisms were predicted to be anaerobes that degrade long-chain fatty acids and reduce sulfite as a terminal electron acceptor, producing hydrogen sulfide as a metabolic end product. However, it is unclear how well this single genome represents the metabolism of the order and how much metabolic diversity exists between its different lineages.

Several novel groups of Thermoplasmatota have recently been described from a myriad of environments¹⁰⁻¹³. These genomes provide the opportunity to study the complex evolutionary history of this group, which has been proposed to be a critical model of complex archaeal evolution and adaptation to contrasted environments⁵. This analysis aimed at investigating outstanding questions in the evolution of Thermoplasmatota, including whether aerobic respiration and acid tolerance in specific lineages were acquired either vertically from a common ancestor or laterally (either once or multiple times), and whether the last common ancestor of Thermoplasmatota was a thermophile whose descendants transitioned to mesophily on numerous occasions independently or vice versa.

We assembled 35 TMEG-related novel genomes from acidic soils and reclassified the group as a novel order Lutiacidiplasmatales. In contrast to hypotheses based on analysis of the first available genome for the group, these organisms are predicted to be capable of aerobic respiration and oxidise, rather than reduce, sulfite generated from thiosulfate. Additionally, we reveal that essential metabolic genes, such as those involved in aerobic respiration, have been laterally acquired independently by multiple lineages during the Thermoplasmatota evolutionary history, rather than having been acquired vertically from a common ancestor. Therefore, this work contributes to our understanding of niche adaptation in archaea, particularly in lineages with complex evolutionary histories involving multiple environmental transitions.

Results

Assembly and classification of Terrestrial Miscellaneous Euryarchaeota Group genomes

Nine surface (0-15 cm) and six subsurface (30-60 cm) acidic soils were sampled from 11 locations around Scotland (Supplementary Data 1). Environmental DNA was extracted from

these samples and subjected to high-throughput sequencing. Reconstruction of metagenome-resolved genomes recovered 35 genomes representing novel lineages of terrestrial archaea, based on GTDB relative evolutionary divergence scores. These genomes were related to the Terrestrial Miscellaneous Euryarchaeota Group (TMEG), an order-level group of the Thermoplasmatota⁷. Altogether, we obtained 22 and 13 genome sequences from surface and subsurface soils, respectively (Table 1). A single genome (TMEG-bg1) has been previously investigated from subsurface anoxic peat layers⁸, and another genome (UBA184¹⁴) has been released in GTDB. In addition, there is presently no TMEG cultured representative. The newly assembled TMEG genomes were of relatively high quality, with average completeness of 78% (range: 45-98 %) and average redundancy of 3% (range: 0-9%) (Table 1). These TMEG genomes appeared to be of relatively low abundance within their environments, with an average relative abundance of 0.4% (Range: 0.1-1.4%), estimated by metagenomics sequence read recruitment (Supplementary Data 2). In addition, we also assembled a genome from the related order Lunaplasmatales¹⁰.

These 35 new TMEG genomes and the two publicly available ones (TMEG-bg1 and UBA184) were affiliated with the same order and family, based on their relative evolutionary divergence (Supplementary Data 2). They were divided into 16 genera and 24 species based on their average amino acid identity (Supplementary Data 3). The 16S rRNA gene similarities agree with the genomes representing a single family with multiple genera (Supplementary Data 4). However, the recovery rate of 16S rRNA genes for these genomes was only 41% (15 of the 37 genomes).

We selected the genome AcS3-62 as the type material for classifying this novel order as this genome meets the quality criteria for type material suggested for MIMAGs^{15, 16}. Indeed, the AcS3-62 genome has high genome completeness (97% complete), has a low level of genome fragmentation (29 contigs), possesses the 5S, 16S and 23S rRNA genes, encodes for all 20 regular amino acids tRNAs and is a member of the most represented genus in our dataset (Supplementary Data 5).

The full-length 16S rRNA gene sequence query from AcS3-62 against the NCBI nr nucleotide database (www.ncbi.nlm.nih.gov) revealed sequences with family-level similarity (> 90 %) in a wide variety of environments, including forest, fen and peat soils, mine drainage, hot springs and rivers over a wide thermal range (10-78°C) (Supplementary Data 6). Despite the large variety of habitats, they all appear to be strongly acidic (pH < 4.5), indicating that this order-level group occupies a wide range of acidic environments. The AcS3-62 16S rRNA gene

was additionally queried against the extensive collection of 16S rRNA libraries in IMNGS¹⁷, an integrated platform for analysing raw sequence read archives. Similar sequences (>90 % similarity) were present in diverse environments but were particularly prevalent in acidic environments such as mine drainage and peat soils (32 and 31% of libraries, respectively) (Figure S1, Supplementary Data 7).

We propose the name *Ca. Lutiacidiplasma sylvanus* ord. nov. sp. nov. ('Luti' and 'acidi' referring to its prevalence in acidic soil environments, 'plasma' referring to its classification within the Thermoplasmatota, 'sylvanus' as it was sequenced from a (pine) forest soil) for AcS3-62 and make it the type species of the novel family *Ca. Lutiacidiplasmataceae* and the novel order Lutiacidiplasmatales.

Metabolic traits of the Lutiacidiplasmatales

In contrast to the previously investigated TMEG-bg1⁸, the aerobic respiration terminal oxidase (Complex IV) was detected in almost all Lutiacidiplasmatales presenting high genome completeness (19 of 22 genomes), suggesting a common aerobic metabolism in Lutiacidiplasmatales (Figure 1; Supplementary Data 8). The complex IV genes are adjacent to genes encoding complex III of the electron transfer chain, biosynthesis of heme A and O, twin-arginine translocation system, cupredoxin and six other proteins of unknown function (with 5 of them possessing transmembrane domains) (Supplementary Data 9). In addition, the microaerobic respiration terminal oxidase, cytochrome bd ubiquinol oxidase genes *cydA* and *cydB*, were only present in two closely related genomes (TMEG-bg1 and UBA184), suggesting adaptation of these organisms to environments where molecular oxygen is scarce. The cytochrome bd ubiquinol oxidases identified in this study are members of the rarer quinol:O₂ oxidoreductase, qOR3 family¹⁸. Several Lutiacidiplasmatales genomes also contain the aldehyde dehydrogenase (ALDH) and aldehyde-alcohol dehydrogenase (*adhE*), which are responsible for producing acetate and ethanol, respectively. The presence of these genes suggests that these organisms may also be facultative anaerobes.

Lutiacidiplasmatales may be obligate heterotrophs. Indeed, a common feature among Lutiacidiplasmatales genomes is the absence of complete carbon fixation pathways, including the dicarboxylate-hydroxybutyrate, the hydroxypropionate-hydroxybutyrate, the reductive acetyl-CoA, the reductive hexulose-phosphate or the Wood-Ljungdahl pathways (Supplementary Data 8). Lutiacidiplasmatales genomes also lack methanogenesis genes. Additionally, many Lutiacidiplasmatales genomes possess a complete glycolytic pathway, notably having the limiting 6-phosphofructokinase gene. Most genomes contain multiple

copies of the *gloA* gene, which detoxifies the glycolysis by-product methylglyoxal, suggesting glycolysis as an essential pathway in those genomes. A complete propionate to succinate pathway is also present and includes the marker genes propionyl-CoA carboxylase beta chain *pccB*, methylmalonate-semialdehyde dehydrogenase *mmsA* and methylmalonyl-CoA/ethylmalonyl-CoA epimerase MCEE, suggesting that propionate could be another carbon and energy source under aerobic environmental conditions.

Degradation of LCFA via β -oxidation was previously proposed for TMEG-bgl genome⁸. This function appears to be a common feature in Lutiacidiplasmatales, with most genomes possessing the necessary genes for utilising long/medium-chain fatty acids (Figure 1). Lutiacidiplasmatales genomes also contain multiple alpha-glucan degrading glycoside hydrolase (GH) genes involved in the degradation of carbohydrate storage, such as amylose (GH57), trehalose (GH15) and glycogen (GH133) (Supplementary Data 10). Extracellular peptidases were also detected, with the S53 family, thermopsin and A4 family peptidases being the most prevalent (Supplementary Data 11). Notably, these peptidases are active at low pH¹⁹⁻²¹, and extracellular S53 and thermopsin peptidases are present in the acidophilic Thermoplasmatales species *Cuniculiplasma divulgatum*²². Lutiacidiplasmatales also possess the peptide/nickel ABC transport system and genes for the degradation of amino acids, such as histidine, glutamate, glutamine, glycine, valine, leucine and isoleucine (Supplementary Data 8).

Lutiacidiplasmatales have an acidophilic lifestyle, as demonstrated by multiple genetic pathways. They encode an *arcA* arginine deiminase, which is involved in the acid tolerance in several bacteria²³⁻²⁵. They also possess the acid-tolerant form of the V/A-ATPase (see section Results: Origination and evolution of critical metabolisms in Thermoplasmatales). This clade of V-type-like ATPase is adapted to function under highly acidic conditions and has been laterally acquired by many distantly related acidophilic archaea²⁶. Other mechanisms also support the acidophilic specialisation of Lutiacidiplasmatales. In particular, most Lutiacidiplasmatales possess the *Kdp* potassium transporter (EC:3.6.3.12), which is involved in pH homeostasis in acidophiles by generating reverse membrane potential^{27, 28}. Many acidophiles contain genes involved in stress resistance, and Lutiacidiplasmatales genomes contain the Uvr excinuclease, involved in DNA repair from ultraviolet DNA damage²⁹, the *MutLS* endonuclease, mitigating the mismatch repair during DNA replication, and the nickel-containing superoxide dismutase, conferring an efficient oxidative stress resistance in environments with high-nickel concentrations³⁰.

The sulfite reduction genes, *dsrAB*, of TMEG-bg1 were not detected in other Lutiacidiplasmatales genomes. In contrast, most Lutiacidiplasmatales genomes possess genes homologous to putative archaeal sulfite oxidases previously identified in Thaumarchaeota³¹. These genes may be involved in converting sulfite into sulfate, allowing the generation of ATP in oxidative phosphorylation. Instead of the transmembrane domains found in Thaumarchaeota sulfite oxidases, the N-terminal region of the Lutiacidiplasmatales sulfite oxidases possess an intrinsic disorder (i.e. natively unfolded) region. Phylogenetic analysis of these genes in various prokaryotes and eukaryotes indicates that these novel archaeal enzymes form the closest prokaryotic sister clade to eukaryotic sulfite oxidases and nitrate reductases (Figures S2,3). This suggests that the eukaryotic enzymes may have evolved from an ancestral archaeal sequence found today in Lutiacidiplasmatales, or that the Asgard archaeal ancestor of eukaryotes might have acquired this ancestral gene from Lutiacidiplasmatales by LGT.

Lutiacidiplasmatales are predicted to be non-motile due to the lack of archaellum genes, including those present in closely related Thermoplasmatota, such as Thermoplasmatales or Poseidoniales^{12, 32}. Members of the Lutiacidiplasmatales possess a high level of genome novelty (average 65%; range 54 – 76%) (Figure 1, Table 1). For example, the near-complete genome sequences of AcS3-62 encode 2,044 predicted proteins, but only 37% of these genes possessed a close homolog in the arCOG database. Given the high level of genome novelty, it is likely that our functional predictions underestimate the physiological repertoire of these organisms.

Phylogenomic tree of Thermoplasmatota

Several contrasting evolutionary histories have been previously proposed for the Thermoplasmatota^{5, 10, 13, 33}, particularly with the divergent placement of the clades Poseidoniales and Thermoprofundales. Therefore, our study provides an updated and robust phylogeny of the Thermoplasmatota (Figure 2). Using a thorough phylogenomic investigation (see SI Phylogenomics; Figure S4; Supplementary Data 12), we reconstructed a phylogeny including our new 35 Lutiacidiplasmatales genomes and representative genomes from this phylum (Supplementary Data 2). The selected topology largely agrees with the 16S rRNA gene phylogenetic tree (Figure S5). The Lutiacidiplasmatales order forms a monophyletic clade with the methanogenic Methanomassiliicoccales³⁴ and the recently described order-level groups Gimiplasmatales¹³, SG8-5³⁵ and Lunaplasmatales¹⁰ (Figure 2).

Diversification and mechanisms of genome evolution in Thermoplasmatota

Ancestral temperature preference was predicted by combining the *in silico* prediction of optimal growth temperature of extant Thermoplasmatota genomes with the reconstructed phylogenomic tree. It indicated that the last common ancestor of Thermoplasmatota was a thermophile, which transitioned to moderate temperature environments on several occasions independently during the subsequent diversification of Thermoplasmatota (Fig S6).

The Thermoplasmatota genome size is relatively consistent across the phylum, with an average of 1,954 protein-coding genes per genome (after adjustment for the completeness) [range 1,027-4,052] (Table 1). In contrast to Thaumarchaeota³¹ and Cyanobacteria³⁶, genome expansion did not occur during the transition of Thermoplasmatota to terrestrial environments (Figure S7). An increase in GC-content also happened during the early evolutionary history of the Lutiacidiplasmatales, Methanomassiliicoccales, SG8-5 order, Gimiplasmatales and Lunaplasmatatales (Figure S7).

A selection of 96 high-quality Thermoplasmatota genomes plus four outgroup genomes from the Archaeoglobales were used to infer the evolutionary history of 6,050 gene trees throughout Thermoplasmatota evolution. The probabilistic ancestral genome reconstructions for every branch of the Thermoplasmatota phylogeny enabled the characterisation and quantification of gene content changes across their history (Supplementary Data 13). The majority of the 21,183 gene content gains in Thermoplasmatota occurred through 13,025 intra-phylum gene transfers (intra-LGT) (61% of gains), 4,947 originations (including inter-phyla gene transfers and de novo gene formation) (23% of gains) and 3,211 duplications of existing genes (15% of gains) (Figure 3A). There were also 16,374 gene losses predicted (Figure 3A). Therefore, gene transfer within the Thermoplasmatota phylum and gene loss were the two main mechanisms of gene content change in Thermoplasmatota evolution.

The distribution of the gene gains and losses across the Thermoplasmatota evolutionary history was determined by estimating a punctuation score index for each of the four gene content change mechanisms. A higher punctuation score represented a lower dispersion of events across the phylogeny, with most events happening on a few branches of the phylogenomic tree. Gene family originations was the most punctuated mechanism of gene content change in Thermoplasmatota, with a punctuation score (PS) of 4.04, followed by intra-LGT (PS = 2.90), losses (PS = 2.69) and duplications (PS = 2.22) (Figure 3B). Indeed, 16 % of all gene family originations in Thermoplasmatota were predicted to have occurred in the last common ancestors (LCA) of Lutiacidiplasmatales (383 originations) and Poseidoniales (392 originations) (Figure 3; Supplementary Data 13). These originating gene families possessed a

high degree of novelty, with only 22 and 8 % having close homologs in the arCOG database for Lutiacidiplasmatales and Poseidoniales, respectively (Supplementary Data 14). These originations were not followed by high duplication and loss rates in the subsequent lineages (see SI Gene duplication), as observed for some Thaumarchaeota lineages³¹. In contrast, gene duplication appears as a more important mechanism for the evolution of the Methanomassiliicoccales or the Thermoplasmatales (15 and 21% of the gene gains throughout these lineages, respectively). As gene families with less than four sequences were not included in gene tree - species tree reconciliation, the relatively low number of extant genomes in the order-level clades Acidiprofundales, Thermoprofundales, Lunaplasmatatales, Gimiplasmatales and SG8-5 may lead to underestimations of the number of originations for their LCAs.

Origination and evolution of key metabolisms in Thermoplasmatota

Following the probabilistic ancestral gene content reconstruction, we examined progressive gain and loss of genes throughout the evolutionary history of Thermoplasmatota by comparing the set of functional genes predicted in each parent ancestor reconstruction to their descendants (expanded description in Figure S8). We complemented these broad-scale systematic analyses with targeted phylogenetics of the genes underpinning key metabolic pathways, including homologues from other archaeal and bacterial genomes (Supplementary Data 15). We used approximate unbiased testing and other statistical methods to test the likelihood of single acquisitions into the phylum (Supplementary Data 16). This approach enables the prediction of gene family acquisitions even when the same gene family has been laterally acquired multiple times into the phylum. We focused on specific metabolic traits such as oxidative phosphorylation, acid tolerance, autotrophy and heterotrophy.

Oxidative phosphorylation

Genes indicating aerobic respiration via complex IV were present in Lutiacidiplasmatales, Thermoplasmatales, Poseidoniales and Lunaplasmatatales (Supplementary Data 8, Figure S9). The gene arrangement of the complex IV locus is highly variable between orders (Figure 5). Still, it is conserved within each order, except for the Thermoplasmatales genera *Ferroplasma*, *Acidiplasma* and *Picrophilus*, which possess a divergent *coxAC*-encoding complex IV locus. Genes encoding cytochrome c cytochromes (PF00034 or PF01322) were detected in the Archaeoglobales genomes *Archaeoglobus fulgidus* DSM4304, *Archaeoglobus veneficus* SNP6 and *Ferroglobus placidus* DSM10642, but were absent from all Thermoplasmatota genomes. The complex IV subunits in Lutiacidiplasmatales, Thermoplasmatales and Lunaplasmatatales are adjacent to blue copper proteins (PF06525,

PF00127). These BCP redox proteins shuttle electrons from an electron donor to an electron acceptor³⁷ and have been proposed as an alternative electron carrier in the respiratory chain of the Thermoplasmatales species *Cuniculiplasma divulgatum*²² and some other acidophiles^{7, 38}.

The high gene arrangement divergence between orders suggests that the individual genes of complex IV may have markedly different evolutionary histories. Therefore, the phylogenies of the subunits *coxA* and *coxB* and the biogenesis genes *ctaA* and *ctaB* were compared with homologs from an expanded inter-domain set of prokaryotic genomes (Supplementary Data 15; Figures 5 and S10-13). The ancestral gene content reconstruction indicates that this trait (characterised by the marker gene *coxA*) was gained by Thermoplasmatota lineages on at least three independent occasions (Figure 4). Targeted gene tree analysis revealed that the *coxA* genes of Poseidoniales were highly divergent from those of Thermoplasmatales, Lutiacidiplasmatales and Lunaplasmatatales, clustering more closely to a clade of bacterial genes (Figures 5B and S10). This indicates that there were at least two independent acquisitions of *coxA* into Thermoplasmatota, and a single acquisition of *coxA* into Thermoplasmatota was rejected by an approximately unbiased (AU) test ($P < 0.01$). Despite the inter-domain *coxA* divergence, all Thermoplasmatota *coxA* genes were members of the D- and K-channel possessing A1 subfamily of heme-copper oxygen reductases³⁹. Multiple acquisitions during evolutionary history were also predicted for *coxB*, *ctaA* and *ctaB* genes (see SI Complex IV evolution)).

The microaerophilic oxygen reductase cytochrome bd ubiquinol oxidase is present in most members of the Thermoplasmatales and in the Lutiacidiplasmatales genomes TMEG-bgl and UBA184 (Figure S9). Ancestral gene content reconstruction indicates that the cytochrome bd ubiquinol oxidase subunit *cydA* originated in the LCA of Thermoplasmatales and has undergone two duplications in its descendants (Figure S14, Supplementary Data 17). The *cydA* present in the two Lutiacidiplasmatales members was likely transferred to their LCA from a member of the Thermoplasmatales order (0.9 TPP (posterior probability of transfers)) (Supplementary Data 17).

Acid tolerance

The energy-yielding V/A-type H⁺/Na⁺-transporting ATPases are present in all orders of the Thermoplasmatota (Supplementary Data 8). The V/A-type ATPase of most of the Thermoplasmatota orders possesses the H/GIKECFABD gene arrangement of the previously described halophilic V-type ATPase²⁶ (Figure 6A), with Acidiprofundales differing only by

subunit D being encoded in a different part of the genome than the rest of the ATPase complex. The Thermoplasmatales possess the KECFABDH/GI arrangement of the acidophilic V-type ATPase²⁶. In contrast, the Lutiacidiplasmatales possess the gene arrangement FABDH/GICEK, which has not been previously identified in archaea to the best of our knowledge.

A phylogenetic analysis of a concatenation of the three largest subunits of this complex (subunits A, B and I) revealed that a single acquisition into the phylum could be statistically rejected ($P < 0.01$). The tree showed that the ATPases of all Thermoplasmatales and Lutiacidiplasmatales genomes clustered with the acidophilic V-type ATPases²⁶ (Figures 6B and S15). In addition, the V/A-ATPases of Acidiprofundales, Thermoprofundales and Thermoplasmatota archaeon JdFR-43 were likely acquired independently of the other ATPases in Thermoplasmatota. Their clustering with other archaea present in deep-sea hydrothermal sediments probably reflects adaptation to conditions specific to this environment (such as pressure), as observed for the acid-tolerant V-type-like ATPases. Individual gene trees of the three subunits indicated the same clustering as the concatenated approach (Figure S16-18).

The acid tolerance associated arginine deiminase, *arcA*, gene, is present in most genomes of the Thermoplasmatales and Lutiacidiplasmatales (Figure S9). It was predicted by ancestral gene content reconstruction to have been gained on two independent occasions into the LCAs of these orders. However, there is some evidence that the *arcA* present in the Thermoplasmatales LCA may have been acquired from another Thermoplasmatota, such as the clustering of all Thermoplasmatota genes together to the exclusion of bacterial homologs (Figure S19) and the indication of an *arcA* intra-phylum transfer into the Thermoplasmatales LCA by gene tree-species tree reconciliation (TPP 0.58) (Supplementary Data 18). However, an intra-LGT from Lutiacidiplasmatales (the most likely donor) was poorly supported (TPP 0.27), so two independent acquisitions remain the most likely scenario.

Autotrophy and heterotrophy

Genes indicative of autotrophy were detected in some (but not all) lineages of Thermoplasmatota, and it is unclear whether Thermoplasmatota transitioned from a heterotrophic ancestor to autotrophy in some lineages or vice versa. The key genes acetyl-CoA decarbonylase, subunit gamma and delta (*cdhED*) in the CO₂ assimilating Wood-Ljungdahl pathway were present in genomes of the Gimiplasmatales and Methanomassiliicoccales but absent from other members of the phylum. The ancestral gene content reconstruction suggested that these genes were acquired once in the LCA of these two orders (Figure 4). This scenario

is supported by an expanded phylogenetic analysis of these genes (Figure S20, S21). The key methanogenesis genes, *mcrAB*, were only detected in the Methanomassiliicoccales and are predicted to have been acquired by the LCA of that order (Figure 4).

The CO₂ assimilating RuBisCO system key gene, *rbcL*, was present in the genomes of the Acidiprofundales, SG8-5 order, and Methanomassiliicoccales and ancestral gene content reconstruction suggests that this gene was gained independently in the LCAs of these orders (Figure 4). Phylogenetic analysis of this gene also supports at least three independent acquisitions, with Methanomassiliicoccales genes forming a divergent cluster and SG8-5 order genes clustering more closely with archaea from other phyla than with the Acidiprofundales genes (Figure S22). Additionally, a single acquisition of *rbcL* into the phylum was statistically rejected ($P < 0.01$).

The autotrophic carbon assimilating pathways, 3-hydroxypropionate bicycle, 3-hydroxypropionate/4-hydroxybutyrate cycle, reverse TCA and methylamine-formate reaction, were not detected in any members of the Thermoplasmatota or their ancestors (Supplementary Data 8).

The key glycolytic gene ATP-dependent phosphofructokinase *pfk* was present in most Thermoprofundales and Lutiacidiplasmatales genomes but absent from most other genomes of the phylum (Figure S9, Supplementary Data 8). The ancestral gene content reconstruction predicted that this gene was gained twice independently, once into the Lutiacidiplasmatales LCA and once into an ancestor pre-dating the Thermoprofundales LCA (Figure 4). Additionally, *pfk* genes from Lutiacidiplasmatales and Thermoprofundales were more closely related to bacterial homologs than each other, indicating two independent acquisitions into the phylum. A single acquisition of *pfk* into the phylum was statistically rejected by an AU test ($P = 0.02$) (Figure S23).

Discussion

An expanded genome sampling of the Lutiacidiplasmatales reveals that many group members are capable of aerobic respiration. This capacity also includes the previously analysed genome TMEG-bg1, which lacked complex IV but possessed a cytochrome bd ubiquinol oxidase (in our analysis). Several Lutiacidiplasmatales also encode acetate and ethanol production genes, which indicate a facultative anaerobic strategy. This strategy is likely advantageous in the frequently waterlogged soils they inhabit, where molecular oxygen can be quickly depleted. Additionally, the presence of a putative sulfite oxidase capable of producing

sulfate from sulfite and thiosulfate may explain their ecological distribution, such as in a water sample in which Lutiacidiplasmatales predominate possessing 28-180 times higher sulfate (40 mM) than in instances where Lutiacidiplasmatales was absent⁴⁰. The prevalence of Lutiacidiplasmatales in various acidic environments seems to be facilitated by several metabolic adaptations to low pH environments, such as an acid-tolerant ATPase and acidophilic peptidases (as noted in other acidophilic Thermoplasmatales species^{22, 26}), and a putatively acid tolerance-conferring arginine deiminase.

The evolutionary history of archaea is marked with several expansions into new environmental habitats and subsequent metabolic adaptation. That adaptation is likely driven by *de novo* gene origination, lateral acquisition of niche-relevant genes⁴¹⁻⁴³, and expansion of ancestral gene repertoires by duplication^{31, 44}, but the relative contributions of these processes across the archaeal tree, and the interplay between them, remain unclear. For example, it has been postulated that the evolutionary transition of the ammonia-oxidising Thaumarchaeota from hot springs to terrestrial environments co-occurred with the lateral acquisition of many new gene families and the subsequent duplication of these genes in specific terrestrial lineages³¹. Adam *et al.* noted that Thermoplasmatota represents a key model system for studying the processes underlying archaeal evolution and adaptation to contrasting environments due to the variety of habitats and metabolisms observed within this group⁵. Our study contributes to deciphering that complex history. Our analyses highlight the role of lateral gene transfer in driving metabolic transitions within Thermoplasmatota, including the convergent, independent acquisition of key genes from different (possibly bacterial) donors in different clades. For example, genes underpinning metabolisms such as oxidative phosphorylation, acid tolerance and heterotrophy were acquired multiple times from multiple donors. This suggests that Thermoplasmatota lineages have assimilated aspects of the indigenous microbial community's genetic repertoire to facilitate their niche adaption following expansion into a new habitat. Additionally, some major metabolisms that appear conserved in divergent lineages of the phylum have likely arisen from convergent evolution rather than through vertical inheritance. Due to the large variety of environments inhabited by the Thermoplasmatota, this group represents an exciting model for studying the gene content change mechanisms of habitat adaptation in archaeal lineages with complex evolutionary histories. It provides an excellent contrast to the more straightforward history of environmental transition adaption predicted for the ammonia-oxidising Thaumarchaeota^{31, 45, 46}.

Methods.

Sampling, sequencing and genome assembly

Soils samples were collected from 11 sites around Scotland (UK) (Supplementary Data 1), and the environmental DNA was extracted using Griffith's protocol⁴⁷. DNA libraries were prepared using Illumina TruSeq DNA PCR-Free Library Prep Kit, and sequencing was performed on the Illumina NovaSeq S2 platform (8.7 X 10¹⁰ reads per sample on average, Supplementary Data 19).

Reads were filtered using the READ_QC module⁴⁸, and high quality reads for each metagenome were assembled using MEGAHIT⁴⁹. Binning of resulting contigs was performed with MaxBin2⁵⁰ and metaBAT2⁵¹, and the results were consolidated using the Bin_refinement module⁴⁸. Completeness and contamination of bins were estimated with CheckM⁵², and only bins with a completeness of > 45 % and contamination of < 10 % were retained for further analysis. Genome coverage was calculated using CoverM v0.6.1 (<https://github.com/wwood/CoverM>). Bins were initially characterised using the classify_wf function in GTDB-Tk v1.3.0⁵³ using the R95 GTDB release. Genomes classified as p__Thermoplasmatota were extracted for further analysis. Defaults settings were used for all software if not otherwise stated.

Collection of public genomes

Genome sequences classified p__Thermoplasmatota and designated as the species representative genome by GTDB was downloaded from Genbank (June 2020) (Supplementary Data 2). TMEG-bg1 was also downloaded from IMG (<https://img.jgi.doe.gov/>). Representative genomes for the Poseidoniales were chosen based on the phylogenetic diversity and genome quality described previously¹² and downloaded from Genbank. Only genomes with a completeness of > 45 % and contamination of < 10 % were retained.

Genome properties

All genomes were annotated using Prokka v1.14⁵⁴, and several genome characteristics were estimated, including GC content (using QUAST⁵⁵), total predicted genomic size (measured genome size (using QUAST⁵⁵) corrected by the completeness score) and predicted optimal growth temperature (based on a machine learning model Tome⁵⁶). Optimal growth temperature in ancestors of extant Thermoplasmatota was inferred with RidgeRace⁵⁷ and using the Tome predictions as leaf values. Environmental source information and genome sequence type (i.e. culture, SAG, MAG, etc) were retrieved from NCBI or the associated

published study. The 5S, 16S and 23S rRNA and tRNA genes were identified using Barrnap v0.9 (--kingdom arc, archaeal rRNA) (<https://github.com/tseemann/barrnap>) and tRNAscan-SE v2.0.5⁵⁸ (-A, archaeal tRNA), respectively. 16S rRNA genes were compared using BLASTn v2.9.0⁵⁹.

Ortholog selection and phylogenomic reconstruction

Datasets

This study used three datasets to build a full Thermoplasmatota tree, a Thermoplasmatota tree containing only high-quality genomes, and a tree restricted to the Lutiacidiplasmatales (see below) (SI Phylogenomics).

Phylogenomic tree construction

Phylogenomic tree reconstruction involved concatenating core ortholog groups (OGs) alignments into a supermatrix. A maximum-likelihood tree of this supermatrix was constructed with IQ-TREE 2.0.3⁶⁰ using either the best-fitting protein model predicted in ModelFinder⁶¹ and an edge-linked partition model or the complex mixture model LG+C60+G+F using a non-partitioned supermatrix. Branch supports were computed using the SH-aLRT test⁶² and 2,000 UFBoot replicates, and a hill-climbing nearest-neighbor interchange (NNI) search was performed to reduce the risk of overestimating branch supports.

Taxonomic and relative abundance affiliation

Order and family level classifications were based on relative evolutionary divergence (RED)⁶³, and genus and species stratifications were defined by amino acid identity (AAI) amalgamative clustering with 70 and 95% thresholds for genus and species, respectively. AAI clusters that were polyphyletic by the species tree topology were considered to be two separate genera. AAI between sets of genomes were calculated using CompareM (<https://github.com/dparks1134/CompareM>). Relative abundance of each genome within their metagenome was estimated by mapping each genome sequence to their original assembled metagenome using the Quant_bins module in metaWRAP⁴⁸.

Predicting gene content changes across evolutionary history

For the Thermoplasmatota dataset containing only high-quality genomes, gene families were detected with Roary⁶⁴ with low stringency (-i 35, -iv 1.3, -s), and sequences shorter than 30 amino acids and families with less than four sequences were removed from further analysis. All remaining sequences within each family were aligned using MAFFT L-INS-i⁶⁵, processed

with trimAl (automated1)⁶⁶ and ML phylogenetic trees were constructed for each alignment as described above. The majority of the gene family trees (6050 of the 6059) could be probabilistically reconciled against the supermatrix tree using the ALEml_undated algorithm of the ALE package⁶⁷ to infer the numbers of duplications, intra-LGTs, losses and originations on each branch of the supermatrix species tree. Genome incompleteness was probabilistically accounted for within ALE using the genome completeness values estimated by CheckM⁵². The mechanism of gene content change on every branch of the species tree was estimated using branchwise_numbers_of_events.py, as described before³¹. The number of intra-LGTs to and from every branch of the species tree was estimated with calc_from_to_T.sh (see https://github.com/SheridanPO/ALE_analysis). All phylogenomic trees were visualised using iTOL⁶⁸.

A punctuation score was given to each mechanism of gene content change to measure the extent to which the branches with the highest numbers of events influence the total number of events for a given genome dataset. For a given mechanism, the punctuation score is calculated by dividing the sum of event numbers in the 10% of branches with the highest event numbers by 10% of the sum of events on all branches.

$$\text{Punctuation score} = \frac{\sum_{\text{events in top 10\% branches}}}{(\sum_{\text{events in all branches}} * 0.1)}$$

Gene family originations were predicted to result from inter-phyla lateral gene transfer if the medoid of the gene family possesses homologues in a database of UniRef90⁶⁹ sequences with strain-level designations and excluding Thermoplasmatota matches. The origins of these putatively laterally acquired gene families were estimated using the best match (as determined by bit score) in this database.

Probabilistic ancestral genome reconstructions were created for each branch of the species tree using gene_copies_at_node.py, as described previously³¹, giving a list of gene families predicted to be present in each ancestor. When the KEGG ortholog numbers corresponding to these gene families are present in a parent but not in the descendant branch genome reconstruction, they are considered functional gene loss. In the reverse situation, they are regarded as functional gene gains. This method enabled the prediction of multiple independent gene family gains and was combined with individual gene tree analysis to infer whether the transfer was from outside of the phylum.

Functional annotation of gene families

For each protein family, a medoid sequence (the sequence with the shortest summed genetic distances to all sequences in the family) was calculated under the BLOSUM62 substitution matrix using DistanceCalculator in Phylo (<https://biopython.org/wiki/Phylo>). Medoids were annotated against the KEGG database⁷⁰ using GhostKOALA⁷¹, against the arCOG database⁷² using Diamond BLASTp⁷³ (best-hit and removing matches with e-value >10⁻⁵, % ID < 35, alignment length < 80 or bit score < 100) and against the Pfam⁷⁴ and TIGRFAM⁷⁵ databases using hmmsearch⁷⁶ (-T 80). Annotation of specific gene families of interest is described in SI: Functional annotation of gene families.

Evolutionary history of selected gene families

A selection of 204 phylogenetically diverse genome sequences (176 bacterial and 28 archaeal) was downloaded from GenBank (Supplementary Data 16n). These genomes were annotated against the KEGG database⁷⁷ using GhostKOALA⁷¹. Protein sequences of genes annotated as particular K numbers of interest were extracted from the genomes and combined with corresponding protein sequences in the Thermoplasmatota genomes. Sequences less than 50 amino acids long were removed, and the remaining sequences were aligned using MAFFT L-INS-I⁷⁸, processed with trimAL⁶⁶ (automated1) and sequences whose alignments were composed of > 50 % gaps were removed. An ML phylogenetic tree was constructed using IQ-TREE 2.0.3⁶⁰ with 1000 UFBoot replicates, an NNI search and the best substitution model selected by ModelFinder⁶¹. For the phylogeny of the V/A-ATPase, genes from *Nitrosocosmicus oleophilis* MY3, *Nitrosotalea okcheonensis* CS and *Nitrososphaera viennensis* EN76 were added to the analysis. Single acquisitions of genes into the phylum were assessed by comparing the unconstrained gene tree with a tree constrained to form a monophyletic Thermoplasmatota clade. These topologies were then compared with approximately unbiased tests and other statistical tests implemented in IQ-TREE 2.0.3⁶⁰ (-zb 10000 -zw -au).

Note

Thermoplasmatales, Acidiprofundales, Thermoprofundales, Poseidoniales, Lunaplasmatales, Gimiplasmatales and Lutiacidiplasmatales are not validly published names under the International Code of Nomenclature of Prokaryotes and thus can be considered as candidatus taxa. The *Candidatus* prefix was omitted from these taxa in the manuscript for brevity.

Acknowledgements

UKRI financially supported P.O.S. and Y.M. through the NERC grant (NE/R001529/1). In addition, C.G.-R. and T.A.W. were supported by a Royal Society University Research Fellowship (URF150571 and UF140626). We thank Tony Travis for his support with Biolinux. The authors would also like to acknowledge the support of the Maxwell computer cluster funded by the University of Aberdeen.

Author contributions

P.O.S., T.A.W. and C.G.-R. designed the study and developed the theory. P.O.S. assembled the 36 new genomes. P.O.S. collected the samples, and Y.M. performed DNA extraction. P.O.S., T.A.W. and C.G.-R. interpreted the results and wrote the manuscript. All authors accepted the final version of the manuscript.

Code availability

Scripts for general manipulation of ALE outputs have been deposited at <https://github.com/Tancata/phylo/tree/master/ALE>, and additional scripts specific to this work have been deposited at https://github.com/SheridanPO/ALE_analysis.

Data availability

Accession numbers for the 36 new genomes presented in this study can be found in Supplementary Data 2 and under the NCBI BioProject xxx.

References

1. Baker, B. J. *et al.* Diversity, ecology and evolution of Archaea. *Nature microbiology* **5**, 887-900 (2020).
2. Solden, L., Lloyd, K. & Wrighton, K. The bright side of microbial dark matter: lessons learned from the uncultivated majority. *Curr. Opin. Microbiol.* **31**, 217-226 (2016).
3. Rinke, C. *et al.* A standardised archaeal taxonomy for the Genome Taxonomy Database. *Nature Microbiology*, 1-14 (2021).
4. Petitjean, C., Deschamps, P., López-García, P., Moreira, D. & Brochier-Armanet, C. Extending the conserved phylogenetic core of archaea disentangles the evolution of the third domain of life. *Mol. Biol. Evol.* **32**, 1242-1254 (2015).

- 544 5. Adam, P. S., Borrel, G., Brochier-Armanet, C. & Gribaldo, S. The growing tree of
545 Archaea: new perspectives on their diversity, evolution and ecology. *The ISME journal* **11**,
546 2407-2425 (2017).
- 547 6. Takai, K., Moser, D. P., DeFlaun, M., Onstott, T. C. & Fredrickson, J. K. Archaeal
548 diversity in waters from deep South African gold mines. *Appl. Environ. Microbiol.* **67**, 5750-
549 5760 (2001).
- 550 7. Korzhnikov, A. A. *et al.* Archaea dominate the microbial community in an ecosystem with
551 low-to-moderate temperature and extreme acidity. *Microbiome* **7**, 1-14 (2019).
- 552 8. Lin, X., Handley, K. M., Gilbert, J. A. & Kostka, J. E. Metabolic potential of fatty acid
553 oxidation and anaerobic respiration by abundant members of Thaumarchaeota and
554 Thermoplasmata in deep anoxic peat. *The ISME journal* **9**, 2740-2744 (2015).
- 555 9. Teske, A. & Sørensen, K. B. Uncultured archaea in deep marine subsurface sediments:
556 have we caught them all? *The ISME journal* **2**, 3-18 (2008).
- 557 10. Zinke, L. A. *et al.* Evidence for non-methanogenic metabolisms in globally distributed
558 archaeal clades basal to the Methanomassiliicoccales. *Environ. Microbiol.* **23**, 340-357
559 (2021).
- 560 11. Zhou, Z. *et al.* Genomic and transcriptomic insights into the ecology and metabolism of
561 benthic archaeal cosmopolitan, Thermoprofundales (MBG-D archaea). *The ISME journal* **13**,
562 885-901 (2019).
- 563 12. Tully, B. J. Metabolic diversity within the globally abundant Marine Group II
564 Euryarchaea offers insight into ecological patterns. *Nature communications* **10**, 1-12 (2019).
- 565 13. Hu, W. *et al.* Metagenomic insights into the metabolism and evolution of a new
566 Thermoplasmata order (Candidatus Gimiplasmatales). *Environ. Microbiol.* (2020).
- 567 14. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes
568 substantially expands the tree of life. *Nature microbiology* **2**, 1533-1542 (2017).
- 569 15. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG)
570 and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.*
571 **35**, 725-731 (2017).
- 572 16. Chuvochina, M. *et al.* The importance of designating type material for uncultured taxa.
573 *Syst. Appl. Microbiol.* **42**, 15-21 (2019).
- 574 17. Lagkouvardos, I. *et al.* IMNGS: a comprehensive open resource of processed 16S rRNA
575 microbial profiles for ecology and diversity studies. *Scientific reports* **6**, 1-9 (2016).
- 576 18. Murali, R., Gennis, R. B. & Hemp, J. Evolution of the cytochrome bd oxygen reductase
577 superfamily and the function of CydAA' in Archaea. *The ISME Journal*, 1-15 (2021).
- 578 19. Oda, K., Takahashi, S., Ito, M. & Dunn, B. M. in *Aspartic Proteinases* 349-353 (Springer,
579 1998).

20. Lin, X. & Tang, J. Purification, characterisation, and gene cloning of thermopsin, a thermostable acid protease from *Sulfolobus acidocaldarius*. *J. Biol. Chem.* **265**, 1490-1495 (1990).
21. Rawlings, N. D. & Barrett, A. J. [13] Evolutionary families of metallopeptidases. *Meth. Enzymol.* **248**, 183-228 (1995).
22. Golyshina, O. V. *et al.* Biology of archaea from a novel family Cuniculiplasmataceae (Thermoplasmata) ubiquitous in hyperacidic environments. *Scientific reports* **6**, 1-12 (2016).
23. Cunin, R., Glansdorff, N., Pierard, A. & Stalon, V. Biosynthesis and metabolism of arginine in bacteria. *Microbiol. Rev.* **50**, 314-352 (1986).
24. Marquis, R., Bender, G., Murray, D. & Wong, A. Arginine deiminase system and bacterial adaptation to acid environments. *Appl. Environ. Microbiol.* **53**, 198-200 (1987).
25. Fulde, M. *et al.* ArgR is an essential local transcriptional regulator of the arcABC operon in *Streptococcus suis* and is crucial for biological fitness in an acidic environment. *Microbiology* **157**, 572-582 (2011).
26. Wang, B. *et al.* Expansion of Thaumarchaeota habitat range is correlated with horizontal transfer of ATPase operons. *The ISME journal* **13**, 3067-3079 (2019).
27. Baker-Austin, C. & Dopson, M. Life in acid: pH homeostasis in acidophiles. *Trends Microbiol.* **15**, 165-171 (2007).
28. Herbold, C. W. *et al.* Ammonia-oxidising archaea living at low pH: Insights from comparative genomics. *Environ. Microbiol.* **19**, 4939-4952 (2017).
29. Crowley, D. J. *et al.* The uvrA, uvrB and uvrC genes are required for repair of ultraviolet light induced DNA photoproducts in *Halobacterium* sp. NRC-1. *Saline systems* **2**, 1-13 (2006).
30. Kim, E., Chung, H., Suh, B., Hah, Y. C. & Roe, J. Transcriptional and post-transcriptional regulation by nickel of sodN gene encoding nickel-containing superoxide dismutase from *Streptomyces coelicolor* Müller. *Mol. Microbiol.* **27**, 187-195 (1998).
31. Sheridan, P. O. *et al.* Gene duplication drives genome expansion in a major lineage of Thaumarchaeota. *Nature communications* **11**, 1-12 (2020).
32. Faguy, D. M., Bayley, D. P., Kostyukova, A. S., Thomas, N. A. & Jarrell, K. F. Isolation and characterisation of flagella and flagellin proteins from the Thermoacidophilic archaea *Thermoplasma volcanium* and *Sulfolobus shibatae*. *J. Bacteriol.* **178**, 902-905 (1996).
33. Diamond, S. *et al.* Soils and sediments host Thermoplasmata archaea encoding novel copper membrane monooxygenases (CuMMOs). *The ISME Journal*, 1-15 (2022).
34. Iino, T. *et al.* Candidatus *Methanogranum caenicola*: a novel methanogen from the anaerobic digested sludge, and proposal of *Methanomassiliicoccaceae* fam. nov. and

- Methanomassiliicoccales ord. nov., for a methanogenic lineage of the class Thermoplasmata. *Microbes and environments*, ME12189 (2013).
35. Lazar, C. S., Baker, B. J., Seitz, K. W. & Teske, A. P. Genomic reconstruction of multiple lineages of uncultured benthic archaea suggests distinct biogeochemical roles and ecological niches. *The ISME journal* **11**, 1118-1129 (2017).
36. Chen, M. *et al.* Comparative genomics reveals insights into cyanobacterial evolution and habitat adaptation. *The ISME Journal* **15**, 211-227 (2021).
37. De Rienzo, F., Gabdoulline, R., Menziani, M. C. & Wade, R. Blue copper proteins: a comparative analysis of their molecular interaction properties. *Protein Science* **9**, 1439-1454 (2000).
38. Komorowski, L., Verheyen, W. & Schäfer, G. The archaeal respiratory supercomplex SoxM from *S. acidocaldarius* combines features of quinole and cytochrome c oxidases. (2002).
39. Sousa, F. L., Alves, R. J., Pereira-Leal, J. B., Teixeira, M. & Pereira, M. M. A bioinformatics classifier and database for heme-copper oxygen reductases. *PLoS One* **6**, e19117 (2011).
40. Dorador, C., Vila, I., Remonsellez, F., Imhoff, J. F. & Witzel, K. Unique clusters of Archaea in Salar de Huasco, an athalassohaline evaporitic basin of the Chilean Altiplano. *FEMS Microbiol. Ecol.* **73**, 291-302 (2010).
41. Méheust, R. *et al.* Hundreds of novel composite genes and chimeric genes with bacterial origins contributed to haloarchaeal evolution. *Genome Biol.* **19**, 1-12 (2018).
42. Nelson-Sathi, S. *et al.* Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* **517**, 77-80 (2015).
43. Deschamps, P., Zivanovic, Y., Moreira, D., Rodriguez-Valera, F. & López-García, P. Pangenome evidence for extensive interdomain horizontal transfer affecting lineage core and shell genes in uncultured planktonic thaumarchaeota and euryarchaeota. *Genome biology and evolution* **6**, 1549-1563 (2014).
44. Spang, A. *et al.* The genome of the ammonia-oxidising Candidatus Nitrososphaera gargensis: insights into metabolic versatility and environmental adaptations. *Environ. Microbiol.* **14**, 3122-3145 (2012).
45. Alves, R. J. E., Minh, B. Q., Urich, T., von Haeseler, A. & Schleper, C. Unifying the global phylogeny and environmental distribution of ammonia-oxidising archaea based on amoA genes. *Nature communications* **9**, 1-17 (2018).
46. Ren, M. *et al.* Phylogenomics suggests oxygen availability as a driving force in Thaumarchaeota evolution. *The ISME journal* **13**, 2150-2161 (2019).
47. Griffiths, R. I., Whiteley, A. S., O'Donnell, A. G. & Bailey, M. J. Rapid method for coextraction of DNA and RNA from natural environments for analysis of ribosomal DNA-

652 and rRNA-based microbial community composition. *Appl. Environ. Microbiol.* **66**, 5488-
653 5491 (2000).

654 48. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-
655 resolved metagenomic data analysis. *Microbiome* **6**, 1-13 (2018).

656 49. Li, D., Liu, C., Luo, R., Sadakane, K. & Lam, T. MEGAHIT: an ultra-fast single-node
657 solution for large and complex metagenomics assembly via succinct de Bruijn graph.
658 *Bioinformatics* **31**, 1674-1676 (2015).

659 50. Wu, Y., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to
660 recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605-607 (2016).

661 51. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient
662 genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).

663 52. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM:
664 assessing the quality of microbial genomes recovered from isolates, single cells, and
665 metagenomes. *Genome Res.* **25**, 1043-1055 (2015).

666 53. Chaumeil, P., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to
667 classify genomes with the Genome Taxonomy Database. *Bioinformatics* (2018).

668 54. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069
669 (2014).

670 55. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for
671 genome assemblies. *Bioinformatics* **29**, 1072-1075 (2013).

672 56. Li, G., Rabe, K. S., Nielsen, J. & Engqvist, M. K. Machine learning applied to predicting
673 microorganism growth temperatures and enzyme catalytic optima. *ACS synthetic biology* **8**,
674 1411-1420 (2019).

675 57. Kratsch, C. & McHardy, A. C. RidgeRace: ridge regression for continuous ancestral
676 character estimation on phylogenetic trees. *Bioinformatics* **30**, i527-i533 (2014).

677 58. Chan, P. P. & Lowe, T. M. in *Gene prediction* 1-14 (Springer, 2019).

678 59. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment
679 search tool. *J. Mol. Biol.* **215**, 403-410 (1990).

680 60. Nguyen, L., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and
681 effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol.*
682 *Evol.* **32**, 268-274 (2015).

683 61. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., von Haeseler, A. & Jermiin, L. S.
684 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods* **14**,
685 587 (2017).

62. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307-321 (2010).
63. Parks, D. H. *et al.* A standardised bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996-1004 (2018).
64. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691-3693 (2015).
65. Katoh, K. & Toh, H. Recent developments in the MAFFT multiple sequence alignment program. *Briefings in bioinformatics* **9**, 286-298 (2008).
66. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009).
67. Szöllősi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E. & Daubin, V. Efficient exploration of the space of reconciled gene trees. *Syst. Biol.* **62**, 901-912 (2013).
68. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127-128 (2007).
69. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282-1288 (2007).
70. Kanehisa, M. *The KEGG database* (Novartis Foundation Symposium, Wiley Online Library, 2002).
71. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG tools for functional characterisation of genome and metagenome sequences. *J. Mol. Biol.* **428**, 726-731 (2016).
72. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Archaeal clusters of orthologous genes (arCOGs): an update and application for analysis of shared features between Thermococcales, Methanococcales, and Methanobacteriales. *Life* **5**, 818-840 (2015).
73. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature methods* **12**, 59 (2015).
74. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427-D432 (2019).
75. Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371-373 (2003).
76. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755-763 (1998).
77. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**, D277-80 (2004).

78. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059-3066 (2002).

79. Tria, F. D. K., Landan, G. & Dagan, T. Phylogenetic rooting using minimal ancestor deviation. *Nature ecology & evolution* **1**, 1-7 (2017).

Figures and tables

Figure 1. Phylogenomic tree of Lutiacidiplasmatales and associated metabolism. This tree includes 37 Lutiacidiplasmatales genomes assembled from surface and subsurface soils (35 genomes obtained in this study and two published sequences). The tree was inferred by maximum likelihood from 738 concatenated phylogenetic markers, which were aligned separately and analysed using the best-fitting model for each alignment. The tree was rooted with the Methanomassiliicoccales strains *Methanoplasma termitum* MpT1, Methanomethylophilus archaeon BRNA1 and *Methanomethylophilus alvus* Mx1201. Dots indicate branches with >95% UFBoot and SH-aLRT support. Intermittent blue and white colours on the leaf labels indicate the 16 known genera of Lutiacidiplasmatales, as determined by an amino acid identity of 70 % or greater. *coxA* (heme-copper oxygen reductases, subunit A; K02274), *coxB* (heme-copper oxygen reductases, subunit B; K02275), *ctaA* (heme a synthase; K02259), *ctaB* (heme o synthase; K02257), *cydA* (cytochrome bd ubiquinol oxidase, subunit A; K00425), *cydB* (cytochrome bd ubiquinol oxidase, subunit B; K00426), *adhE* (alcohol dehydrogenase; K0407), *dsrAB* (dissimilatory sulfite reductase, subunits A and B; K11180 and K11181), *pfk/pfp* (ATP-dependent phosphofructokinase; K21071), *gloA* (lactoylglutathione lyase; K01759), *pccB* (propionyl-CoA carboxylase, beta chain, K01966), MCEE (methylmalonyl-CoA/ethylmalonyl-CoA epimerase; K05606), *mmsA* (methylmalonate-semialdehyde dehydrogenase; K00140), *fadD* (long-chain acyl-CoA synthetase; K01897), *acd* (acyl-CoA dehydrogenase; K00249), *crt* (enoyl-CoA hydratase; K01715), *fadB* (3-hydroxybutyryl-CoA dehydrogenase; K00074), *fadA* (acetyl-CoA acyltransferase; K00632), *atoB* (acetyl-CoA C-acetyltransferase; K00626), GH# (glycoside hydrolase family #; dbCAN), Pen amidase (Penicillin amidase), *arcA* (arginine deiminase; K01478), *atpABI* (acid) (V/A-type atpase A, B and I subunits; K02117, K02118 and K02123, plus gene tree analysis), *kdpABC* (K⁺ transporting ATPase subunits A, B and C; K01546, K01547 and K01548), *uvrABC* (excinuclease subunits A, B and C; K03701, K03702 and K03703), *mutLS* (DNA mismatch repair proteins L and S; K03572 and K03555), SOD (Ni) (nickel superoxide dismutase; K00518), SUOX (sulfite oxidase; PF00174, plus gene tree

analysis) and TST (thiosulfate/3-mercaptopyruvate sulfurtransferase; K01011). Similar metabolic analysis expanded to the whole Thermoplasmatota genome database is presented in Figure S8. The predicted completeness, redundancy and proteome novelty of each genome are indicated in the bar charts on the far right of the figure.

Figure 2. Phylogenomic tree of Thermoplasmatota. This tree includes 120 Thermoplasmatota genomes from a variety of environments. The tree was inferred by maximum likelihood from 108 concatenated marker genes, aligned separately and analysed using the best-fitting model for each alignment. The tree was rooted with four Archaeoglobales genomes. Dots indicate branches with >95% UFBoot and SH-aLRT support. As previously published studies have recovered different phylogenies for Thermoplasmatota, species trees inferred using a variety of approaches were compared. This phylogeny was the most likely, based on approximately unbiased testing and this topology being reconstructed in six of the seven approaches applied here (see SI Phylogenomics). Protein novelty is defined as the percentage of encoded proteins that lack a close homologue in the arCOG database. Detailed genome information is given in Supplementary Data 1.

Figure 3. Inferred gene originations, duplications, intra-phylum transfers and losses during the evolution of Thermoplasmatota. The quantitative and qualitative predictions of the genome content changes were estimated across the Thermoplasmatota history and reported on the circular cladograms (panel **a**) possessing the topology of the Thermoplasmatota tree (see Figure S6). The mechanisms of gene content changes quantified were duplication (copying of a gene within a genome), loss (loss of a gene within a genome), intra-phylum transfers (defined as the acquisition of a gene from another member of the same phylum) (Intra-LGT) and origination (defined as the acquisition of a gene from members of phyla outside the sampled genome set or by de novo gene formation). Scale numbers indicate the range of the predicted number of events for each given mechanism. The order-level classification of the genomes is indicated by the coloured bar surrounding the circle. The boxplot (panel **b**) represents normalised events per branch $((\text{events per branch} - \mu)/\sigma)$ for each mechanism. Numbered circles mark branches of the tree with the highest numbers of events, and the position of these branches is also indicated with numbered circles on the trees (panel **a**). A punctuation score is measured for each given mechanism. It represents the sum of events in the 10% of branches

with the highest event numbers divided by 10% of the sum of events into all branches ($\Sigma_{\text{events in top10\%}} / (\Sigma_{\text{events in all branches}} * 0.1)$).

Figure 4. Evolution of Thermoplasmatota metabolisms. The origins of key metabolisms were mapped on the phylogeny of Thermoplasmatota. Progressive gain and loss of genes between ancestral reconstructions and individual gene tree analysis were used to detect multiple independent lateral acquisitions of the same metabolisms into Thermoplasmatota lineages. The triangles represent collapsed clades. A dotted triangle indicates that this clade consists of less than four representative genomes, so originations may be underrepresented in the LCA of this clade (dotted branch). Marker genes of the selected metabolisms are in parenthesis. *coxA* (heme-copper oxygen reductase, subunit A; K02274), *cydA* (cytochrome bd ubiquinol oxidase, subunit A; K00425), *atpABI* (acid) (V/A-type atpase A, B and I subunits; K02117, K02118 and K02123), *arcA* (arginine deiminase; K01478), *cdhDE* (acetyl-CoA decarbonylase/synthase complex D; K00194 and E; K00197 subunits), *rbcL* (ribulose-bisphosphate carboxylase large chain; K01601) and *pfk* (ATP-dependent phosphofructokinase; K21071).

Figure 5. Complex IV genetic organisation and phylogeny. Panel a: The physical genetic organisation of complex IV genes in Thermoplasmatota is conserved at the order level but highly variable at the phylum level. The complex IV loci in the genera *Ferroplasma*, *Acidiplasma* and *Picrophilus* (Thermoplasmatales 2, panel a), are different from those present in the rest of the Thermoplasmatales order (Thermoplasmatales 1, panel a). The black backslash (/) between genes indicate sections that are physically separated in the genome sequence, and white arrows represent hypothetical proteins. Panel b: the ML tree of the *coxA* subunit of heme-copper oxygen reductase shows that the Poseidoniales *coxA* gene clusters more closely with bacterial than other archaeal sequences, indicating an independent acquisition from a bacterial donor. Circles indicate collapsed clades. The Thaumarchaeota, Lunaplasmatales and SW-10-69-26 branches are represented by the single genomes *Cenarchaeum symbiosum* A, Lunaplasmatales archaeon RBG-16-67-27 and Thermoplasmatota archaeon SW-10-69-26, respectively. Dots indicate branches with >75% UFBoot support. All analysed *coxA* genes are from A1 subfamily, apart from one bacterial clade belonging to A2 (labelled on the tree). The

expanded *coxA* phylogenetic tree and trees of the genes *coxB*, *ctaA* and *ctaB* are presented in Figures S10-13.

Figure 6. V/A-type ATPase locus gene order and phylogeny. Panel a: The gene order of V/A-type ATPase genes is conserved at the order level and broadly conserved at the phylum level, except for Thermoplasmatales and Lutiplasmatales. The tree indicates the phylogeny of the species in which these loci are found. Four gene orders presented in Wang *et al.* 2019²⁶ are shown in parenthesis for comparison. The arrow colour indicates genes that tend to be adjacent in all gene orders and is added for comparison. Panel b: The ML tree of the three largest subunits (*atpA*, B and I) concatenated (panel b). Numbers on branches represent the percentage of UFBoot support. The tree was rooted by minimal ancestral deviation (MAD)⁷⁹. The expanded *atpABI* tree is presented in Figure S15.

Table 1. Genome characteristics of newly sequenced metagenome-assembled genomes.

Figure 1

Figure 2

Figure 3

Figure 4

Figure 5

a

b

Figure 6

a Species phylogeny and V/A-ATPase locus

b V/A-ATPase phylogeny

Table 1

Genome ID	Completeness	Redundancy	Closest relative in GTDB	Adjusted genome size (bp)	# Contigs	tRNAs (total)	tRNAs (# aa encoded)	Adjusted CDS#	Proteome novelty (%)	Genome type	Environment source	Ecosystem
AcS10-2	48	0	f__UBA184	1.85E+06	263	24	15	1984	69	MAG	Surface soil	Pine forest
AcS11-139	58	3	f__UBA184	1.78E+06	591	16	12	2234	75	MAG	Surface soil	Pine forest
AcS11-62	95	2	f__UBA184	2.28E+06	129	47	20	2062	65	MAG	Surface soil	Pine forest
AcS1-3	59	0	f__UBA184	3.23E+06	219	41	20	3238	64	MAG	Surface soil	Pine forest
AcS13-54	93	2	f__UBA184	2.16E+06	17	44	20	1027	62	MAG	Surface soil	Pine forest
AcS1-36	96	4	g__UBA184	9.68E+05	28	24	15	1861	56	MAG	Surface soil	Pine forest
AcS14-35	55	2	f__UBA184	1.84E+06	213	31	14	1915	67	MAG	Surface soil	Marsh
AcS1-45	83	2	f__UBA184	2.30E+06	222	34	17	2259	66	MAG	Surface soil	Pine forest
AcS1-67	84	2	f__UBA184	1.90E+06	37	41	20	1850	62	MAG	Surface soil	Pine forest
AcS2-54	55	9	f__UBA184	1.92E+06	139	22	12	1945	64	MAG	Surface soil	Pine forest
AcS3-62	97	2	f__UBA184	2.16E+06	29	46	20	2044	63	MAG	Surface soil	Pine forest
AcS3-69	54	4	f__UBA184	1.96E+06	411	8	8	2277	74	MAG	Surface soil	Pine forest
AcS4-12	45	3	f__UBA184	1.98E+06	216	26	17	2029	65	MAG	Surface soil	Birch forest
AcS4-16	77	9	f__UBA184	2.19E+06	356	32	16	2218	69	MAG	Surface soil	Birch forest
AcS4-93	94	7	f__UBA184	2.38E+06	205	50	19	2272	67	MAG	Surface soil	Birch forest
AcS5-107	98	2	f__UBA184	2.50E+06	52	49	20	2284	66	MAG	Surface soil	Pine forest
AcS5-109	58	6	f__UBA184	2.37E+06	400	10	8	2707	75	MAG	Surface soil	Pine forest
AcS5-116	98	3	f__UBA184	2.18E+06	51	42	20	1982	63	MAG	Surface soil	Pine forest
AcS5-34	92	2	f__UBA184	2.06E+06	146	39	18	2023	65	MAG	Surface soil	Pine forest
AcS5-58	97	2	f__UBA184	2.18E+06	27	45	20	2037	65	MAG	Surface soil	Pine forest
AcS5-77	74	5	f__UBA184	2.33E+06	309	17	13	2426	69	MAG	Surface soil	Pine forest
AcS5-86	46	4	f__UBA184	2.53E+06	411	7	4	3014	76	MAG	Surface soil	Pine forest
SubAcS10-22	95	2	f__UBA184	1.60E+06	5	44	20	1492	55	MAG	Subsurface soil	Pine forest
SubAcS10-31	94	3	f__UBA184	1.94E+06	105	44	19	1863	63	MAG	Subsurface soil	Pine forest
SubAcS11-52	89	2	f__UBA184	1.59E+06	6	41	20	1438	54	MAG	Subsurface soil	Pine forest
SubAcS13-24	94	1	f__UBA184	1.72E+06	95	39	20	1655	60	MAG	Subsurface soil	Pine forest
SubAcS13-35	65	1	f__UBA184	1.66E+06	164	25	16	1723	64	MAG	Subsurface soil	Pine forest
SubAcS13-64	93	3	f__UBA184	1.93E+06	97	45	20	1802	61	MAG	Subsurface soil	Pine forest
SubAcS14-22	51	6	f__UBA184	2.96E+06	569	37	17	3302	75	MAG	Subsurface soil	Marsh
SubAcS15-120	92	2	f__UBA184	2.19E+06	106	43	20	2076	65	MAG	Subsurface soil	Grassland
SubAcS15-131	60	0	g__RBG-16-68-12	1.31E+06	301	14	11	1633	64	MAG	Subsurface soil	Grassland
SubAcS15-25	88	3	f__UBA184	1.42E+06	161	33	15	1454	59	MAG	Subsurface soil	Grassland
SubAcS15-31	96	2	f__UBA184	2.16E+06	5	47	20	2111	62	MAG	Subsurface soil	Grassland
SubAcS15-47	56	2	f__UBA184	2.64E+06	373	35	16	2798	70	MAG	Subsurface soil	Grassland
SubAcS15-67	93	3	f__UBA184	1.97E+06	159	45	20	1902	62	MAG	Subsurface soil	Grassland
SubAcS9-61	89	4	f__UBA184	1.96E+06	311	35	18	1981	66	MAG	Subsurface soil	Peaty gley