





Gene duplication drives genome expansion in a major lineage of Thaumarchaeota

Paul O. Sheridan ^{1,2}, Sebastien Raguideau³, Christopher Quince^{3,4,5}, Jennifer Holden⁶, Lihong Zhang ⁷, Thames Consortium*, Tom A. Williams ² & Cécile Gubry-Rangin ¹✉

Ammonia-oxidising archaea of the phylum Thaumarchaeota are important organisms in the nitrogen cycle, but the mechanisms driving their radiation into diverse ecosystems remain underexplored. Here, existing thaumarchaeotal genomes are complemented with 12 genomes belonging to the previously under-sampled Nitrososphaerales to investigate the impact of lateral gene transfer (LGT), gene duplication and loss across thaumarchaeotal evolution. We reveal a major role for gene duplication in driving genome expansion subsequent to early LGT. In particular, two large LGT events are identified into Nitrososphaerales and the fate of these gene families is highly lineage-specific, being lost in some descendant lineages, but undergoing extensive duplication in others, suggesting niche-specific roles. Notably, some genes involved in carbohydrate transport or coenzyme metabolism were duplicated, likely facilitating niche specialisation in soils and sediments. Overall, our results suggest that LGT followed by gene duplication drives Nitrososphaerales evolution, highlighting a previously under-appreciated mechanism of genome expansion in archaea.

¹School of Biological Sciences, University of Aberdeen, Aberdeen, UK. ²School of Biological Sciences, University of Bristol, Bristol, UK. ³Warwick Medical School, University of Warwick, Coventry, UK. ⁴Organisms and Ecosystems, Earlham Institute, Norwich, UK. ⁵Gut Microbes and Health, Quadram Institute, Norwich, UK. ⁶School of Life Sciences, University of Warwick, Coventry, UK. ⁷European Centre for Environment and Human Health, Medical School, University of Exeter, Exeter, UK. *A list of authors and their affiliations appears at the end of the paper. ✉email: c.rangin@abdn.ac.uk

Thaumarchaeota have attracted much attention since their discovery, with accumulating evidence of their ubiquitous distribution and crucial ecosystem function in many environments^{1–3}. Isolation and cultivation demonstrated that many of these organisms derive energy from the oxidation of ammonia under aerobic conditions. Ammonia oxidation is limited to a small number of microbial phyla but is an important microbial process as it implements an essential and rate-limiting step in the global nitrogen cycle, the conversion of ammonia to nitrite (via hydroxylamine). Within Thaumarchaeota, ammonia-oxidising archaea (AOA) are ubiquitous and abundant in soils and oceans. At present, all identified AOA are members of the class Nitrososphaeria⁴ and are classically placed in four order-level phylogenetic lineages¹: the Nitrososphaerales⁵, Nitrosopumilales⁶, *Candidatus* Nitrosotaleales⁷ and *Candidatus* Nitrosocaldales⁸.

While a substantial number of ecological studies and several cultures and enrichments confirm the ammonia-oxidising activity of Thaumarchaeota, there is also evidence that ammonia oxidation is not universal in this phylum^{9–13}. In particular, there is evidence that the early-diverging lineages of Thaumarchaeota (including the Group 1.1c, Group 1.3 and pSL12 lineages) do not require ammonia oxidation for growth^{10,13}, and the ammonia oxidation machinery may have been acquired subsequently to the origin of Thaumarchaeota during, or after, the Great Oxygenation Event 2300 My ago¹². While oxygen and pH have been recognised as driving forces of massive diversification through their evolutionary history^{12,14}, the evolutionary mechanisms leading to such high phylogenetic and metabolic diversity have received little attention.

A large lateral gene transfer event in the last common ancestor (LCA) of AOA is proposed to have played a major role in their transition to a chemolithoautotrophic lifestyle^{12,15}. However, there is currently little evidence for other cases of large lateral gene acquisition in Thaumarchaeota evolution^{16,17}, and little is known of the relative contributions of gene duplications and genes losses. Large-scale events of lateral gene transfer (LGT) have been suggested as major drivers of proteome evolution in diverse archaeal lineages¹⁸, including Thaumarchaeota¹⁶. However, the long evolutionary history of Thaumarchaeota (>2.3 billion years of evolution) and their diversification in a range of ecosystems^{2,12,14} raises the possibility that distinct evolutionary mechanisms have shaped thaumarchaeotal genomes across their history. In particular, we hypothesise that an ongoing process of LGT spread across thaumarchaeotal evolution (in contrast, or addition to acquisitions at their origin) was the major source of molecular innovation. Phylogenomic methods based on reconstructed ancestral gene contents have been used previously throughout the archaeal radiation to explicitly model gene family acquisitions, duplications and losses using a recently-developed approach for probabilistic gene mapping¹⁹ by amalgamated likelihood estimation (ALE)²⁰. However, the aforementioned study only included a limited number of Thaumarchaeota. Similarly, a recent comparative analysis of a diverse set of thaumarchaeotal genomes examining the influence of oxygen availability on diversification did not include the large diversity of terrestrial thaumarchaeotal genomes presented here, nor disentangle the contributions of gene transfer and gene duplication to genomic diversification¹⁵.

Here we present 12 high-quality metagenome-assembled genomes (MAGs), providing the first genome representatives of three of the six known major lineages of the prevalent terrestrial order of AOA, the Nitrososphaerales. This enables the reconstruction of a strongly supported phylogeny for the Nitrososphaerales, which was not previously achieved using single-gene markers. Then, we investigated the mechanisms of genome evolution in the

Thaumarchaeota phylum, quantifying the relative contributions of lateral gene transfers, gene duplications and gene losses. Phylogenomic analysis reveals two previously uncharacterised large lateral gene transfer events followed by subsequent extensive gene duplication mechanisms that have shaped the evolution of Nitrososphaerales.

Results

Thaumarchaeotal phylogenomic diversity. Twelve thaumarchaeotal genomes of 63–97% completeness (Supplementary Data 1) were reconstructed by large-scale co-assembly of 171 metagenomes from sediment samples collected from 23 sites along the River Thames, UK. Phylogenomic analysis of 75 concatenated single-copy orthologues retrieved from 152 available thaumarchaeotal genomes (including the 12 genomes assembled in this study) (Supplementary Data 1) provided a well-supported thaumarchaeotal phylogenomic tree (Fig. 1), with most nodes with ultrafast bootstrap (UFBoot) values >95% and Shimodaira and Hasegawa-like approximate likelihood ratio test (SH-aLRT) values >95%. This phylogenomic tree was the best-supported tree after comparison of several marker gene selections (SI: Extended phylogenomics). The tree is highly congruent with previously published work¹², with the exception of several early-diverging lineages (SI: Extended phylogenomics). Previous classifications of Thaumarchaeota have been mainly based on the single marker 16S rRNA gene^{5,21–23} with the exception of a few studies using concatenated protein trees^{12,15,24}, while the ammonia monooxygenase (*amoA*) gene was used as a proxy marker for classification of AOA^{2,4,14}, as *amoA* gene is not present in all members of the phylum. Classification of non-AOA within Thaumarchaeota is based on monophyletic placement rather than unified metabolism at the phylum level. We therefore suggest a possible thaumarchaeotal taxonomic ranking based on our phylogenomic analysis that maintains maximum consistency with previous work while incorporating the early-diverging lineages that lack *amoA*^{2,25,26} (Supplementary Data 2). Our thaumarchaeotal genome dataset represents a diverse phylum comprising 8 classes (1 AOA, 7 non-AOA), 10 orders (3 AOA, 7 non-AOA), 28 families (18 AOA, 10 non-AOA), 31 genera (20 AOA, 11 non-AOA) and 103 species (87 AOA, 16 non-AOA). While the classically used thaumarchaeotal nomenclature is congruent with the taxonomic stratification, a few exceptions were observed. For example, the order Nitrosopumilales encompasses the formally defined order *Candidatus* Nitrosotaleales²⁷, whose classification may now need to be reconsidered. In addition, this taxonomic stratification indicates that Nitrososphaerales contains a minimum of 8 genera, with *Ca. Nitrososphaera gargensis* belonging to a different genus to *Nitrososphaera viennensis* and *Ca. Nitrososphaera evergladensis*.

We also used the Genome Taxonomy Database Toolkit (GTDB-Tk) to evaluate the genomic diversity of our 12 river sediment MAGs (Supplementary Data 3). While one of these genomes (TH1173) is a close relative of a published Nitrososphaerales genome, *Ca. Nitrosocosmicus oleophilus* MY3, the range of relative evolutionary divergence (RED)²⁸ values for the other 11 MAGs was 0.72–0.82, suggesting that some of these MAGs represent new family-level lineages (Supplementary Data 3). These results are consistent with our phylogenomic analysis, in which these genomes represent novel deep branches within Nitrososphaerales (Fig. 1). Assessment of the diversity of the genome set revealed that specific lineages are disproportionately represented in terms of genome sequence based on the diversity indicated by *amoA* gene studies (Supplementary Fig. 1), as observed previously². While the Nitrososphaerales genomic diversity is now relatively well represented with the addition of

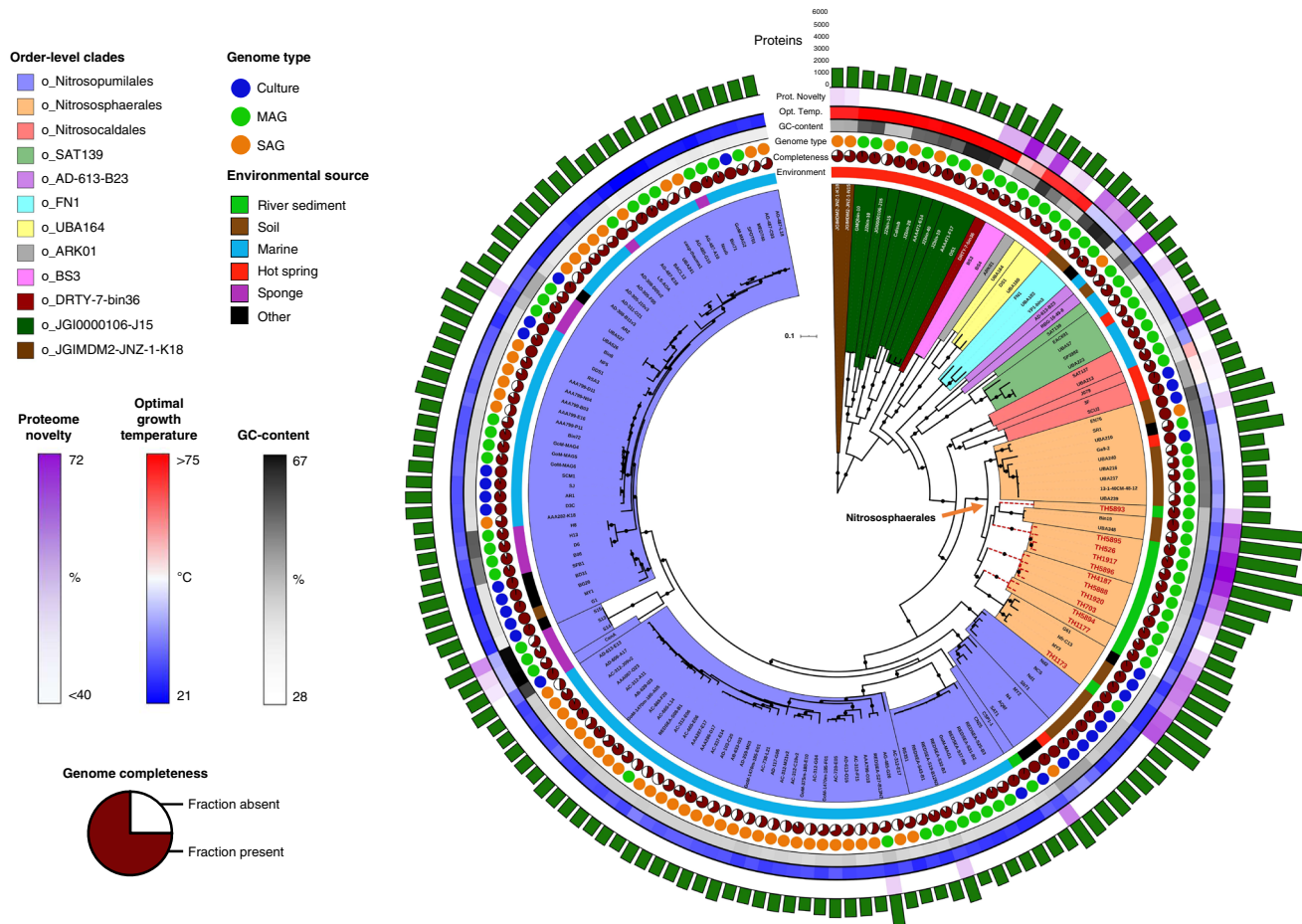


Fig. 1 Phylogenomic tree of Thaumarchaeota. This tree includes 152 thaumarchaeotal genomes (29 culture genomes, 51 single-cell-assembled genomes (SAG) and 72 metagenome-assembled genomes (MAGs), including 12 novel MAGs obtained in this study) from a range of environments and 11 Aigarchaeota genomes and is rooted with two Bathyarchaeota genomes. The tree was inferred by maximum likelihood from 75 concatenated phylogenetic markers, which were aligned separately and analysed using the best-fitting model for each alignment. The 12 new genomes assembled in this study are indicated by red tip labels. Clades that previously lacked representative genomes are indicated with dotted red branches. Dots indicate branches with >95% UFboot and SH-aLRT support. Coloured order-level clades are further divided into family-level clades with black clade borders. Protein novelty is defined as the percentage of encoded proteins that lack a close homologue (e -value $<10^{-5}$, % ID >35 , alignment length >80 and bit score >100) in the arCOG database. Detailed genome information is given in Supplementary Data 1.

these genomes, several major clades of the Nitrosopumilales lack representative genomes. Previously, broad groups loosely reflecting the origins of thaumarchaeotal isolates and genomes were described as deep-water AOA, shallow-water AOA, terrestrial AOA and basal (non-ammonia-oxidising) Thaumarchaeota¹². The deep-water AOA group (clade NP-Alpha-2) is the least diverse set of genomes, with 33 genomes affiliating to a single genus (max. divergence 26%), closely followed by the shallow-water AOA group (clades NP-Epsilon-2; NP-Gamma-2; NP-Theta-4; NP-Delta-2), with 62 genomes forming 4 genera (max. divergence 44%) (Supplementary Fig. 2). The terrestrial AOA group (clades NP-Eta, NT-Alpha and NS) represent the most diverse AOA set of genomes, with 40 genomes forming 15 genera (max. divergence 48%). However, the basal thaumarchaeotal group appears to be even more diverse with 17 genomes forming 11 genera (max. divergence 57%) (Supplementary Fig. 2).

The genomes used in this study presented relatively high completeness (average of 82%) with lowest values being reported for single-cell genomes from the marine environment (Fig. 1), low contamination (average of 1%) and variable strain heterogeneity (average of 34%) (Supplementary Data 1). Despite the wide range of genome size across the phylum (0.8–5.2 Mbp), the data clearly support higher genome sizes in some of the terrestrial

Nitrososphaerales. This higher genome size was noted previously in several cultured Nitrososphaerales²⁹ and is particularly pronounced in the MAGs within the genera *g_Bin19* and *g_TH5895*. In agreement with the findings of Alves et al.², GC content is closely related to phylogeny in most cases, with a notable exception being the sponge-associated Thaumarchaeota, which present a higher GC content, potentially due to adaptation to environmental stresses such as nutrient and energy limitation³⁰. Prediction of optimal growth temperatures (OGTs) confirmed that most of the Nitrososphaerales and Nitrosopumilales are mesophilic with OGTs between 21 and 37 °C, with a few exceptions such as organisms affiliating to three closely related terrestrial genera (*g_EN76*, *g_UBA210* and *g_Nitrososphaera*) or to two genera containing exclusively sponge-associated Thaumarchaeota (*g_Cenarchaeum* and *g_S14*) with slightly higher temperature optimum (range 33–42 °C) (see SI: Extended optimal growth temperature). As predicted, Nitrosocaldes have higher OGTs (range 32–56 °C), but these estimates are lower than measured OGT in the two cultures available (difference of 15 and 19 °C, SI: Extended optimal growth temperature). Predicted OGTs among the deeply-rooted non-AOA Thaumarchaeota are consistent with the hypothesis that Thaumarchaeota evolved from a hyperthermophilic common ancestor (Fig. 1 and

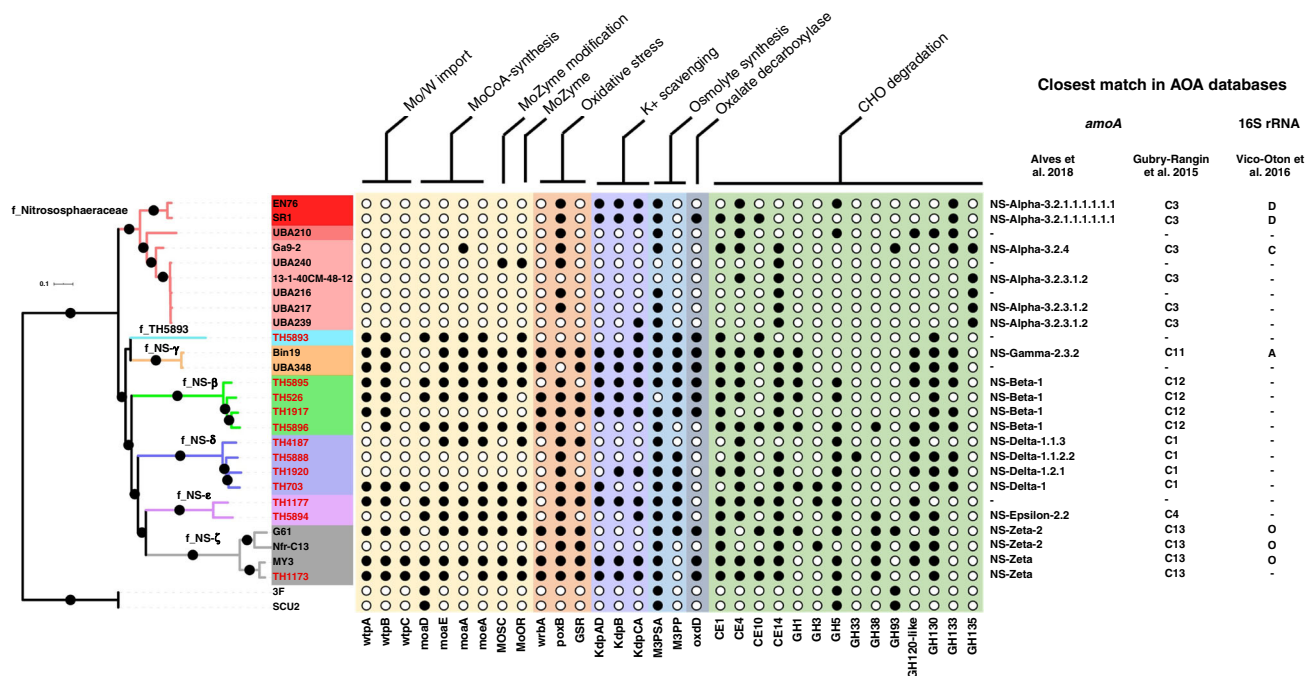


Fig. 2 Phylogeny and distinctive metabolic traits of Nitrososphaerales. This tree includes 26 Nitrososphaerales genomes and was inferred by maximum likelihood reconstruction from 188 concatenated marker genes. The 12 new genomes obtained in this study are indicated by red tip labels and the tree is rooted with the *Ca. Nitrosocaldus* strains 3F and SCU2. This tree has similar topology to the phylum-wide tree, but with strong support for almost all branches. Dots indicate branches with >95% UFBoot and SH-aLRT support. Colours on branches and leaf labels indicate families and genera, respectively. While many of the genes acquired in early stages of Nitrososphaerales evolution are uncharacterised, several genes involved in molybdoenzymes synthesis and stress response could be identified, among which several are represented on the figure. The presence and absence of those genes are indicated by a filled or empty circle, respectively. The following genes were used in the figure: *wtpABC*, Wtp transport system subunits A, B and C; MOSC, molybdenum cofactor sulfurase; MoOR, Molybdopterin oxidoreductase; *wrbA*, NAD(P)H quinone oxidoreductase; *poxB*, pyruvate oxidase; GSR, glutathione-disulfide reductase; *kdpABC*, potassium transport system subunits A, B and C; M3PSA, mannosyl-3-phosphoglycerate synthase; M3PP, mannosyl-3-phosphoglycerate phosphatase; CE, carbohydrate esterase family; GH, glycoside hydrolase family. Closest BLASTn⁶² matches in the Alves², Gubry-Rangin¹⁴ and Vico-Oton²¹ databases are indicated for the Nitrososphaerales genomes on the right of the figure. Genomes in which *amoA* or 16S rRNA genes were not detected are indicated with a hyphen (-).

Supplementary Data 1), but the presence of the marine mesophiles SAT137 and UBA213 in Nitrosocaldus (both sharing 84% amino acid identity on the *amoA* gene with the thermophilic *Ca. Nitrosocaldus islandicus* 3F) and mesophiles in their closest non-AOA relatives (*o_FN1*, *o_AD-613-B23* and *o_SAT139*) are consistent with the hypothesis that the last common ancestor of AOA was a mesophile, placing the hot origin of Thaumarchaeota further back in evolutionary history than previously proposed^{15,23}.

Phylogeny and metabolic traits of the Nitrososphaerales lineage. The 12 MAGs reconstructed in this study represent the first genome representatives of three of the six Nitrososphaerales families, enabling a detailed analysis of this order. As the whole-phylum analysis did not result in a highly supported phylogeny of the Nitrososphaerales order, a more targeted phylogenomic analysis based on 188 Nitrososphaerales markers was performed to increase resolution (Fig. 2). Previously available genomes and the new MAGs represent 22 species and 9 genera within Nitrososphaerales. The basal tree topologies of this order based on the widely-used single-gene marker *amoA*^{2,14} or on 188 single-copy core markers are different, but a phylogenomically-informed rooting of the *amoA* phylogeny resolved many of the incongruencies (Supplementary Fig. 3). This phylogenomic approach also provides robust resolution (>95% ultrafast bootstrapping) of the basal relationship between the lineages NS-alpha, NS-beta and NS-gamma, which was not obtained in previous studies using the *amoA* phylogeny alone. Moreover, the genome TH5893 appears

to represent a previously undiscovered seventh family of Nitrososphaerales, but this remains unclear as its relationship to NS-gamma was not resolved with high confidence (Fig. 2) and *amoA* and 16S rRNA genes were not detected in this MAG. The genomes in this work represent the known family-level diversity of Nitrososphaerales, but many of the lower taxonomic ranks identified in *amoA*-based studies have no genome sequenced. Despite the greater statistical power of the phylogenomic data compared to the *amoA* gene approach, the limited diversity coverage may have potential phylogenetic biases and future increased genome representation in these ranks may improve phylogeny prediction in this group.

AOA are a highly metabolically conserved group of chemolithoautotrophic organisms, which use ammonia for energy³¹ and fix CO₂ as a carbon source³². They also have the capacity to synthesise many of the cofactors³³ and amino acids²⁹ required for their cellular function. Despite this high level of metabolic conservation, Nitrososphaerales AOA possess additional metabolic capacities that likely facilitated their expansion into a wide variety of soil, wastewater and river sediment environments³⁴ (Fig. 2). Among the traits acquired early in the Nitrososphaerales diversification, many likely contribute to protection against osmotic and oxidative stress, production of molybdoenzymes and carbohydrate utilisation (Supplementary Data 7).

Carbohydrate active enzymes (CAZymes) are ubiquitous in the tree of life, allowing organisms to degrade environmental carbohydrates as energy and carbon sources (such as the utilisation of monosaccharides from cellulose or hemicellulose

degradation) and for the formation of more complex compounds from simpler substrates (such as extracellular polysaccharide synthesis)³⁵. The number of CAZyme genes in the Nitrososphaerales lineage is large and greater than that of other members of the phylum, as the other Thaumarchaeota encode an average of less than two CAZymes. The last common ancestor of the Nitrososphaerales is predicted to have acquired the amylo- α -1,6-glucosidase GH133 (EC3.2.1.33) and polyspecific carbohydrate esterases CE1, CE4 and CE10. In addition, the beta-mannanase GH130, alpha-mannanase GH38, hemicellulases GH1, GH3 and GH5, and a putative chitin disaccharide deacetylase (CE14) were also acquired during Nitrososphaerales diversification.

Since these organisms lack the phosphofructokinase necessary to complete the glycolytic pathway, it is difficult to determine the purpose of these carbohydrate degrading enzymes. One possible explanation is that the resulting monosaccharides are being utilised for the biosynthesis of cellular components such as extracellular polysaccharides or osmolytes. Indeed, the Nitrososphaerales possess mannosyl-3-phosphoglycerate synthase and mannosyl-3-phosphoglycerate phosphatase genes, giving these organisms the potential ability to synthesise the osmoprotectant mannosyl-glycerate from mannose monosaccharides and glycerate 3-phosphate. They have also lost the mannose-6-phosphate isomerase that is present in most other Thaumarchaeota, preventing mannose monosaccharides being shunted into early glycolysis, increasing the amount of mannose available for mannosyl-glycerate synthesis.

Kdp, a high-affinity ATP-driven potassium uptake system that enables K⁺-mediated osmoregulation in potassium limited environments³⁶, is present in several Nitrososphaerales genomes, but almost completely absent from marine Thaumarchaeota (with the exception of REDSEA-S19-B12N3). In the event of osmotic shock, this system could allow the organisms to maintain turgor pressure in soils where there is low potassium concentration or where potassium is bound to negatively charged humus or clay particles by cation exchange.

Nitrososphaerales have also acquired several systems for protection against oxidative stress indicating that there may be additional production of reactive oxygen species (ROS) in the metabolism of this lineage. These include pyruvate oxidase (*PoxB*), NAD(P)H quinone oxidoreductase (*WrbA*) and glutathione-disulfide reductase (*GSR*). *PoxB* catalyses the ubiquinone-dependent oxidative decarboxylation of pyruvate to form acetate and CO₂. This pathway enables the production of acetate in a NAD-independent manner, preventing the accumulation of ROS from resulting NAD⁺ regeneration by NADH dehydrogenase in slow-growing cells where ROS are not diluted by cell growth³⁷. *WrbA* maintains quinones in the fully reduced state, guarding against the production of ROS from one-electron redox cycling³⁸ and *GSR* catalyses the reduction of glutathione disulfide to two reduced molecules of glutathione—an antioxidant that protects cellular components from oxidative stress.

Another notable difference between the Nitrososphaerales and other lineages of AOA is the presence of genes for the uptake of molybdate and synthesis of molybdoenzymes (MoZymes). These transport and synthesis genes were previously detected in three *Ca. Nitrosocosmicus* strains³⁹ and appear to be also present in several of the newly represented Nitrososphaerales lineages (Fig. 2). The *Wtp* transport system, which has high affinity for both molybdate and tungstate, is present in many Nitrososphaerales, but almost completely absent in other Thaumarchaeota (with the exception of the non-AOA Thaumarchaeota ARK01, UBA164 and YP1-bin3). The enzymes *MoaA*, *MoaD/E* and *MoeA* catalyse the formation of molybdenum cofactor (MoCo) of MoZymes and MOSC transfers sulfur to the molybdenum, forming mono-oxo MoCo⁴⁰. MoZymes of the

sulfite oxidase (SO) superfamily were identified in several members of the Nitrososphaerales order (Supplementary Fig. 4). These enzymes appear to be an intermediate between the eukaryotic SO and the bacterial protein-methionine-sulfoxide reductase (Supplementary Fig. 4). The majority of the Nitrososphaerales MoZymes are structurally similar to the eukaryotic SO, possessing both the oxidoreductase molybdopterin binding domain (PF00174) and the MoCo oxidoreductase dimerisation domain (SSF81296), but lacking the N-terminal cytochromes b5 electron transport hemoprotein domain (PF00173). These enzymes, however, possess at their N-terminal several transmembrane helices that are likely to be a novel class of cytochrome domains. We would therefore hypothesise that this protein family is a class of archaeal SO that convert sulfite to sulfate, allowing the generation of ATP in oxidative phosphorylation. This reaction produces hydrogen peroxide as a by-product⁴¹, possibly explaining the need for extra oxidative stress protection mechanisms in this order.

The three MoZymes of the arctic soil MAG, Bin19, and one of the two MoZymes of *Ca. Nitrosocosmicus oleophilus* (MY3_01316) were more similar in structure and sequence to the bacterial protein-methionine-sulfoxide reductase that repairs oxidized periplasmic proteins containing methionine sulfoxide residues than to the eukaryotic sulfite oxidases, indicating a putative role of this enzyme in oxidative stress damage protection.

Influence of evolutionary mechanisms on thaumarchaeotal diversification. The exhaustive creation of probabilistic ancestral reconstructions from every branch of the thaumarchaeotal phylogeny allowed the characterisation and quantification of proteome changes along every lineage (Fig. 3 and Supplementary Fig. 5; Supplementary Data 4). The majority of the predicted 67,400 gene content gains in Thaumarchaeota evolution occurred through 51,653 duplications (77% of gains) of pre-existing genes, with 11,430 intra-phylum gene transfers (intra-LGT) (17% of gains) and 4317 originations (including inter-phyla gene transfers (inter-LGT) and de novo gene formation) (6% of gains). There were also 227,837 gene losses predicted, indicating gene duplication and gene loss as the two most significant drivers of gene content change in Thaumarchaeota evolution. The same ancestral genes were lost along multiple lineages, explaining the higher number (nearly 3.5-fold) of losses than gains.

Four branches within the Nitrososphaerales order (noted A, B, C and D in Fig. 3, panel b) are estimated to undergo particularly high rates of gene duplication, collectively accounting for 11% of duplication events in the phylum (Fig. 3, Supplementary Fig. 5 and Supplementary Data 4), with more than 900 duplication events occurring on each of those branches. Numerically, these are the major mechanisms of predicted proteome expansion in the Thaumarchaeota phylum, resulting in most Nitrososphaerales encoding more than 3000 proteins and possessing an average genome size of 2.7 Mbp (range 1.5–5.2 Mbp). Duplication of existing genes also appears to be the major mechanism of predicted proteome expansion in the *Ca. Nitrosotalea* and *Ca. Nitrosotenuis* genera with their last common ancestors undergoing 1337 and 1010 duplication events, respectively (Supplementary Data 4). This resulted in the extant members of these genera possessing the largest proteomes in the Nitrosopumilales order.

These duplication hotspots resulted in the expansion of several arCOG families (Supplementary Data 5), with copy numbers increasing more than 10-fold in families predicted to be involved in transcription (arCOG04362, arCOG01760 and arCOG01055), signal transduction (arCOG02391), carbohydrate transport (arCOG00144) and coenzyme metabolism (arCOG00972) (Supplementary Data 5). In parallel to these expansions, the copy

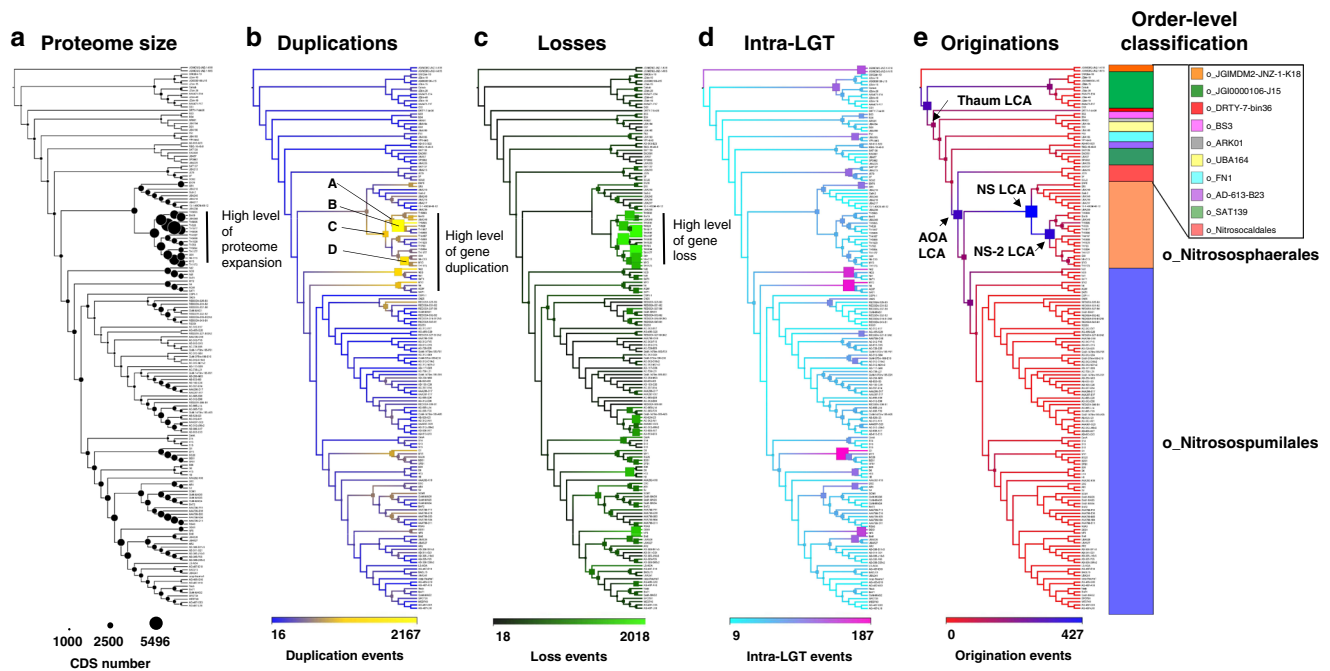


Fig. 3 Quantified mechanisms of proteome evolution. The quantitative and qualitative predictions of the proteome changes were estimated across the thaumarchaeotal history and reported on the thaumarchaeotal cladogram possessing the topology of the ML tree presented in Fig. 1. In addition to the changes in proteome size (panel **a**), four mechanisms of proteome changes were quantified: duplication (panel **b**; copying of a gene within a genome), loss (panel **c**; loss of a gene within a genome), intra-LGT (panel **d**; defined as the acquisition of a gene from other member(s) of the phylum) and origination (panel **e**; defined as the acquisition of a gene from members of phyla outside the sampled genome set (Inter-LGT) or by de novo gene formation). For each cladogram, scale numbers indicate the range of the predicted number of events for a given mechanism. The number of events occurring on each branch is represented by the colour of the associated boxes using the colour legend at the bottom of the cladogram and box sizes increase relatively to the number of events. The order-level classification of the genomes is indicated by the coloured bar on the right. Duplication and loss hotspots within the Nitrososphaerales were highlighted by vertical bars in panels (**b**) and (**c**) and letters A, B, C (which corresponds to NS-2 LCA) and D indicate branches with extensive duplication. Originations occurring in the last common ancestors of all Thaumarchaeota, all ammonia-oxidising archaea, all Nitrososphaerales and a multifamily sub-group within Nitrososphaerales are labelled Thaum LCA, AOA LCA, NS LCA and NS-2 LCA, respectively.

number of other arCOGs decreased (Supplementary Data 6). Interestingly, the three most notable contractions in NS-beta LCA (arCOG01067, NADH dehydrogenase; arCOG04946, lactoylglutathione lyase-like enzyme and arCOG00041, phosphoribosyltransferase) (Supplementary Data 6) actually increased in copy number in its ancestor (NS-2 LCA) (Supplementary Data 5). This indicates a fluctuating expansion and contraction of families throughout evolutionary history, rather than a consistent selective direction.

The evolution of the Nitrososphaerales lineage appears to have involved a significantly larger number of proteome changes than the phylum as a whole ($P < 1e-6$), with greater numbers of duplications ($P < 0.006$) and losses ($P < 5e-5$) explaining much of this difference.

Although duplications are the major mechanism of proteome expansion in Thaumarchaeota, acquisition of a large number of gene families from outside the phylum have also largely contributed. This is particularly true for the LCA of ammonia-oxidising Thaumarchaeota (AOA LCA), where the origination of 320 gene families contributed to the transition to a chemolithoautotrophic ammonia-oxidising lifestyle (Fig. 3 and Supplementary Data 4). Such large gene family gains in AOA LCA have also been predicted in earlier studies^{12,15}. All major origination events (182–427 gene families) have occurred early in the evolutionary history of ammonia-oxidising Thaumarchaeota and accounted for 48% of all gene family originations in the class Nitrososphaeria. In addition, large gene family origination (166 gene families) also occurred in the last common ancestor of the entire thaumarchaeotal phylum (Thaum LCA in Fig. 3).

Two large origination events were predicted in the Nitrososphaerales introducing 427 and 353 new gene families into the last common ancestor of Nitrososphaerales (NS LCA) and an early descendant of NS LCA (NS-2 LCA) (Fig. 4), making these the two largest origination events detected in the phylum (Fig. 3). The majority of the genes originating on these branches (71 and 83% in NS LCA and NS-2 LCA, respectively, Supplementary Data 7) lack close matches in the arCOG database¹², reflecting a high level of proteome novelty (Fig. 1). The source of these originating gene families may be inter-LGT or de novo gene formation. Homologues of the majority of these gene families (84 and 59% in NS LCA and NS-2 LCA, respectively, Supplementary Data 8) were detected in a larger database of archaea, bacteria and eukaryote genomes, indicating their lateral acquisition. Most gene families acquired by NS LCA (Fig. 3) are predicted to have been mainly transferred laterally from other archaea (archaea 95%, bacteria 5%) (Supplementary Data 8), whereas gene families acquired in NS-2 LCA (Fig. 3) were predicted to have been transferred from a variety of archaea (64%), bacteria (34%) and eukaryotes (2%) (Supplementary Data 8). Gene transfer from other archaea was also previously demonstrated in the sister phylum Aigarchaeota¹⁵.

These originating gene families encountered different fates in the different lineages of Nitrososphaerales (Fig. 4), with some families being subject to duplications or losses. Interestingly, half of originating gene families (47 and 54% for NS LCA and NS-2 LCA, respectively) were duplicated within at least one extant member of this order (Fig. 4). The combination of these gene family originations followed by intensive duplications have

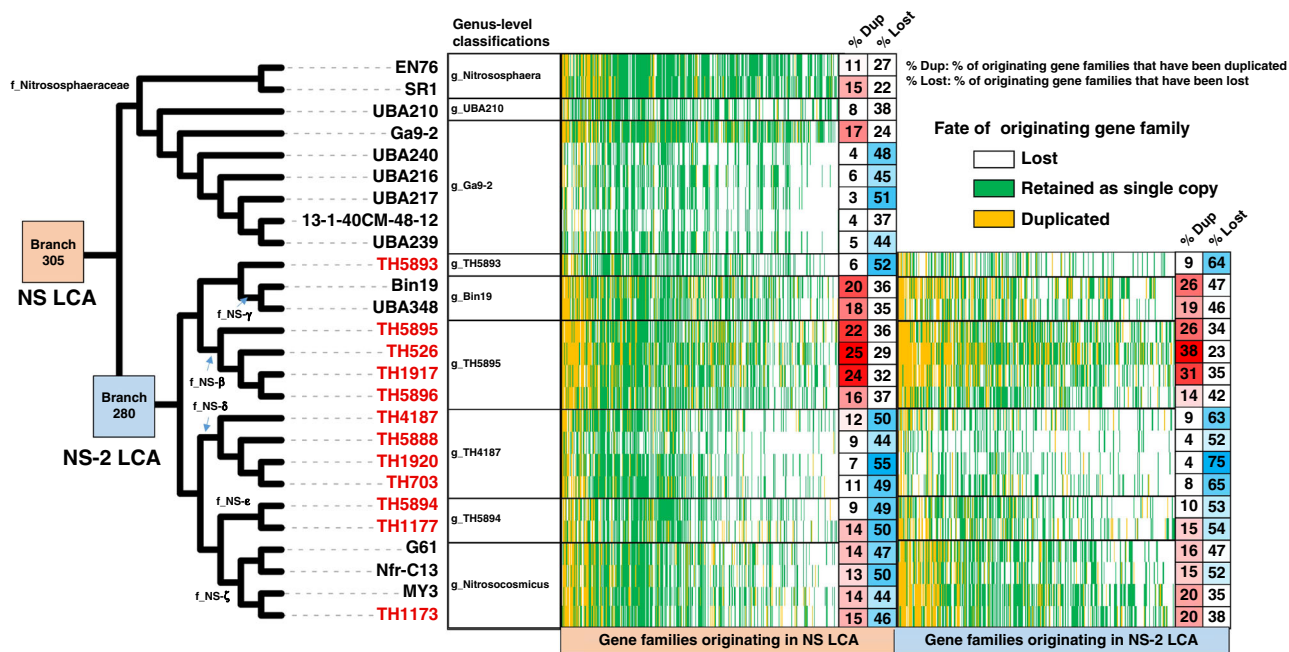


Fig. 4 Fate of the originating gene families in Nitrososphaerales lineages: evidence for subsequent extensive duplication and loss. The cladogram highlights two branches that experienced large origination events, NS LCA and NS-2 LCA. The originating gene families are extensively duplicated in some lineages, such as *g_TH5895*, but also extensively lost in other lineages, such as *g_TH4187*. The ‘% Dup’ and ‘% Lost’ values have been corrected for genome completeness. The 12 new genomes assembled in this study are indicated by red tip labels. Genus-level taxonomic affiliations are presented next to the cladogram tips.

resulted in members of the Nitrososphaerales having among the highest levels of proteome novelty in Thaumarchaeota (Fig. 1). A large number of losses (up to 75% of the originating gene families) occurred in *g_TH4187* and *g_Ga9-2* (Fig. 4). Conversely, there was intensive duplication of originating gene families (up to 29% of the originating gene families) in the *g_TH5895* and *g_Bin19*, the genera with the largest genomes in the phylum, indicating that originating genes are duplicated rather than lost in case of proteome expansion. Similar trends of extensive loss in *g_TH4187* and *g_Ga9-2*, and duplication in *g_TH5895* and *g_Bin19* were observed in ancestral gene families (originating prior to the LCA of Nitrososphaerales) (Supplementary Data 9), but were significantly higher for originating gene families in *g_TH4187* ($P < 0.01$), *g_Ga9-2* ($P < 0.01$) and *g_TH5895* ($P < 0.05$). The lineage-specific nature of gene duplication and loss of these originating gene families also suggests that they may be more important for specialisation of the different Nitrososphaerales lineages into distinct ecological niches than for more general soil and sediment environmental colonisation.

Major transitions in thaumarchaeotal evolution. Thaumarchaeota proteomes have undergone extensive gene content changes during their evolution alongside important environmental niche transitions; this has resulted in a highly metabolically diverse phylum. The nature and influence of these gene content changes were analysed through ancestral reconstruction of the metabolic pathways at key nodes in the thaumarchaeotal phylogeny (Supplementary Data 10).

The most extreme transition in the thaumarchaeotal history was probably the transition from the thaumarchaeotal LCA (Thaum LCA) to the ammonia-oxidising archaeal LCA (AOA LCA). Based on the genomic predictions, Thaum LCA was a heterotrophic hyperthermophilic organism utilising carbohydrates as energy and carbon sources, with a complete glycolytic

pathway (Supplementary Data 10). The absence of genes for biosynthesis of cobalamin, riboflavin and biotin indicate that it acquired vitamins from extracellular organic sources. The transition from the Thaum LCA to AOA LCA was accompanied by several functional gain and loss events (Supplementary Data 11 and 12). Most notably, this transition involved the acquisition of the ammonia monooxygenase genes (*amoABC*), which generate energy by oxidising ammonia to hydroxylamine. This event coincided with the acquisition of urease (EC:3.5.1.5), which converts urea to ammonia, and genes for the biosynthesis of cobalamin, riboflavin and biotin. AOA LCA also acquired the ability to biosynthesise asparagine, alanine and threonine from aspartate through the acquisition of aspartate 4-decarboxylase (EC:4.1.1.12), asparagine synthase (EC:6.3.5.4), aspartate kinase (EC:2.7.2.4) and threonine synthase (EC:4.2.3.1) (Supplementary Data 11 and 12).

The evolutionary transition from Thaum LCA to the ammonia-oxidising lifestyle of AOA LCA involved the loss of glucokinase (EC:2.7.1.2) and phosphofructokinase (EC:2.7.1.11), preventing them from generating ATP by glycolysis. They also lost the glycolysis-associated carbohydrate degrading enzymes alpha-N-arabinofuranosidase (EC:3.2.1.55), alpha-amylase (EC:3.2.1.1), alpha-mannosidase (EC:3.2.1.24) and the CO₂ fixing genes phosphoenolpyruvate carboxylase (EC:4.1.1.31) and pyruvate ferredoxin oxidoreductase (EC:1.2.7.1). These losses were concomitant with the acquisition of the 3-hydroxypropionate/4-hydroxybutyrate CO₂ fixation pathway genes methylmalonyl-CoA epimerase (EC:5.1.99.1) and 4-hydroxybutyryl-CoA dehydratase (EC:4.2.1.120) (Supplementary Data 11 and 12)³².

The establishment of AOA LCA as a chemolithoautotroph was followed by diversification into three order-level groups: Nitrosocaldales, present in hot spring and marine environment; Nitrososphaerales, present mostly in river sediments and soil; and Nitrosopumilales present mostly in marine, soil and sponge-associated environments.

The transition from AOA LCA to the last common ancestor of Nitrosocaldales (NC LCA) involved several functional gains, including geranylgeranyl-glycerol-phosphate geranylgeranyltransferase (EC:2.5.1.42), which is involved in the formation of polar membrane lipids in many thermophilic archaea, and the vitamin-B12-independent methionine synthase (EC:2.1.1.14), functionally replacing the vitamin-B12-dependent methionine synthase (EC:2.1.1.13), which is absent from every member of this order. The DNA polymerase D (EC:2.7.7.7) was also lost in NC LCA. Both the presence of vitamin-B12-independent methionine synthase and the absence of DNA polymerase D were previously reported in *Ca. N. islandicus* 3F⁴² and *Ca. N. cavascurensis* SCU2⁴³, but present results expand those findings to all members of the order. Other notable losses in this transition are the Kdp, high-affinity ATP-driven K⁺-transport system (EC:3.6.3.12), which is involved in osmotic stress resistance in low potassium environments and Uvr excinuclease, which is involved in DNA repair from ultraviolet DNA damage. Kpd is absent from almost all marine Thaumarchaeota in this analysis, indicating that this system is not essential in this environment. It has been proposed that the absence of Uvr in deep-water Nitrosopumilales is due to the lack of light in this environment¹². This theory agrees with the absence of these genes in Nitrosocaldales, which derive from hot spring (in the case of 3F, SCU2 and J079) and deep ocean (in the case of SAT137 and UBA231) environments in which light is also likely limited.

During the transition from AOA LCA to the last common ancestor of Nitrososphaerales (LCA NS), many genes were acquired including the mismatch repair genes DNA adenine methylase (EC:2.1.1.72) and DNA helicase II (EC:3.6.4.12), possibly conferring additional resistance to DNA damage as these organisms expanded into environments with more light exposure such as surface soil and sediment. This transition also involved the loss of the *Phn* phosphonate transport system, which provides a source of phosphate in environments where alternative phosphorus sources are scarce⁴⁴, which may be required less in more nutrient-rich soil and sediments (Supplementary Data 11 and 12).

The evolution of AOA LCA to the last common ancestor of Nitrosopumilales (NP LCA) involved gain of ubiquinone-dependent (EC:7.1.1.2) and ubiquinone-independent (EC:1.6.99.3) NADH dehydrogenases (Supplementary Data 11 and 12), giving NP LCA all three families of respiratory NADH dehydrogenases, which regenerate NAD⁺ from NADH by catalysing the transfer of electrons from NADH to coenzyme Q10 in oxidative phosphorylation. The advantage of encoding all three families remains largely unknown^{45–47}, but could help maintain respiratory function under changing metabolic conditions.

Discussion

As for many prokaryotic lineages, the metabolic diversity of Thaumarchaeota remains to be fully characterized, especially in underexplored environments. The 12 genomes presented in this work offer the first genome representatives for three of the six family-level clades of Nitrososphaerales. This enabled a resolution of Nitrososphaerales phylogeny that was not possible from *amoA* gene-based phylogenies and provided insights into their predicted physiology. Although most of the presently available thaumarchaeotal genomes belong to the marine groups, the majority of thaumarchaeotal phylogenetic diversity is present in non-ammonia-oxidising Thaumarchaeota and in AOA from terrestrial environments. In addition to phylogenetic diversity, substantial genome novelty is also predicted in these organisms, with the Nitrososphaerales possessing the most novel gene families in the ammonia-oxidising Thaumarchaeota, resulting

from two large lateral gene transfer events in their history. The class Nitrososphaeria represent the most environmentally widespread archaea on the planet, despite their apparent specific metabolic niche for ammonia oxidation. The present work confirmed that core metabolic capacities are maintained for ammonia oxidation and autotrophy but, importantly, revealed that their diversification is linked to acquisition of many different traits to survive in diverse environments, explaining the previously suggested vast protein family differences required for environment-specific adaptation^{17,29,39,43}. The transition of Thaumarchaeota to an autotrophic ammonia-oxidising lifestyle has been proposed to result from a large lateral gene transfer event in the last common ancestor of AOA¹². The genomes presented in this study enabled the detailed analysis of the evolutionary history of Nitrososphaerales gene families, identifying two additional LGT events associated with large gene gains and extensive molecular innovation. These findings are relevant to understand the diversification of an entire phyla and suggest that further genome sampling from this group would probably allow identification of other molecular innovations.

It has been proposed that archaeal genome evolution is driven by punctuated episodes of extensive gene acquisition, followed by lineage-specific gene loss⁴⁸. While our analysis indicates that gene transfer contributed to the evolution of Nitrososphaerales, the predominant mode of genome expansion was gene duplication of both ancestral and originating gene families, with higher duplication rates occurring in the latter. As suggested previously⁴⁹, the duplication of newly acquired genes by LGT may facilitate higher gene dosage of a novel metabolic function, which can enable increased production of a gene product in high demand⁵⁰. Duplicated genes can diverge sufficiently to perform a novel function^{51–53} (neofunctionalisation) or specialise to perform the same function under different environmental conditions, such as pH⁵⁴ or temperature⁵⁵ (subfunctionalisation). Therefore, duplication may have a role in Nitrososphaerales adaptation, allowing the maintenance of essential functions in fluctuating and heterogeneous terrestrial environments⁵⁶. While gene duplication has been previously reported in archaea and bacteria, its importance in comparison to LGT may have been underappreciated, at least for some lineages⁴⁹. This unexpected importance of duplication in the evolution of Nitrososphaerales led us to reinterpret the results of a previously archaea-wide evolutionary analysis¹⁹, which revealed that a similar pattern of large origination events (>500 gene families) and subsequent duplications (>350 duplications) may have also occurred in other archaeal lineages, especially for the lineages with the largest known archaeal genomes (including *Haloarcula marismortui* and *Haloferax volcanii*)¹⁹. Taken together, these results suggest that duplication of ancestrally acquired genes may be an important mechanism of genome expansion across a number of archaeal lineages. Larger genome sizes have been reported for mesophilic soil prokaryotes (both archaea and bacteria) compared to marine relatives⁵⁷, but it is unclear whether this is due to genome expansion in soil lineages from ancestors with smaller genomes, or from genome reduction in marine lineages. This work provides an example of a case in which soil-marine genome size differences are driven by genome expansion in soil lineages. While gene duplication is a well-documented mechanism in eukaryotic evolution⁵⁸, demonstration of its importance in thaumarchaeotal evolution requires further investigation across other microbial lineages.

Methods

Sampling, sequencing and genome assembly. Sediment samples were taken from 23 sites spread across 7 tributaries of the River Thames, UK, at three different times, with duplicate samples taken in the summer of 2015 and triplicate samples

taken in the summer and winter of 2016. DNA extraction was performed from 0.5 g of sample within 12 h of collection using the FastDNA™ spin kit for soil, following manufacturer's protocol. Libraries were constructed with TruSeq® DNA library preparation kit and sequenced on the HiSeq2500 platform (2 × 150 bp reads). Sequences were obtained from 171 samples at an average read depth of 45 million, resulting in 2.3 Tbp. The 171 samples were co-assembled using MEGAHIT⁵⁹ with default parameters giving a 59.3 Gbp assembly with an N50 of 999 bp and a total of 4,340,028 contigs >2000 bp in length. Reads from every sample were mapped back onto these contigs using bwa-mem⁶⁰ with default parameters and coverage depths were calculated. These coverage depths and the sequence composition of the contigs were used by the CONCOCT clustering algorithm⁶¹ to produce 6008 genomic bins. Twelve of these bins were found to possess at least one of the *amoA* or *amoB* genes that are characteristic of ammonia-oxidizing archaea, detected using BLASTn⁶² against custom databases of *amoA* and *amoB* sequences²¹.

Sequence database. Thaumarchaeotal (both AOA and non-AOA lineages) genomes were searched in literature and downloaded from NCBI (www.ncbi.nlm.nih.gov) and IMG (<https://img.jgi.doe.gov/>) genome reference databases (final data collection in May 2019). In addition, 12 MAGs from the Thames river were assembled in this study. Thirteen diverse genomes of the related archaeal lineages Aigarchaeota (11 genomes) and Bathyarchaeota (two genomes) were also added to this database. All genomes were annotated using Prokka v1.11⁶³ to ensure uniformity between annotation methods and genome completeness and contamination were estimated using CheckM⁶⁴. Only genomes with completeness higher than 45% and presenting <10% contamination were kept in the final dataset (to avoid bias in phylogenomic analyses), reaching a total of 152 thaumarchaeotal genomes (29 culture genomes, 51 single-cell genomes and 72 MAGs (Supplementary Data 1)). These completeness and contamination thresholds were based on a balance between genome completeness and taxonomic coverage, and were chosen specifically for this work, and may not be appropriate benchmarks for other evolutionary genomic studies. Several genome characteristics were estimated, including GC content (using QUAST⁶⁵), total predicted genomic size (measured genome size corrected by the completeness score) and predicted optimal growth temperature (based on a machine learning model Tome⁶⁶). Environmental sources were also retrieved from the genome reference databases or from the associated published study (Supplementary Data 1). As detailed current phylogeny of Thaumarchaeota is based on the *amoA* or 16S rRNA genes^{2,21,25}, these two phylogenetically congruent markers²¹ were retrieved using RNAMmer⁶⁷ or BLASTp⁶² against two representative *amoA* databases^{2,25}. Their classification towards the most up-to-date Thaumarchaeota *amoA* and 16S rRNA gene databases^{2,21} was performed using BLASTn⁶². Initial classification of genome sequences was performed using classify_wf in the GTDB-Tk package⁶⁸.

Phylogenomic reconstruction and proteome content changes across evolutionary history

Datasets. Two datasets were used in this study: a phylum-level phylogenomic reconstruction including 165 genomes with a completeness >45% and contamination lower than 10%, and Nitrososphaerales-level phylogenomic reconstruction performed on the 26 genomes belonging to this lineage.

Ortholog selection. The best of three ortholog selection methodologies was selected (see SI: Extended phylogenomics). Ortholog groups (OGs) were detected using Roary (-i 50, -iv 1.5)⁶⁹ and core OGs were defined as those that were present in a single copy in each genome and were present in at least 85% of the near complete genomes. Core OGs were aligned individually using MAFFT L-INS-i⁷⁰ and spurious sequences and poorly aligned regions were removed with trimal (automated1, resoverlap 0.55 and seqoverlap 60)⁷¹. Alignments were removed from further analysis if they presented evidence of recombination using the PHITest⁷². Such ortholog prediction using the MCL algorithm resulted in 75 and 188 marker genes for the phylum-level and the Nitrososphaerales-level phylogenomic reconstruction, respectively (workflow illustrated in Supplementary Fig. 6).

Phylogenomic tree construction. Phylogenomic reconstruction was performed on each dataset, using a concatenated supermatrix of core OG alignments. Each maximum likelihood tree was constructed with IQ-TREE⁷³ using the best-fitting protein model predicted in ModelFinder⁷⁴ and an edge-linked partition model (model details in Supplementary Data 13). Branch supports were computed using the SH-aLRT test⁷⁵ and 2000 UFBoot replicates and a hill-climbing nearest neighbour interchange (NNI) search was performed to reduce the risk of over-estimating branch supports. Trees were rooted with the Bathyarchaeota genomes.

Proteome content changes across evolutionary history. For the phylum-level dataset, protein families were detected with Roary with reduced stringency (-i 35, -iv 1.3, -s) and sequences shorter than 30 amino acids and families with <4 sequences were removed from further analysis. All remaining sequences within each family were aligned using MAFFT L-INS-I, processed with trimal (automated1) and ML phylogenetic trees were constructed for each alignment (using IQ-TREE⁷³ with 1000 UFBoot replicates, a NNI search and the best substitution model selected by

ModelFinder⁷⁴). Ninety-seven percent (5683 of the 5850) of the protein family trees could be probabilistically reconciled against the supermatrix tree using the ALEml_unated algorithm of the ALE package²⁰ to infer the numbers of duplications, losses, intra-LGT (transfer within the sampled genome set) and originations (including both inter-LGT, e.g. transfer from other phyla outside the sampled genome set or de novo gene formation) on each branch of the supermatrix tree. Genome incompleteness was probabilistically accounted for within ALE using the genome completeness values indicated by CheckM⁶⁴. For the purposes of this analysis the small number of genes transferred between the related phyla Thaumarchaeota, Aigarchaeota and Bathyarchaeota genomes studied here are deemed intra-LGT events, not origination events. Ancestral proteome content (both size and predicted metabolic pathways) was estimated for each node of the tree. All phylogenomic trees were visualised using FigTree (<http://tree.bio.ed.ac.uk/software/figtree/>) and iTOL⁷⁶.

Taxonomic affiliation. Internally consistent taxonomic levels were calculated using an agglomerative clustering approach. Average amino acid identities (AAIs) between pairwise sets of genomes were calculated using CompareM (<https://github.com/dparks1134/CompareM>) and AAIs hierarchical clustering was performed in MeV (<https://sourceforge.net/projects/mev-tm4/>) using Euclidean distance with complete linkage. Maximum divergence is defined as the lowest AAI between any two members of the same group, subtracted from 100%. Pairwise relative phylogenetic distances between each genome were subjected to hierarchical clustering using single linkage Pearson correlation and taxonomic levels corresponding to class, order and family were defined as Pearson's distances of less than 0.34, 0.13 and 0.015, respectively. These cut-off distances were chosen by comparison to existing thaumarchaeotal taxonomy^{5-8,42,43,77-81}. The taxonomic levels for genus and species levels were defined with AAI thresholds higher than 70% and 95%, respectively^{82,83}. The nomenclature of the different taxonomic stratifications relied primarily on previously existing classifications, with attribution of the names being primarily based on the strains whose cultivation was first reported. For clusters with no reported culture, the family-level nomenclature inferred from previously ammonia monooxygenase phylogenetic tree² was used. For groups not matching those affiliations, such as genomes lacking the ammonia monooxygenase, the first listed genome of the group was used to name the group. In addition, the topology of a subset of the phylogenomic tree (focusing on ammonia-oxidising Thaumarchaeota) was compared to previously published *amoA*-based phylogenies^{2,14} at the family level.

Functional annotation and metabolic reconstruction. For each protein family, a medoid sequence (the sequence with the shortest summed genetic distances to all other 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 database⁸⁴ using GhostKOALA⁸⁵, against the arCOG database⁸⁶ using Diamond BLASTp⁸⁷ (best-hit and removing matches with e-value >10⁻⁵, % ID < 35, alignment length < 80 or bit score < 100) and against the Pfam database⁸⁸ using hmmscan⁸⁹ (-T 80). Carbohydrate active enzymes were annotated using HMM models from dbCAN (<http://ccb.unl.edu/dbCAN2/>) (filtered with hmmscan-parser.sh and by removing matches with mean posterior probability < 0.7). The predicted protein families present within each ancestor genome was inferred from the ALE results, allowing for ancestral metabolic reconstruction at each node of the phylogenetic tree. Proteome novelty in extant genomes was defined as the percentage of proteins encoded by a genome that do not have a homologue in the arCOG database. Gene family originations were predicted to be the result of inter-LGT if the medoid of the family possesses homologues in a database of UniRef90⁹⁰ sequences with strain-level designations and excluding thaumarchaeotal matches. The origins of these putatively laterally acquired gene families were estimated using the best match in this database.

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

Data availability

Genome sequences assembled in this work are available from Genbank under the accession numbers JAATV1000000000 (Nitrososphaerales archaeon TH1173), JAATVJ0000000000 (Nitrososphaerales archaeon TH1177), JAATVK0000000000 (Nitrososphaerales archaeon TH5894), JAATVL0000000000 (Nitrososphaerales archaeon TH703), JAATVM0000000000 (Nitrososphaerales archaeon TH1920), JAATVN0000000000 (Nitrososphaerales archaeon TH5888), JAATVO0000000000 (Nitrososphaerales archaeon TH4187), JAATVP0000000000 (Nitrososphaerales archaeon TH5896), JAATVQ0000000000 (Nitrososphaerales archaeon TH1917), JAATVR0000000000 (Nitrososphaerales archaeon TH526), JAATVS0000000000 (Nitrososphaerales archaeon TH5895) and JAATVT0000000000 (Nitrososphaerales archaeon TH5893). Public data is available from NCBI (www.ncbi.nlm.nih.gov), IMG (<https://img.jgi.doe.gov/>), KEGG (<https://www.genome.jp/kegg/>), dbCAN (<http://ccb.unl.edu/dbCAN2/download/>) and arCOG (<https://ftp.ncbi.nih.gov/pub/wolf/COGs/arCOG/>).

Code availability

Custom scripts for manipulating ALE outputs have been deposited on GitHub at <https://github.com/Tancata/phylo/tree/master/ALE> (<https://doi.org/10.5281/zenodo.4012549>)⁹¹.

Received: 20 April 2020; Accepted: 21 September 2020;

Published online: 30 October 2020

References

- Gubry-Rangin, C., Williams, W. & Prosser, J. I. Approaches to understanding the ecology and evolution of understudied terrestrial archaeal ammonia-oxidisers. *Emerg. Top. Life Sci.* **2**, 619–628 (2018).
- Alves, R. J. E., Minh, B. Q., Ulrich, T., von Haeseler, A. & Schleper, C. Unifying the global phylogeny and environmental distribution of ammonia-oxidising archaea based on amoA genes. *Nat. Commun.* **9**, 1–17 (2018).
- Kitzinger, K. et al. Single cell analyses reveal contrasting life strategies of the two main nitrifiers in the ocean. *Nat. Commun.* **11**, 1–12 (2020).
- Kerou, M. & Schleper, C. *Bergey's Manual of Systematics of Archaea and Bacteria*, 1–10 (2015).
- Stieglmeier, M. et al. *Nitrososphaera viennensis* gen. nov., sp. nov., an aerobic and mesophilic, ammonia-oxidizing archaeon from soil and a member of the archaeal phylum Thaumarchaeota. *Int. J. Syst. Evol. Microbiol.* **64**, 2738–2752 (2014).
- Qin, W. et al. *Nitrosopumilus maritimus* gen. nov., sp. nov., *Nitrosopumilus cobalaminigenes* sp. nov., *Nitrosopumilus oxycliniae* sp. nov., and *Nitrosopumilus ureiphilus* sp. nov., four marine ammonia-oxidizing archaea of the phylum Thaumarchaeota. *Int. J. Syst. Evol. Microbiol.* **67**, 5067–5079 (2017).
- Prosser, J. I. & Nicol, G. W. *Bergey's Manual of Systematics of Archaea and Bacteria*, 1–7 (2015).
- Qin, W. et al. *Bergey's Manual of Systematics of Archaea and Bacteria*, 1–2 (2015).
- Mussmann, M. et al. Thaumarchaeotes abundant in refinery nitrifying sludges express amoA but are not obligate autotrophic ammonia oxidizers. *Proc. Natl Acad. Sci. USA* **108**, 16771–16776 (2011).
- Weber, E. B., Lehtovirta-Morley, L. E., Prosser, J. I. & Gubry-Rangin, C. Ammonia oxidation is not required for growth of Group 1.1 c soil Thaumarchaeota. *FEMS Microbiol. Ecol.* **91**, fiv001 (2015).
- Lin, X., Handley, K. M., Gilbert, J. A. & Kostka, J. E. Metabolic potential of fatty acid oxidation and anaerobic respiration by abundant members of Thaumarchaeota and Thermoplasmata in deep anoxic peat. *ISME J.* **9**, 2740–2744 (2015).
- Ren, M. et al. Phylogenomics suggests oxygen availability as a driving force in Thaumarchaeota evolution. *ISME J.* **13**, 2150–2161 (2019).
- Kato, S. et al. Isolation and characterization of a thermophilic sulfur- and iron-reducing thaumarchaeote from a terrestrial acidic hot spring. *ISME J.* **13**, 2465–2474 (2019).
- Gubry-Rangin, C. et al. Coupling of diversification and pH adaptation during the evolution of terrestrial Thaumarchaeota. *Proc. Natl Acad. Sci. USA* **112**, 9370–9375 (2015).
- Hua, Z. et al. Genomic inference of the metabolism and evolution of the archaeal phylum Aigarchaeota. *Nat. Commun.* **9**, 1–11 (2018).
- López-García, P., Zivanovic, Y., Deschamps, P. & Moreira, D. Bacterial gene import and mesophilic adaptation in archaea. *Nat. Rev. Microbiol.* **13**, 447–456 (2015).
- Herbold, C. W. et al. Ammonia-oxidising archaea living at low pH: insights from comparative genomics. *Environ. Microbiol.* **19**, 4939–4952 (2017).
- Nelson-Sathi, S. et al. Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* **517**, 77–80 (2015).
- Williams, T. A. et al. Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc. Natl Acad. Sci. USA* **114**, E4602–E4611 (2017).
- Szöllösi, 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).
- Oton, E. V., Quince, C., Nicol, G. W., Prosser, J. I. & Gubry-Rangin, C. Phylogenetic congruence and ecological coherence in terrestrial Thaumarchaeota. *ISME J.* **10**, 85–96 (2016).
- Fuhrman, J. A., McCallum, K. & Davis, A. A. Novel major archaeobacterial group from marine plankton. *Nature* **356**, 148–149 (1992).
- De, L. A. et al. Cultivation of a thermophilic ammonia oxidizing archaeon synthesizing crenarchaeol. *Environ. Microbiol.* **10**, 810–818 (2008).
- Brochier-Armanet, C., Boussau, B., Gribaldo, S. & Forterre, P. Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nat. Rev. Microbiol.* **6**, 245–252 (2008).
- Gubry-Rangin, C. et al. Niche specialization of terrestrial archaeal ammonia oxidizers. *Proc. Natl Acad. Sci. USA* **108**, 21206–21211 (2011).
- Pester, M., Schleper, C. & Wagner, M. The Thaumarchaeota: an emerging view of their phylogeny and ecophysiology. *Curr. Opin. Microbiol.* **14**, 300–306 (2011).
- Prosser, J. I. & Nicol, G. W. *Bergey's Manual of Systematics of Archaea and Bacteria* (2016).
- Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
- Kerou, M. et al. Proteomics and comparative genomics of *Nitrososphaera viennensis* reveal the core genome and adaptations of archaeal ammonia oxidizers. *Proc. Natl Acad. Sci. USA* **113**, E7937–E7946 (2016).
- Zhang, S. et al. Comparative genomics reveals ecological and evolutionary insights into sponge-associated Thaumarchaeota. *mSystems* **4**, <https://doi.org/10.1128/mSystems.00288-19> (2019).
- Vajrala, N. et al. Hydroxylamine as an intermediate in ammonia oxidation by globally abundant marine archaea. *Proc. Natl Acad. Sci. USA* **110**, 1006–1011 (2013).
- Konneke, M. et al. Ammonia-oxidizing archaea use the most energy-efficient aerobic pathway for CO₂ fixation. *Proc. Natl Acad. Sci. USA* **111**, 8239–8244 (2014).
- Doxey, A. C., Kurtz, D. A., Lynch, M. D., Sauder, L. A. & Neufeld, J. D. Aquatic metagenomes implicate Thaumarchaeota in global cobalamin production. *ISME J.* **9**, 461–471 (2015).
- Spang, A. et al. The genome of the ammonia-oxidizing *Candidatus Nitrososphaera gargensis*: insights into metabolic versatility and environmental adaptations. *Environ. Microbiol.* **14**, 3122–3145 (2012).
- Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490–D495 (2014).
- Epstein, W. The Kdp system: a bacterial K transport ATPase. *Curr. Top. Membr. Transp.* **23**, 153–175 (1985).
- Moreau, P. L. Diversion of the metabolic flux from pyruvate dehydrogenase to pyruvate oxidase decreases oxidative stress during glucose metabolism in nongrowing *Escherichia coli* cells incubated under aerobic, phosphate starvation conditions. *J. Bacteriol.* **186**, 7364–7368 (2004).
- Patridge, E. V. & Ferry, J. G. WrbA from *Escherichia coli* and *Archaeoglobus fulgidus* is an NAD(P)H:quinone oxidoreductase. *J. Bacteriol.* **188**, 3498–3506 (2006).
- Alves, R. J. E. et al. Ammonia oxidation by the arctic terrestrial thaumarchaeote *Candidatus Nitrosocosmicus arcticus* is stimulated by increasing temperatures. *Front. Microbiol.* **10**, 1571 (2019).
- Wahl, R. C., Warner, C. K., Finnerty, V. & Rajagopalan, K. V. *Drosophila melanogaster* ma-1 mutants are defective in the sulfuration of desulfo Mo hydroxylases. *J. Biol. Chem.* **257**, 3958–3962 (1982).
- Cohen, H. J. & Fridovich, I. Hepatic sulfite oxidase. *Purif. Prop. J. Biol. Chem.* **246**, 359–366 (1971).
- Daebeler, A. et al. Cultivation and genomic analysis of “*Candidatus Nitrosocaldus islandicus*,” an obligately thermophilic, ammonia-oxidizing thaumarchaeon from a hot spring biofilm in Graendalur Valley, Iceland. *Front. Microbiol.* **9**, 193 (2018).
- Abby, S. S. et al. *Candidatus Nitrosocaldus cavascurensis*, an ammonia oxidizing, extremely thermophilic archaeon with a highly mobile genome. *Front. Microbiol.* **9**, 28 (2018).
- Gebhard, S., Tran, S. L. & Cook, G. M. The Phn system of *Mycobacterium smegmatis*: a second high-affinity ABC-transporter for phosphate. *Microbiology* **152**, 3453–3465 (2006).
- Torres, A. et al. NADH dehydrogenases in *Pseudomonas aeruginosa* growth and virulence. *Front. Microbiol.* **10**, 75 (2019).
- Kerscher, S., Dröse, S., Zickermann, V. & Brandt, U. *Bioenergetics* 185–222 (Springer, 2007).
- Ito, T. et al. Genetic and biochemical analysis of anaerobic respiration in *Bacteroides fragilis* and its importance in vivo. *mBio* **11**, <https://doi.org/10.1128/mBio.03238-19> (2020).
- Wolf, Y. I. & Koonin, E. V. Genome reduction as the dominant mode of evolution. *Bioessays* **35**, 829–837 (2013).
- Treangen, T. J. & Rocha, E. P. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet.* **7**, e1001284 (2011).
- Riehle, M. M., Bennett, A. F. & Long, A. D. Genetic architecture of thermal adaptation in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **98**, 525–530 (2001).
- Zhang, J., Dean, A. M., Brunet, F. & Long, M. Evolving protein functional diversity in new genes of *Drosophila*. *Proc. Natl Acad. Sci. USA* **101**, 16246–16250 (2004).
- Corbin, C. J. et al. Paralogs of porcine aromatase cytochrome P450: a novel hydroxylase activity is associated with the survival of a duplicated gene. *Endocrinology* **145**, 2157–2164 (2004).
- Weitzel, C. S. et al. Duplication of leucyl-tRNA synthetase in an archaeal extremophile may play a role in adaptation to variable environmental conditions. *J. Biol. Chem.* **295**, 4563–4576 (2020).
- Plaitakis, A., Spanaki, C., Mastorodemos, V. & Zaganas, I. Study of structure–function relationships in human glutamate dehydrogenases reveals novel molecular mechanisms for the regulation of the nerve tissue-specific (GLUD2) isoenzyme. *Neurochem. Int.* **43**, 401–410 (2003).
- Hu, C., Lin, S. Y., Chi, W. T. & Charng, Y. Y. Recent gene duplication and subfunctionalization produced a mitochondrial GrpE, the nucleotide exchange factor of the Hsp70 complex, specialized in thermotolerance to chronic heat stress in *Arabidopsis*. *Plant Physiol.* **158**, 747–758 (2012).

56. Fierer, N. Embracing the unknown: disentangling the complexities of the soil microbiome. *Nat. Rev. Microbiol.* **15**, 579–590 (2017).
57. Cobo-Simón, M. & Tamames, J. Relating genomic characteristics to environmental preferences and ubiquity in different microbial taxa. *BMC Genomics* **18**, 499 (2017).
58. Zhang, J. Evolution by gene duplication: an update. *Trends Ecol. Evolution* **18**, 292–298 (2003).
59. Li, D., Liu, C., Luo, R., Sadakane, K. & Lam, T. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
60. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
61. Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
62. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
63. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
64. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
65. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
66. Li, G., Rabe, K. S., Nielsen, J. & Engqvist, M. K. Machine learning applied to predicting microorganism growth temperatures and enzyme catalytic optima. *ACS Synth. Biol.* **8**, 1411–1420 (2019).
67. Lagesen, K. et al. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108 (2007).
68. Chaumeil, P., Mussig, A. J., Hugenholz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2018).
69. Page, A. J. et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
70. Katoh, K. & Toh, H. Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinforma.* **9**, 286–298 (2008).
71. 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).
72. Bruen, T. & Bruen, T. *PhiPack: PHI Test and Other Tests of Recombination* (McGill University, Montreal, Quebec, 2005).
73. Nguyen, L., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
74. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., von Haeseler, A. & Jermini, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587 (2017).
75. 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).
76. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127–128 (2007).
77. Sauder, L. A., Engel, K., Lo, C. C., Chain, P. & Neufeld, J. D. “*Candidatus Nitrosotenuis aquarius*,” an ammonia-oxidizing archaeon from a freshwater aquarium biofilter. *Appl. Environ. Microbiol.* **84**, <https://doi.org/10.1128/AEM.01430-18> (2018).
78. Sauder, L. A. et al. Cultivation and characterization of *Candidatus Nitrosocosmicus esaqueare*, an ammonia-oxidizing archaeon from a municipal wastewater treatment system. *ISME J.* **11**, 1142–1157 (2017).
79. Santoro, A. E. et al. Genomic and proteomic characterization of “*Candidatus Nitrosopelagicus brevis*”: an ammonia-oxidizing archaeon from the open ocean. *Proc. Natl Acad. Sci. USA* **112**, 1173–1178 (2015).
80. Lehtovirta-Morley, L. E. et al. Isolation of ‘*Candidatus Nitrosocosmicus franklandus*’, a novel ureolytic soil archaeal ammonia oxidiser with tolerance to high ammonia concentration. *FEMS Microbiol. Ecol.* **92**, <https://doi.org/10.1093/femsec/fiw057> (2016).
81. Jung, M., Islam, M. A., Gwak, J., Kim, J. & Rhee, S. *Nitrosarchaeum koreense* gen. nov., sp. nov., an aerobic and mesophilic, ammonia-oxidizing archaeon member of the phylum Thaumarchaeota isolated from agricultural soil. *Int. J. Syst. Evol. Microbiol.* **68**, 3084–3095 (2018).
82. Konstantinidis, K. T. & Tiedje, J. M. Towards a genome-based taxonomy for prokaryotes. *J. Bacteriol.* **187**, 6258–6264 (2005).
83. Rodriguez-R, L. M. & Konstantinidis, K. T. Bypassing cultivation to identify bacterial species. *Microbe* **9**, 111–118 (2014).
84. Kanehisa, M. *The KEGG Database* (Novartis Foundation Symposium, Wiley Online Library, 2002).
85. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.* **428**, 726–731 (2016).
86. 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).
87. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59 (2015).
88. El-Gebali, S. et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
89. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).
90. 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).
91. Sheridan, P. O. et al. Gene duplication drives genome expansion in a major lineage of Thaumarchaeota. *ALE*, <https://doi.org/10.5281/zenodo.4012549> (2020).

Acknowledgements

This work and P.O.S. were financially supported by UKRI through the NERC grant NE/R001529/1. In addition, C.G.-R. and T.A.W. were both supported by a Royal Society University Research Fellowship (URF150571 and URF140626). C.Q. was funded through an MRC fellowship (MR/M50161X/1) as part of the CLOUD Infrastructure for Microbial Genomics (CLIMB) consortium (MR/L015080/1). S.R. was funded through the BBSRC grant BB/R015171/1. The Thames Metagenome Database was funded through the NERC grants NE/M011674/1, NE/M011259/1 and NE/M01133X/1. We thank Dr Tony Travis for his support with Biolinux and acknowledge Prof Jim Prosser for his critical reading of the manuscript. 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. C.Q. and S.R. assembled the 12 metagenomes. The Thames consortium, including J.H. and L.Z., collected samples and performed DNA extraction for the metagenome assembly. 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.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41467-020-19132-x>.

Correspondence and requests for materials should be addressed to C.G.-R.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020

Thames Consortium

William H. Gaze⁷, Jennifer Holden⁶, Andrew Mead⁸, Sebastien Raguideau³, Christopher Quince^{3,4,5}, Andrew C. Singer⁹, Elizabeth M. H. Wellington⁴ & Lihong Zhang⁷

⁸Rothamsted Research, Harpenden, UK. ⁹UK Centre for Ecology & Hydrology, Wallingford OX10 8BB, UK.