Soils and sediments host novel archaea with divergent monooxygenases implicated in ammonia oxidation

Spencer Diamond[1,2*], Adi Lavy[2], Alexander Crits-Christoph[3], Paula B. Matheus Carnevali[2], Allison Sharrar[2], Kenneth H. Williams[3], Jillian F. Banfield[1,2,5,6*]

[1]Innovative Genomics Institute, University of California, Berkeley, CA, USA.
[2]Department of Earth and Planetary Science, University of California, Berkeley, CA, USA.
[3]Department of Plant and Microbial Biology, University of California, Berkeley, CA, USA.
[4]Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[5]Environmental Science, Policy and Management, University of California, Berkeley, CA, USA.
[6]School of Earth Sciences, University of Melbourne, Melbourne, Victoria, Australia.

***Corresponding authors:**
Spencer Diamond
Associate Project Scientist
Earth and Planetary Science Department
University of California - Berkeley
Innovative Genomics Institute Building
2151 Berkeley Way, Berkeley, CA 94720
Phone: (818) 268-5776
Email: sdiamond@berkeley.edu

Jillian F. Banfield
Professor
Earth and Planetary Science Department
University of California - Berkeley
Innovative Genomics Institute Building
2151 Berkeley Way, Berkeley, CA 94720
Phone: (510) 316-4334
Email: jbanfield@berkeley.edu

ABSTRACT

Copper membrane monooxygenases (CuMMOs) play critical roles in the global carbon and nitrogen cycles. Organisms harboring these enzymes perform the first, and rate limiting, step in aerobic oxidation of ammonia, methane, or other simple hydrocarbons. Within archaea, only organisms in the order Nitrososphaerales (Thaumarchaeota) encode CuMMOs, which function exclusively as ammonia monooxygenases. From grassland and hillslope soils and aquifer sediments, we identified 20 genomes from distinct archaeal species encoding divergent CuMMO sequences. These archaea are phylogenetically clustered in a previously unnamed Thermoplasmatota order, herein named the Ca. Angelarcheales. The CuMMO proteins in Ca. Angelarcheales are more similar in structure to those in ammonia-oxidizing archaea than those of bacteria, and they contain all functional residues required for activity. Similarly to the Nitrososphaerales, Ca. Angelarcheales genomes are significantly enriched in blue copper proteins (BCPs) relative to sibling lineages, including plastocyanin-like electron carriers and divergent nitrite reductase-like (nirK) 2-domain cupredoxin proteins co-located with electron transport machinery. Angelarcheales do not have identifiable genes for methanol oxidation or carbon fixation, encode significant capacity for peptide/amino acid uptake and degradation, and share numerous electron transport mechanisms with the Nitrososphaerales. In the studied soils and sediments Ca. Angelarcheales were at least as abundant as ammonia-oxidizing Nitrososphaerales. Thus, we predict that Angelarcheales live a mixotrophic lifestyle based on oxidation of ammonia liberated from peptide and amino acid degradation. This work expands the known diversity of Thermoplasmatota and of CuMMO enzymes in archaea and suggests that these organisms are important and previously unaccounted for contributors to nitrogen cycling.

INTRODUCTION

Copper membrane monooxygenases (CuMMOs) are a family of phylogenetically diverse and ecologically widespread but highly conserved enzymes that function in the aerobic oxidation of ammonia, methane, and potentially other hydrocarbon substrates[1–3]. Organisms encoding CuMMO are critically important in global carbon and nitrogen cycles[4,5], as particulate methane monooxygenases (pMMOs) attenuate atmospheric methane release[5], and ammonia monooxygenases (AMOs) can generate nitrous oxide as a byproduct[6]. While the diversity of CuMMOs has been expanded due to cultivation independent studies[7–9], CuMMOs have still only been identified in a few monophyletic clades of Bacteria and Archaea[1]. The distribution and functionality of CuMMOs in Archaea is particularly constrained. Archaeal CuMMOs have only been shown to participate in ammonia oxidation, and Nitrososphaerales (Thaumarchaeota) are the only identified group to date that encode AMOs[8]. These ammonia oxidizing archaea (AOA) occur across a wide array of natural environments[10], and are often far more abundant than ammonia oxidizing bacteria (AOB)[11], resulting in large contributions to the global nitrogen cycle[3,12].

The CuMMO is a membrane bound protein complex that consists of 3 core proteins amoA/pmoA, amoB/pmoB, and amoC/pmoC, that are typically encoded in an operon[1,13]. In AOA, a fourth hypothetical protein, amoX, of unknown function is also present[13]. Despite their ecological importance, the difficulties in sustaining in vitro activity of CuMMO complexes as well as difficulties in protein structure resolution has led to inconsistent reports regarding the co-factors and substrate specificity controls for these enzymes. Indeed the substrate specificity of CuMMOs is promiscuous and purified enzymes have been demonstrated to have activity on methane, ammonia, and small hydrocarbons[14,15]. However, in the organismal context, the substrate

51 specificity is highly preferential to a single substrate, and occurs in a metabolic context of enzymes
52 that support the processing of that specific substrate and its byproducts[3,16].
53
54 In both AOA and AOB, ammonia oxidation proceeds initially by the conversion of ammonia to
55 hydroxylamine and subsequent oxidation of hydroxylamine to nitrite[17]. While the AMO mediated
56 oxidation of ammonia to hydroxylamine has been established in AOA[18], the lack of an identifiable
57 hydroxylamine dehydrogenase (Hao) in Archaea has led to the proposal of at least two
58 hypothetical routes from hydroxylamine to nitrite[17,19]. Evidence is mounting that Hao activity in
59 AOA instead relies on a suite of blue copper proteins (BCPs), which are numerous in AOA
60 lineages[16,17,19,20]. Plastocyanin-like electron transport proteins, divergent Cu-containing nitrite
61 reductases (nirK), and other 1 and 2 domain cupredoxin-like proteins have been shown to have
62 correlated transcriptional activity with AMO proteins[21], and are implicated as substituting for the
63 missing Hao activity.
64
65 Identification of novel CuMMO complexes in microbes, and surveys of CuMMOs in natural
66 environments, has primarily relied on primer based nucleic acid amplification of the amoA/pmoA
67 subunit[22]. However, the requirement of sufficient sequence similarity between environmental
68 sequences and known bacterial and archaeal amoA/pmoA subunits precludes the identification
69 of completely novel sequence variants. Further, primer-based studies do not provide information
70 about genome context, which can place CuMMOs in a metabolic context that supports a specific
71 activity. Here, we applied sensitive homology-based methods to *de novo* reconstructed
72 metagenome assembled genomes (MAGs) obtained from soils, aquifer sediment, and deep
73 ocean environments to identify novel CuMMO variants and describe their genomic, phylogenetic
74 and ecological contexts. We propose that a new group of soil- and aquifer-associated archaea

75 contribute to the ecosystem essential and possibly keystone functionality of ammonia oxidation
76 that has previously only been linked to a phylogenetically narrow range of bacteria and archaea.
77
78 **RESULTS**
79
80 **Divergent CuMMOs identified in MAGs recovered from soil and sediment ecosystems.** In
81 previous work we identified putative divergent amoA/pmoA homologues in 7 Thermoplasmatota
82 genomes recovered from a Mediterranean grassland soil[23]. This was intriguing, given that
83 amo/pmo homologues had not been previously observed in archaea outside of the
84 Nitrososphaerales (Thaumarchaeota). Here we searched for additional genomes encoding
85 related (divergent) amo/pmos using a series of readily available, and custom built, hidden markov
86 models (HMMs) across all archaeal genomes in the Genome Taxonomy Database (GTDB), and
87 in all archaeal MAGs in our unpublished datasets from ongoing studies (**Supplementary Fig. 1**).
88 We found additional amoA/pmoA genes in genomes recovered from soils at the South Meadow
89 and Rivendell sites of the Angelo Coast Range Reserve (CA)[23,24], the nearby Sagehorn site[24], a
90 hillslope of the East River watershed (CO)[25], and in sediments from the Rifle aquifer (CO)[26] and
91 the deep ocean[27]. In total we identified 201 archaeal MAGs taxonomically placed using
92 phylogenetically informative single copy marker genes outside of Nitrososphaerales containing
93 divergent amo/pmo proteins (**Supplementary Table 1 and Supplementary Data 1)**. Genome
94 de-replication with additional references at 95% average nucleotide identity (ANI) resulted in 34
95 species-level genome clusters, 20 of which encoded an amo/pmo homologue (**Supplementary**
96 **Table 2**). Of these genomes, 11 are species not previously available in public databases. In all
97 cases where assembled sequences were of sufficient length, the amoA/pmoA, B, and C protein
98 coding genes were found co-located with each other and with a hypothetical protein considered
99 to be amoX in the order C-A-X-B (**Fig. 1a**, **Supplementary Table 2, and Supplementary Fig.**

100 **2**). The mean sequence identity of the recovered amoA/pmoA, B, and C proteins to known
101 bacterial sequences were 16.7, 8.0, and 14.2 % and 13.8, 9.5, and 20.8 % to known archaeal
102 sequences. As might be expected considering the large sequence divergence between the
103 recovered sequences and known amo/pmo proteins, we found that no pair of typical primers used
104 for bacterial and archaeal amoA/pmoA environmental surveys[28] matched any novel amoA/pmoA
105 gene with less than 7 mismatched bases (**Supplementary Table 3**). This suggests that these
106 sequences would have been missed by previous primer-based pmoA gene surveys.
107
108 **Novel CuMMO subunit sequences contain expected catalytic and metal binding residues.**
109 Alignments were constructed for each predicted amoA/pmoA, B, and C protein in combination
110 with reference sequences that cover the known diversity of these protein subunits[1] (**Fig. 1b and**
111 **Supplementary Data 1**). In the new sequences, all of the expected catalytic and metal binding
112 residues[29] were conserved (**Fig. 1b**). In amoB/pmoB, all three histidine residues for the mono-
113 copper B-site required for enzyme activity were conserved. The C-site in amoC/pmoC, which
114 contains an aspartic acid and two histidine residues important for enzymatic activity[13], is also
115 completely conserved. Although the A-sites in amoB/pmoB and the D-sites in amoC/pmoC and
116 amoA/pmoA were not observed in the new sequences, these sites are not required for catalytic
117 function and are only conserved within bacterial lineages[1,8]. We also note that the new
118 amoB/pmoB sequences share a C-terminal truncation to previously identified Nitrososphaerales
119 pmoB's[30,31], as well as share an N-terminal truncation in the amoC/pmoC protein.
120
121 **Novel CuMMOs form a new group in the CuMMO superfamily.** To infer the evolutionary
122 relationship of our newly identified CuMMO sequences to known CuMMO family members we
123 used a concatenated pmoABC sequence alignment to produce a phylogenetic reconstruction
124 covering known family members **(Fig. 1c)**. Individual protein phylogenies were also constructed

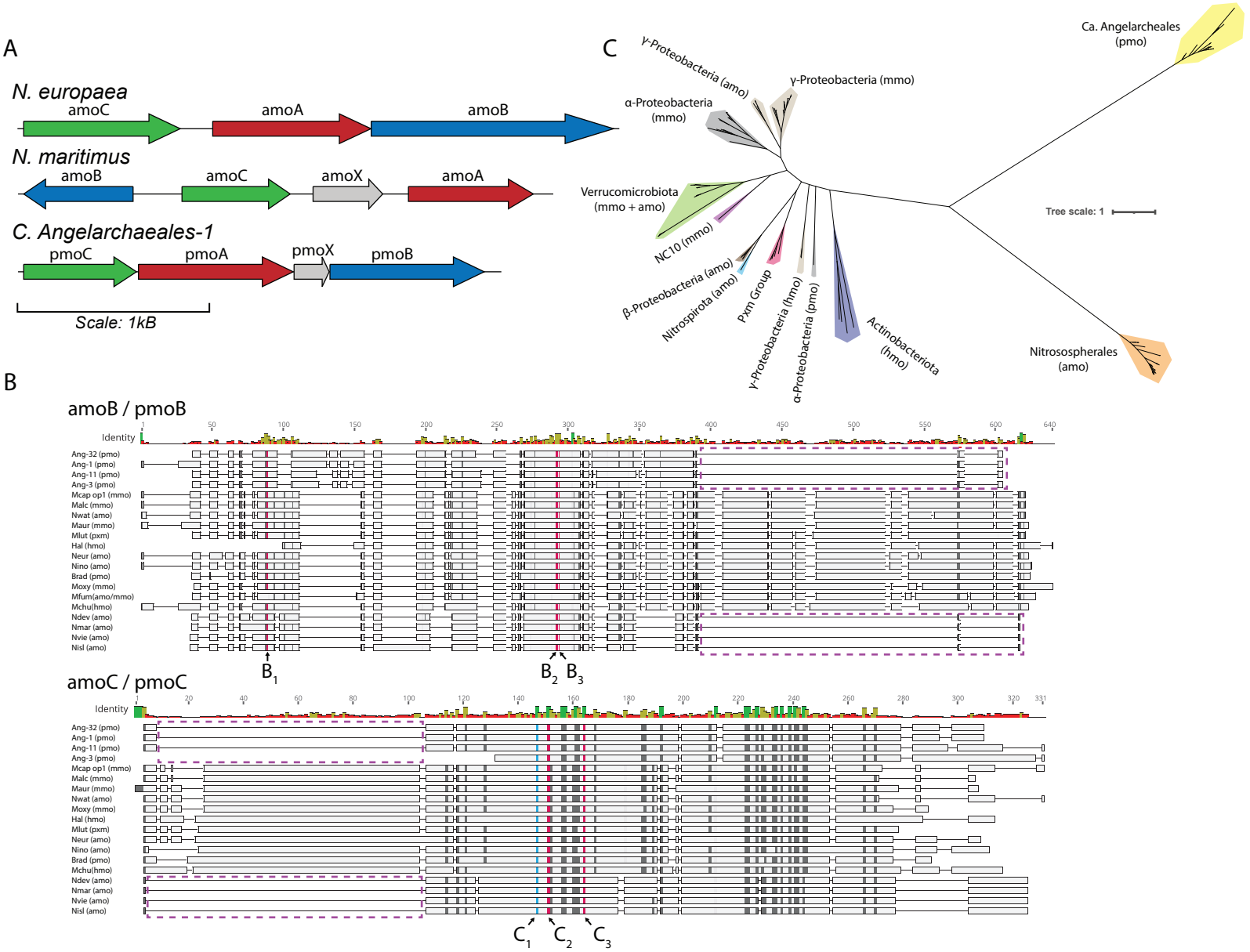

Figure 1 I Comparison of divergent CuMMO proteins to known sequences. (A) Comparison of pmoABC loci and gene arrangement between Nitrosomonas europaea (ammonia oxidizing bacteria), Nitrosopumilus maritimus (ammonia oxidizing archaea), and Ca. Angelarcheales-1. **(B)** Representative sequences from a multiple sequence alignment of amoB/pmoB proteins (top; n = 114 sequences) and amoC/pmoC proteins (bottom; n = 110 sequences). Sequences are named by species abbreviation followed by putative function in parentheses: pmo = particulate monooxygenase (no biochemical evidence), mmo = methane monooxygenase, amo = ammonia monooxygenase, hmo = hydrocarbon monooxygenase. Species abbreviations are described in Supplementary Table 3. Lettering below alignments indicates residues forming part of B or C-sites in these proteins. Histidine residues and aspartic acid residues within these sites are colored red and blue respectively. Residues colored black are identical across >= 65% of sequences in full alignment. Purple boxes indicate truncations shared by CuMMOs identified in this study and known archaeal sequences. All amoABC/pmoABC proteins and alignments are available in Supplementary Data 1. **(C)** Maximum likelihood phylogenetic tree constructed from a concatenated alignment of amoABC/pmoABC proteins (n = 112 sequences; >= 2 sequences per genome). Colored clades are drawn based on a combination of shared function and taxonomy and colored by the taxonomy of genomes encoding the sequences within the clade. Clades are labeled by their shared taxonomy and known/predicted CuMMO function in parentheses (see above). Pxm group is an exception and represents a group of duplicated proteins present in many gammaproteobacterial methanotrophs. Scale bar represents average changes per amino acid position.

125 for each pmo subunit (**Supplementary Fig. 3a-c**). The different subunit reconstructions largely
126 agree in overall topology, and all reconstructions support our newly identified sequences as a
127 highly divergent third major lineage of CuMMOs. Also, similar to previous phylogenetic
128 reconstructions of CuMMO sequences, our sequences form clusters that mirror the phylogenetic
129 relatedness of encoding genomes[26].
130
131 **Archaea with divergent CuMMOs form a novel clade within the Thermoplasmatota.** An initial
132 phylogenetic classification of our archaeal genomes placed them as an order level lineage, RBG-
133 16-18-12, within the Thermoplasmatota phylum (**Supplementary Tables 1-3**). A concatenated
134 alignment of 76 archaeal specific marker proteins[32] resolved them as a well supported order-level
135 monophyletic lineage basal to the *Methanomassilococcales* within *Thermoplasmatota* (**Fig. 2a,**
136 **Supplementary Fig. 4, and Supplementary Table 4**). We reported the first genomes from this
137 clade (RBG-16-68-12 in GTDB) from the RBG dataset from the Rifle, CO aquifer sediments[23].
138 Given that this clade is now represented by 32 species-level genomes (with 11 additional species
139 added in this study), and that 10 genomes satisfy the completeness and contamination
140 requirements to be considered high-quality drafts[1,8], we propose that they define a new candidate
141 order, hereafter referred to as the Ca. Angelarcheales. We propose Ca. Angelarcheales-1
142 (GCA_005878985.1[16]) from the Angelo Coastal Reserve, CA, as the type genome for this clade.
143 Of the Ca. Angelarcheales genome set, 20 contain identifiable pmo/amo gene clusters. The fact
144 that CuMMO sequences are encoded within genomes of a single monophyletic subclade of the
145 Thermoplasmatota is similar to previous observed patterns of CuMMO distribution, as taxa that
146 encode CuMMO systems are often constrained to monophyletic groups scattered across the tree
147 of life[3,17,19].
148

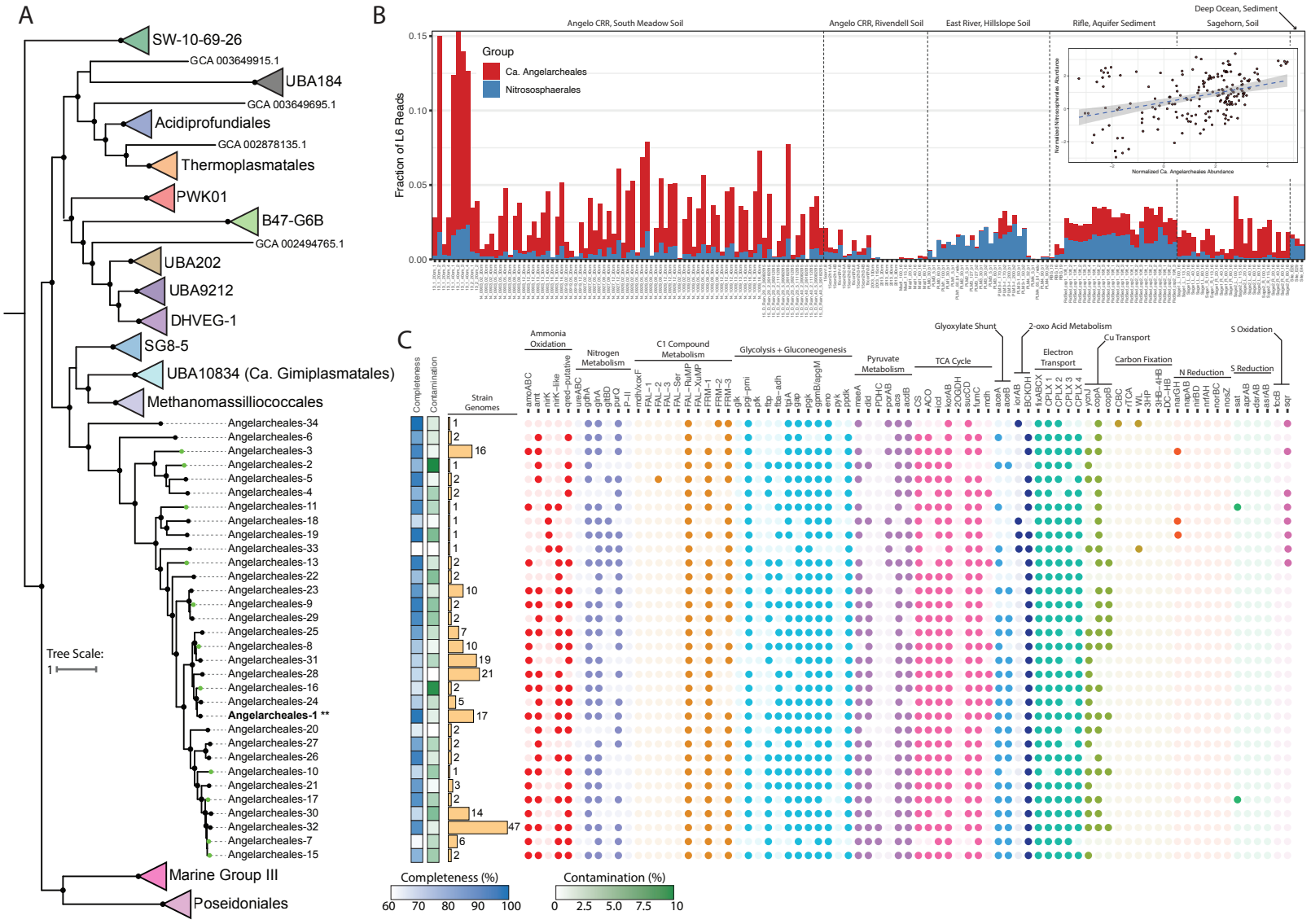

Figure 2 I Taxonomy, metabolism, and distribution of Ca. Angelarcheales. (A) Maximum likelihood phylogenetic tree constructed for the archaeal phylum Thermoplasmatota using a concatenated alignment of 76 archaeal specific marker genes. The tree includes 32 Ca. Angelarcheales genomes and 295 reference genomes. The type strain, Ca. Angelarcheales-1, is marked with two asterisks. Tree was rooted using A. fulgidus (GCF_000008665.1) as an outgroup. Clades were collapsed at the order level if they contained more than one genome and arbitrarily colored. Black dots at nodes indicate >= 95% bootstrap support (ufboot; n = 1000). For the full, un-collapsed, tree see Supplementary Fig. 4. **(B)** Relative abundance information, based on rpL6 read counts, for Ca. Angelarcheales and Nitrososphaerales across 185 shotgun metagenome samples from 6 sites. The x-axis indicates the sample name, and the y-axis indicates the fraction of reads out of the total reads in a given sample that mapped to rpL6 sequences taxonomically associated with each group. Samples are separated by the general sampling location, indicated at the top of the plot. Inset, normalized rpL6 based relative abundance of Ca. Angelarcheales (x-axis) vs Nitrososphaerales (y-axis) for all 185 shotgun metagenome samples. A best fit line is plotted using linear regression, shaded area indicates standard error of the regression. Rho of association is positive and significant (rho = 0.366, FDR < 0.001). **(C)** Genome quality, number of strain-level genomes (genomes with ANI >= 95%), and predicted metabolism for the 32 Ca. Angelarcheales genomes in Fig. 2A. Filled dots indicate the presence of a gene or gene set that executes a specific metabolic function or reaction. Dots are colored based on shared pathways or metabolic functionality as described above the figure, and colors are chosen arbitrarily. For complete explanation of metabolic functions and search criteria see Supplementary Table 8.

149 **Taxonomic marker profiling indicates widespread distribution of Ca. Angelarcheales.**
150 Using ribosomal protein L6 (rpL6) as a taxonomic marker, we determined an average prokaryotic
151 community relative abundance of 1.73 ± 2.25 % for Ca. Angelarcheales in 185 samples from six
152 environments **(Fig. 2b and Supplementary Tables 5-6).** In comparison, the average relative
153 abundance of Nitrososphaerales in the same dataset was 0.65 ± 0.61 %. We observed significant
154 and strong positive associations between the abundances of both Ca. Angelarcheales and
155 Nitrososphaerales with each other (**Fig. 2b;** rho = 0.366, FDR < 0.001), and with the bacterial
156 order Nitrospirales (rho_{Ang} = 0.400, FDR < 0.001; $rho_{Nitroso}$ = 0.613; FDR < 0.001), a group of nitrite
157 reducing bacteria (**Supplementary Fig. 5 and Supplementary Table 7**).
158
159 **Ca. Angelarcheales share similar metabolic features with ammonia oxidizing**
160 **Nitrososphaerales.** We conducted a general metabolic analysis of Ca. Angelarcheales to place
161 the CuMMO into metabolic context (**Fig. 2c, Supplementary Fig. 6a-d, Supplementary Table**
162 **8, and Supplementary Data 2**). Ca. Angelarcheales genomes contained many of the electron
163 transfer and ammonia assimilation components known to be conserved in characterized ammonia
164 oxidizing archaea including: an NADH:ubiquinone oxidoreductase complex with an additional
165 copy of the M protein component (CPLX 1), a four-subunit putative succinate dehydrogenase
166 complex (CPLX II), a complete cytochrome b containing complex III with a plastoquinone-like
167 electron transfer apparatus (CPLX 3), up to two distinct oxygen reducing terminal oxidases (CPLX
168 4), an ammonia transporter (amt), a glutamine synthase (glnA), and glutamate dehydrogenase
169 (gdhA). We found little evidence that Ca. Angelarcheales can use inorganic nitrogen or sulfur
170 containing compounds as alternative electron acceptors, thus it is likely that these organisms are
171 obligate aerobes. We also did not identify carbon fixation pathways within any CuMMO encoding
172 Ca. Angelarcheales.
173

174 Some genomes encode a credible nitrite reductase (nirK) and 2-domain cupredoxins with
175 homology to nirK (nirK-like). However, nirK is not essential for ammonia oxidation in
176 Nitrososphaerales[33]. Also identified were plastocyanin-like proteins, which are common in
177 Nitrososphaerales (**Fig. 3),** and three distinct copper transport systems (**Fig. 2c**). The parallels
178 between the redox enzyme inventories of Ca. Angelarcheales and Nitrososphaerales support the
179 role of the CuMMO in ammonia rather than methane oxidation. Further supporting this deduction,
180 we did not detect any genes with homology to methanol dehydrogenases (mdh/xoxF) across the
181 entire Ca. Angelarcheales clade, which would be expected if methane to methanol oxidation was
182 occurring. Mechanisms to assimilate formaldehyde and formate (breakdown products of
183 methanol) were present, however these compounds can originate from other processes, such as
184 sarcosine and betaine degradation pathways, which are present in Ca. Angelarcheales.
185
186 Ca. Angelarcheales contain numerous transporters for branched chain amino acids, polar and
187 non-polar amino acids, oligopeptides, and many proteases (**Supplementary Fig. 6b-d**). The
188 number of encoded amino acid and peptide transport systems in Ca. Angelarcheales is on
189 average the largest across the phylum Thermoplasmatota (**Supplementary Fig. 6d**). The
190 presence of a branched chain keto acid dehydrogenase complex (BCKDH) enables the
191 degradation of branched chain amino acids to acetyl and propionyl-CoA, and a glyoxylate shunt
192 (aceA and aceB) enables the carbons of acetyl-CoA to be used for biosynthesis. A number of
193 enzymes indicate the capacity for acetate degradation to acetyl-CoA (acetate-CoA ligase (acdB))
194 and lactate degradation to pyruvate (D-lactate dehydrogenase (dld)). These archaea do not have
195 a complete glycolytic pathway (missing core enzymes including glucokinase (glk),
196 phosphofructokinase (pfk), and pyruvate kinase (pfk)), but have gluconeogenesis pathways, thus
197 enabling the biosynthesis of glucose from acetyl-CoA and pyruvate.
198

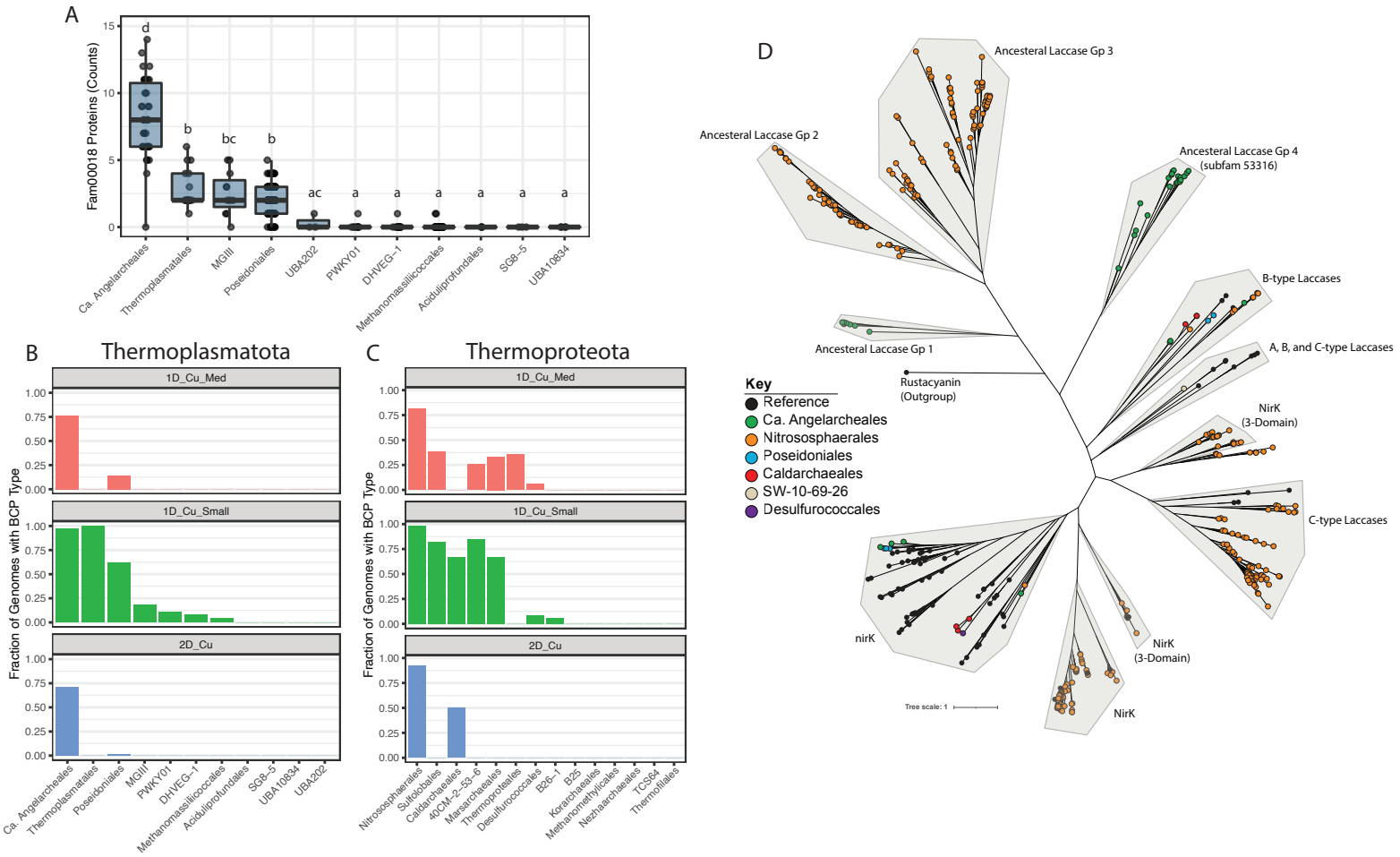

Figure 3 | Analysis of BCPs across CuMMO encoding archaeal phyla. (A) Counts of fam00018 proteins in genomes from each order level lineage of the Thermoplasmatota containing ≥ 3 genomes. Each dot represents one genome. Boxes indicate the first and third quartile of counts, lines in boxes indicate median values, and whiskers indicate 1.5 x IQR in either direction. Letters above boxes indicate statistically significant differences between groups. Groups sharing no letters have statistically significant differences (FDR ≤ 0.05; pairwise Wilcoxon test). **(B)** The fraction of genomes within each order level lineage of the Thermoplasmatota containing ≥ 3 genomes carrying the BCP subtype noted in the plot title. **(C)** The fraction of genomes within each order level lineage of the Thermoproteota containing ≥ 3 genomes carrying the BCP subtype noted in the plot title. **(D)** Maximum likelihood phylogenetic tree containing 349 2-domain and 3-domain BCPs from our analysis and 90 reference nirK and 2-domain laccase proteins. Clades were manually defined, shaded in grey, and named based on their constituent reference sequences or based on their sequence architecture relative to expected ancestral 2-domain BCPs34. Node colors indicate if a sequence was a reference sequence (black) or the order level taxonomy of the encoding genome. Rustacyanin (ACK80662.1) is provided as an outgroup.

199 **Blue copper proteins (BCPs) are enriched in CuMMO containing archaea.** As BCPs have
200 been suggested to serve critical functions in archaeal ammonia oxidation[16], we compared the
201 BCP inventories in genomes of the phyla Thermoplasmatota and Thermoproteota, which include
202 Ca. Angelarcheales and Nitrososphaerales, respectively. The dataset included 34 representative
203 Ca. Angelarcheales genomes and 610 reference genomes (**Supplementary Table 4**). Due to
204 their high primary sequence diversity, the identification and comparison of BCPs across
205 organisms is difficult using standard annotation methods. Thus we clustered 1,103,913 proteins
206 using a previously validated two-step protein clustering approach[33]. This generated 76,216 protein
207 subfamily clusters (subfams), which are groups of proteins sharing global homology, and 19,828
208 protein family clusters (fams), which are groups of protein subfamilies where remote local
209 homology could be confidently detected. We identified 1,927 proteins with BCP-associated
210 (cupredoxin-like) PFAM domains across 30 protein fams (**Supplementary Fig. 7a**). Notably, a
211 single protein family (fam00018) contained 1,738 (90.2%) of these proteins, and the remaining
212 proteins either made up very small fractions of other fams or were part of fams with very few
213 proteins (**Supplementary Fig. 7b**). Analysis of the domain architectures of proteins within
214 fam00018 indicate that this protein family primarily contains BCPs with between 1-3 cupredoxin-
215 like domains. Included in fam00018 are small globular plastocyanin-like proteins, nirK-like
216 proteins, 2-domain laccase-like proteins, and the Cu binding cytochrome c oxidase subunit 2
217 (coxB/COX2) (**Supplementary Fig. 7c**). Fam00018 also contained 671 proteins with no
218 identifiable domain annotations, which was expected given the high sequence diversity of BCPs.
219 However, many proteins with no annotations were clustered into fam00018 subfamilies containing
220 proteins with identifiable BCP domains, allowing the recruitment of these proteins into our
221 analyses. We used the proteins of fam00018 as a broad homology group to quantify and ultimately
222 sub-classify BCP types across genomes (**Supplementary Table 9**). Compared to all other order-
223 level lineages in the Thermoplasmatota, Ca. Angelarcheales genomes are significantly enriched

224 in fam00018 proteins (FDR < 0.05 ; pairwise Wilcoxon test), encoding on average 8.1 per genome
225 **(Fig. 3a)**. This pattern of fam00018 protein enrichment is similarly observed for the ammonia-
226 oxidizing Nitrososphaerales order relative to sibling orders within the Thermoproteota (FDR < 0.05
227 ; pairwise Wilcoxon test) (**Supplementary Fig. 7d)**.
228
229 **Subclassification of fam00018 identifies specific BCP architectures associated with**
230 **lineages carrying CuMMOs.** To more comprehensively understand the subtypes of BCPs that
231 are present across the archaeal orders within Thermoplasmatota and Thermoproteota, we
232 subdivided fam00018 into six manually annotated groups that covered 85.3% of all fam00018
233 proteins (**Fig. 3b, 3c, Supplementary Figs. 8a-b, and Supplementary Table 9**). We observed
234 that small plastocyanin-like 1-domain BCPs (< 250 aa), while present in many lineages, were
235 extremely prevalent in the genomes of Ca. Angelarcheales and Nitrososphaerales, supporting
236 their important role in facilitating electron transport in these groups (**Figs. 3b and 3c**).
237 Alternatively, medium length 1-domain BCPs (250-400 aa) and 2-domain BCPs were encoded by
238 most genomes of Nitrososphaerales and Ca. Angelarcheales, found in few lineages outside them,
239 and if found were not widely present in the genomes of those other lineages (**Figs. 3b and 3c**).
240 This is consistent with these proteins performing functions that are specific to both Ca.
241 Angelarcheales and Nitrososphaerales.
242
243 We speculated that two-domain cupredoxins (2-domain BCPs) generally, and not nirK specifically,
244 may be important in the ammonia oxidation pathway in archaea. These proteins can be
245 differentiated based on phylogenetic relationships, the types of copper centers they contain, and
246 the arrangement of these centers[34,35]. A phylogenetic tree for 2-domain BCPs, known nirK
247 sequences (which include 3-domain BCPs), and 2-domain laccase sequences **(Fig. 3d)** resolves
248 11 discrete clades. The Nitrososphaerales nirK clades are distinct from classic nirK sequences,

249 as has been observed previously[36]. The four, high confidence, nirK sequences identified in Ca.
250 Angelarcheales fall into the classic nirK clade. Four clades are composed of sequences that
251 contain two Type I copper centers but appear to lack Type II or III centers. Such proteins lack
252 functional predictions, and are referred to as ancestral forms of 2-domain BCPs[34]. The 2-domain
253 BCPs of ancestral group 4 (subfam53316) and ancestral group 1 (subfam54500) are exclusively
254 found in Ca. Angelarcheales. We also note that while 2-domain BCPs were found in archaeal
255 orders outside of Ca. Angelarcheales and Nitrososphaerales, these sequences fall into clades of
256 known laccases.
257
258 **BCPs in Ca. Angelarcheales are co-localized with energy generation machinery.** We
259 examined the genomic context, in Angelarcheales, of three gene clusters known to be important
260 for electron transfer and energy generation in ammonia oxidizing archaea: the amoCAXB cluster,
261 the coxAB oxygen utilizing terminal oxidase cluster, and the complex III like cytochrome b gene
262 cluster (**Fig. 4a and Supplementary Figs. 2, 9, and 10)**. Four of the amoCAXB encoding contigs
263 from 20 genomes encode a medium length 1-domain BCP from subfam17112 ~4 genes upstream
264 of the amoCAXB gene cluster. We note that only 5 of 20 contigs have sufficient length upstream
265 of the amoCAXB locus to allow identification of this BCP. The proteins of subfam17112 are all
266 predicted to contain 5 transmembrane helices in their N-terminal region with a cupredoxin-like
267 domain occupying the outer membrane facing C-terminal region (**Supplementary Fig. 11**). The
268 8 proteins of subfam17112 only occur in Ca. Angelarcheales genomes that also encode an amo.
269 Medium length 1-domain BCPs also occur at a high frequency in Nitrososphaerales (**Fig. 3c**) and
270 have been proposed to be associated with hao activity[16]. Thus, subfam17112 may be implicated
271 in hao activity, given its proximity to the amoCAXB locus and the proposal that an archaeal hao
272 would be a divergent membrane bound cupredoxin-like protein[16,37].
273

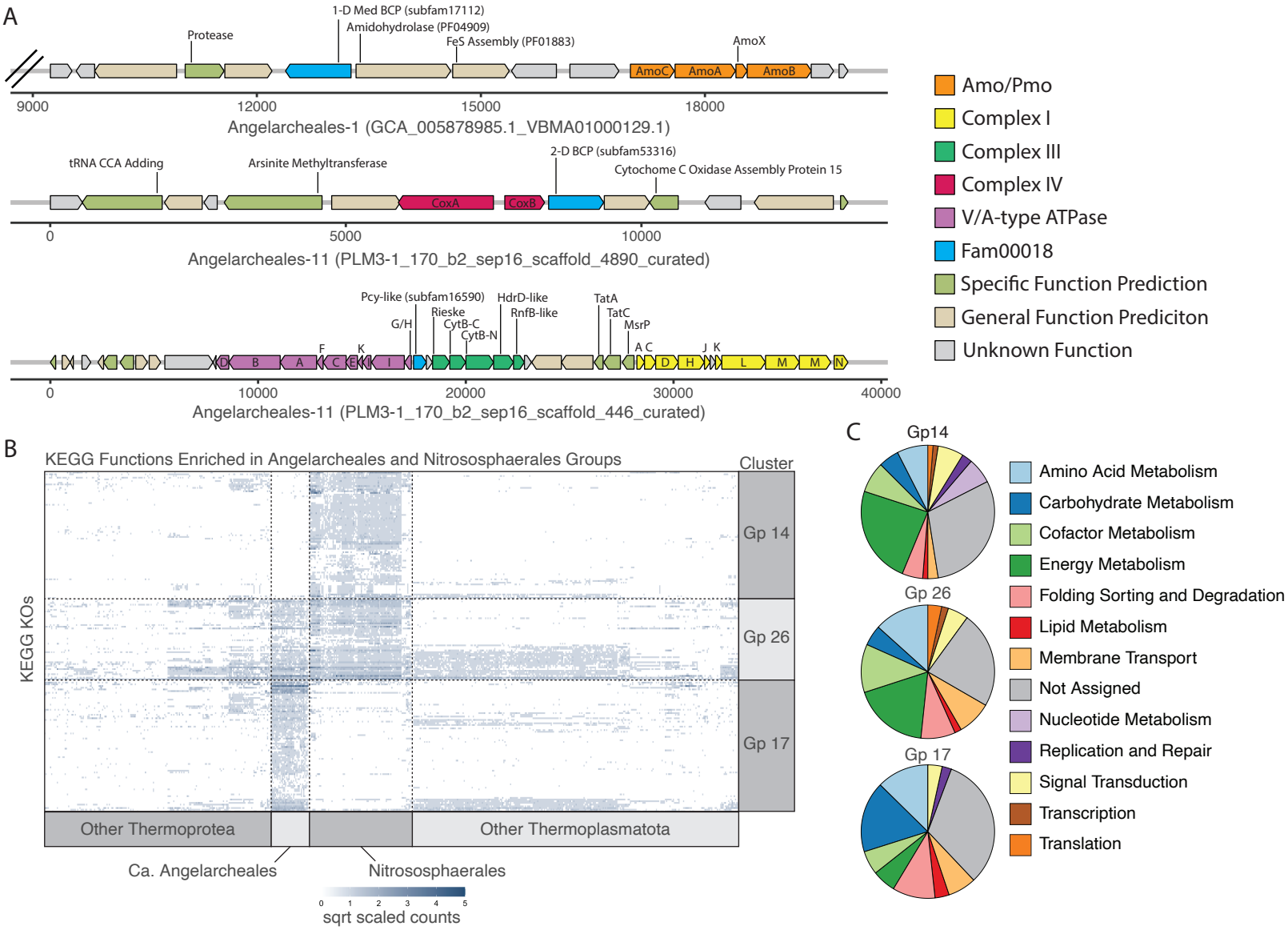

Figure 4 | Genomic context and metabolic function enrichment analysis for Angelarcheales. (A) Example operons showing the gene context surrounding the amoCAXB gene cluster (top), the coxAB oxygen utilizing terminal oxidase gene cluster (middle), and the complex III like cytochrome b gene cluster (bottom). Subfam membership is noted in labels for fam00018 proteins (blue). Scale is in base pairs and double hash on amoCAXB containing contig indicates truncation after 10 genes to the left of the cluster for readability. **(B)** Heatmap showing the number of hits to 201 KEGG KOs found to be significantly enriched in Angelarcheales (Gp 17), the Nitrososphaerales (Gp 14), or showing shared enrichment by both orders relative to all others (Gp 26). Each column represents the hits across one genome, and each row represents the hits for a single KO. Intensity of each spot in the heatmap is based on square root scaled counts of hits to a KO in each genome for ease of readability. Dotted lines are added to segregate clusters for ease of viewing. **(C)** Breakdowns of functional categories associated with KEGG KOs in each enrichment group. Also see Supplementary Table 10.

274 We reconstructed coxAB encoding contigs from 29 genomes. In 19 of these genomes, we could
275 identify a 2-domain BCP directly following the coxB gene, which in 14 of 19 cases was the 2-
276 domain BCP from the nirK-like group subfam53316 (**Fig. 3d)**. Again, in at least 7 cases it was not
277 possible to search for a BCP as contigs were of insufficient length.
278
279 A cytochrome b containing complex III-like locus could be identified and reconstructed in 30
280 genomes. It was commonly co-located with gene clusters encoding other components of the
281 electron transport chain (the V/A-type ATPase and NADPH:Quinone oxidoreductase - complex
282 I). Electron transfer from complex III to downstream electron transport machinery is posited to
283 involve a plastocyanin-like 1 domain BCP, not a soluble cytochrome c, similar to ammonia
284 oxidizing Nitrososphaerales[16,37]. In Ca. Angelarcheales, we found a small 1 domain plastocyanin-
285 like BCP gene upstream of a Rieske iron-sulfur protein in 23 of 30 reconstructed complex III loci
286 (and no cytochrome c). This indicates that the electron transport chain in Ca. Angelarcheales is
287 similar to that found in Nitrososphaerales.
288
289 **Metabolic functions enriched in Ca. Angelarcheales and Nitrososphaerales.** Using indicator
290 analysis, we identified KEGG orthology groups (KOs) that were significantly enriched in the Ca.
291 Angelarcheales (Gp 17), the Nitrososphaerales (Gp 14), and the KOs shared by both orders
292 relative to all other orders of Thermoplasmatota and Thermoproteota (Gp 26). Of the 78 KOs that
293 were significantly enriched in Ca. Angelarcheales (**Fig. 4b and Supplementary Table 10**), the
294 largest functional groups corresponded to carbohydrate metabolism (17.2%), amino acid
295 metabolism (12.6%), and protein folding, sorting, and degradation (10.3%) (**Fig. 4c**). KOs
296 enriched in Ca. Angelarcheales support the use of peptides and amino acids as a carbon and
297 nitrogen source. These included isocitrate lyase of the glyoxylate shunt (K01637), proteins for
298 detoxification of the threonine catabolite methylglyoxal (K10759, K18930, and K23257), 4

299 proteases (K01392, K06013, K07263, and K09640), components of the archaeal proteosome
300 (K13527 and K13571) enzymes for betaine (K00130, K00544, and K00479), proline (K00318),
301 and cysteine (K01760) catabolism, the E1 component of the branched chain keto acid
302 dehydrogenase complex (K00166), and a transport system for polar amino acids (K02028 and
303 K02029).
304
305 The 75 KOs significantly enriched in Nitrososphaerales genomes were largely associated with
306 energy metabolism (23.7%) (**Fig. 4c**). This included functions critical for the
307 hydroxypropionate/hydroxybutyrate carbon fixation pathway known to operate in these organisms
308 (K18593, K18594, K18603, and K18604). The absence of these genes in the Ca. Angelarcheales,
309 thus their inability to fix CO_2, differentiates them from the Nitrososphaerales. Other functions
310 enriched in Nitrososphaerales are involved in electron transfer including plastocyanins (K02638),
311 ferredoxins (K05524), and rieske iron-sulfur proteins (K15878). We also identified enriched
312 capacity for urea utilization (K01429, K01430, K03187, K03188, K03190) and urea transport
313 (K20989), which agrees with the fact that many Nitrososphaerales are thought to use urea as a
314 nitrogen source for ammonia oxidation[38].
315
316 The AMO subunits A and C (K10944 and K10946) were identified among the 48 functions that
317 were significantly enriched in both Ca. Angelarcheales and Nitrososphaerales compared to the
318 other groups. Many shared functions were also associated with energy metabolism and are
319 known to be associated with ammonia oxidizing archaea, including nitrite reductase (K00368),
320 the oxygen utilizing terminal oxidase subunit I (K02274), the ammonium transporter (K03320), a
321 duplicated NADPH:Quinone oxidoreductase subunit M (K00342), and a split cytochrome b-561
322 like protein (K15879), as well as 3 iron-sulfur complex assembly proteins (K09014, K09015, and

323 K13628), a cytochrome c oxidase complex assembly protein (K02259), a high affinity iron
324 transporter (K07243), and a copper transporter (K14166).
325
326 **Metabolic reconstruction supports an amino acid based metabolism.** We undertook a
327 complete metabolic reconstruction for the type strain genome (Angelarchaeales-1) to evaluate
328 the feasibility of a life strategy where amino acid metabolism is coupled to ammonia oxidation.
329 We focused on pathways for the import and catabolism of amino acids and routes by which their
330 products feed into central carbon metabolism, ammonia oxidation, and are interconnected with
331 electron transport and energy generation (**Fig. 5 and Supplementary Tables 11-12)**.
332
333 As is typical of the Angelarcheales, the Angelarchaeales-1 genome lacks key oxidative enzymes
334 of glycolysis and oxidative enzymes of the pentose phosphate pathway. They can interconvert
335 glucose/fructose to mannose and galactose derivatives but appear unable to import or
336 phosphorylate these sugars. The encoded fructose bisphosphatase would allow
337 gluconeogenesis. They have no detectable pyruvate kinase and instead encode a pyruvate
338 orthophosphate dikinase, which is known to be allosterically regulated and reversible in
339 archaea[39,40]. The capacity for production of the compatible solute trehalose is notable, as
340 Angelarchaeales-1 derives from an environment that regularly undergoes large cyclic changes in
341 water content[23].
342
343 Angelarchaeales-1 encodes a full complement of genes for the conversion of pyruvate into Acetyl-
344 CoA, the TCA cycle, and a glyoxylate shunt. However, it lacks genes for both the pyruvate
345 dehydrogenase complex as well as the 2-oxoglutarate dehydrogenase complex. These reactions
346 are likely enabled by pyruvate/2-oxoglutarate ferredoxin oxidoreductase systems, which provide
347 reduced ferredoxin. A glyoxylate shunt allows for catabolic reactions that terminate in 2-carbon

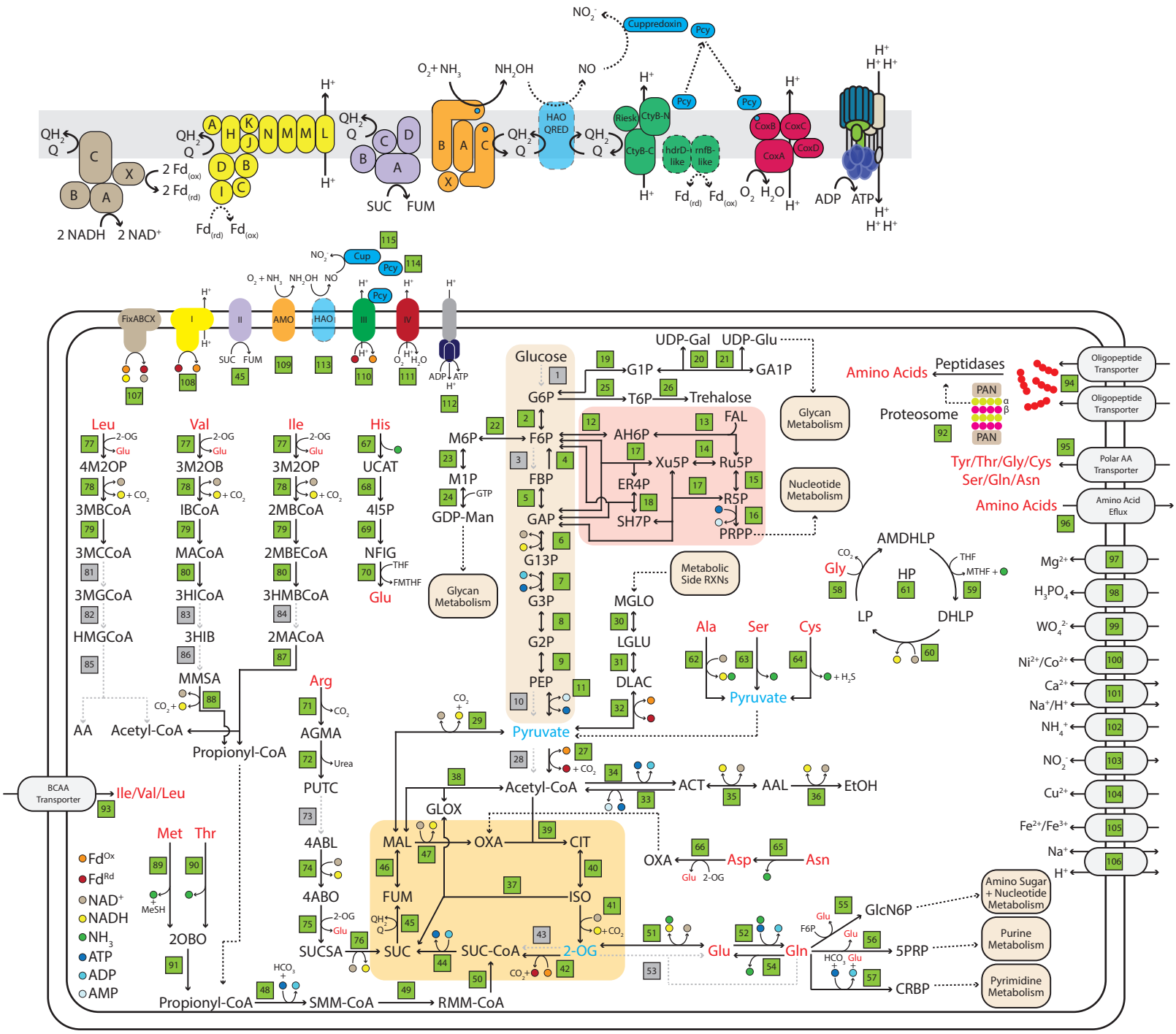

Figure 5 I Metabolic reconstruction of Angelarcheales-1 genome. For full reaction, gene list, and list of compound abbreviations see Supplementary Tables 12 and 13. Green boxes indicate a reaction (and its reference number in Supplementary Table 12) that could be linked to a gene with the predicted metabolic function. Black arrows with solid lines indicate a reaction that could be identified, associated smaller arrows with colored dots indicate consumed, and generated reaction substrates and products are indicated in the key at the bottom left. Black arrows with dotted lines indicate flow of metabolites to other pathways or reactions. Grey arrows with dotted lines and grey boxes indicate reactions that were searched for and could not be identified. Amino acids are in red text to highlight their locations throughout the figure. Metabolites in blue text indicate hubs for carbon derived from amino acid catabolism. For ease of viewing the reactions of glycolysis, the pentose phosphate pathway, and the TCA cycle have been highlighted with beige, red, and orange backgrounds. The upper panel is a blow up of the electron transport reactions showing predicted organizations of subunits in each complex. Reference numbers for each subunit can be found in the larger figure panel, and colors of subunits are the same as those used in that figure panel. Transparent HAO-QRED indicates a putative/proposed functionality. Black arrows with dotted lines indicate putative reactions. Protein subunits with solid color but dotted borders indicate a protein was found but functionality is unclear.

348 compounds (e.g. acetyl-CoA produced by amino acid and acetate catabolism) to be utilized for
349 biosynthetic purposes, as it bypasses the two decarboxylation steps of the TCA cycle.
350
351 We identified reasonably confident catabolic routes for 15 amino acids, including a complete
352 glycine cleavage system, as well as a route for the end product (propionyl-CoA) of at least 4 amino
353 acid catabolic pathways, to be incorporated into the TCA cycle as succinyl-CoA. Genes encoding
354 the terminal reactions of branched chain amino acid degradation were not identified, although the
355 genome encodes numerous acyl-CoA dehydrogenases with unknown specificity that could
356 perform these functions. However, we could confidently identify the branched chain keto-acid
357 dehydrogenase complex (BCKAD) that is critical for the degradation of leucine, isoleucine, and
358 valine, as well as in processing the downstream degradation products of methionine and
359 threonine. Finally, this organism carries multiple independent branched chain amino acid
360 transport systems, as well as a polar amino acid transport system that is enriched in the Ca.
361 Angelarcheales order **(Fig. 4b, Supplementary Fig. 6c-d, and Supplementary Table 12).**
362
363 Angelarchaeales-1 and ammonia oxidizing Nitrososphaerales both have respiratory chains that
364 include a complex I lacking the E, F, and G subunits for NADH binding and a duplicated subunit
365 M that may mediate translocation of an additional proton[41]**.** The electron donor to the complex I
366 may be reduced ferredoxin[41,42]. Both groups encode a four-subunit succinate/fumarate
367 dehydrogenase, a cytochrome b-like complex III with an associated plastocyanin-like electron
368 transfer protein, and an oxygen utilizing cytochrome c terminal oxidase complex.
369
370 Unlike the Nitrososphaerales, Angelarchaeales-1 encodes a multitude of systems for the putative
371 utilization of ferredoxin**.** This includes a FixABCX electron bifurcation system that can couple the
372 reduction of ferredoxin and quinone to the oxidation of NADH. Interestingly the FixABCX complex

373 is co-located with the BCKAD complex in the Angelarchaeales-1 genome. This FixABCX complex
374 may be important for converting the reducing power of NADH derived from BCKAD mediated
375 branched chain amino acid degradation into reducing power in the form of reduced ferredoxin and
376 quinone. Angelarchaeales-1 has 2 genes proximal to the cytochrome b-like complex III, a hdrD-
377 like gene and a rnfB-like gene that are not found in Nitrososphaerales genomes. In rnf complexes,
378 rnfB binds and oxidizes reduced ferredoxin, however the functions of these iron-sulfur cluster
379 containing proteins, which are not normally found associated with complex III, are still unclear.
380

381 **DISCUSSION**

382

383 In this study, we identified a novel CuMMO occurring in a largely unstudied lineage of archaea
384 that we show is widely distributed in soils, terrestrial sediments, and even detected in deep ocean
385 systems. The ability of a CuMMO to function as an ammonia monooxygenase requires a
386 metabolic context that supports the acquisition of ammonia, processing of ammonia oxidation
387 byproducts, and acquisition of energy from these reactions. Thus, the question of whether the
388 Angelarcheales CuMMOs function as ammonia vs. methane monooxygenases can be addressed
389 in part by comparison of the genomic context to those of ammonia monooxygenases in
390 Nitrososphaerales. The Nitrososphaerales encode transporters for ammonia uptake. These are
391 also present in Angelarcheales, but probably are not needed, given the ability to derive ammonia
392 from amino acid degradation. Further, we show (i) Conservation between the protein architecture
393 and active site residues in known archaeal amoA, amoB, and amoC ammonia monooxygenases
394 and those found in Ca. Angelarcheales. (ii) A large inventory of cupredoxin-like BCPs including
395 BCP subtypes (such as 2-domain and medium length 1-domain BCPs) in both Ca.
396 Angelarcheales and Nitrososphaerales relative to sibling lineages of their respective phyla. (iii) A
397 combination of electron transport proteins that is shared by Ca. Angelarcheales and

398 Nitrososphaerales, consistent with similar electron transport chain functionality and energy
399 acquisition strategies.
400
401 The presence of extensive inventories of blue copper proteins is of particular interest, as
402 Nitrososphaerales lack a hao homologue, and BCPs are thought to perform the yet unresolved
403 enzymatic steps that oxidize hydroxylamine to nitrite[16,17]. Work in Nitrososphaerales generally
404 finds that while having a large number of these proteins is a clear signature of this group, the
405 exact gene set of BCPs present in any Nitrososphaerales genome is highly variable[16]. Candidate
406 enzymes including specific plastocyanin-like cupredoxins[19] or nirK homologues[21,43] have been
407 proposed to carry out specific steps in hydroxylamine oxidation, however no individual enzyme is
408 completely conserved across known ammonia oxidizing archaeal genomes. Thus, a variety of
409 diverse BCPs may substitute for the missing hao activity. In Ca. Angelarcheales, we observe a
410 similar pattern, i.e., an extensive inventory of BCPs relative to sister lineages with no individual
411 BCP ortholog being conserved across the entire group. Instead, we observe sets of 2-domain
412 BCPs and 1-domain medium length BCPs with divergent sequences but conserved architecture
413 that are largely restricted to the Ca. Angelarcheales and Nitrososphaerales lineages. BCP
414 architectures are in fact well conserved across the entire lineages, respectively. Overall, the
415 weight of genomic evidence supports the conclusion that the Ca. Angelarcheales are a new and
416 distinct group of ammonia-oxidizing archaea.
417
418 Ecological support for the deduction that Ca. Angelarcheales are ammonia oxidizers is their
419 strong association with, and high abundances in, relatively aerobic grassland soils, where
420 organisms with the capability for methanogenesis have not been detected[23,44]. These archaea
421 also occur in sediments from above the water table and in saturated sediments that periodically
422 receive oxygenated water[45]. In this context, it is interesting to note that the Ca. Angelarcheales

423 can utilize amino acids as a carbon and nitrogen source without oxygen, generating energy via
424 substrate level phosphorylation (as we detect the ability to ferment acetyl-CoA to ethanol), and
425 oxidizing ammonia when O_2 is available.
426
427 While we acknowledge that these data cannot represent all soils and sediments, it is striking that
428 Ca. Angelarcheales can comprise a reasonably large fraction of communities in which they were
429 detected. Also we find that the abundances of Ca. Angelarcheales are positively associated with
430 both Nitrososphaerales and Nitrospirales, two groups that participate in nitrification. We note that
431 while Ca. Angelarcheales shared significant positive associations with a number of other microbial
432 groups, their associations with both Nitrososphaerales and Nitrospirales were some of the
433 strongest detected. While the abundances of Angelarcheales and Nitrososphaerales are
434 positively associated, we predict that the niches of Angelarcheales and Nitrososphaerales do not
435 overlap, given that Angelarcheales can use ammonia that they liberate from breakdown of amino
436 acids whereas Nitrososphaerales likely use ammonia (and urea) directly. The strong association
437 of both Angelarcheales and Nitrososphaerales with the nitrite oxidizing Nitrospirales is consistent
438 with the former groups producing nitrite from ammonia oxidation which is then exchanged with
439 and oxidized by the Nitrospirales.
440
441 The identification of a lineage of archaea outside of the Nitrososphaerales involved in ammonia
442 oxidation has important implications for biogeochemical nitrogen cycling. These archaea are
443 unusual, as they both degrade proteins/amino acids and oxidize the product ammonia.
444 Angelarcheales likely occur in a variety of soils and sediments where ammonia oxidation has
445 gone undetected using primer-based methods. The new clade of CuMMOs provides an
446 opportunity for further study of these enzymes, and for molecular clock-based analyses of their
447 evolution relative Earth history[46].

448

449 **METHODS**

450

451 **Amo/Pmo identification, genome set selection, and dereplication**

452 Previous genome-resolved work at the Angelo Coast Range Reserve[23] identified 7
453 Thermoplasmatota genomes each containing a hit to the KOFAM Hidden Markov Model (HMM)
454 for amoA/pmoA (K10944), below the HMM score threshold, but with a significant E-value (≤ $1E^{-}$
455 ^{4}). Proteins co-localized with these divergent amoA/pmoA proteins were identified. To search for
456 additional similar divergent amoA/pmoA proteins we aligned the divergent proteins using MAFFT
457 v7.471 (--maxiterate 1000 --localpair) and constructed an HMM using hmmbuild in the HMMER v3.3.1
458 package[47] with default parameters. This HMM was scored against all archaeal genomes in the
459 Genome Taxonomy Database (GTDB) release r95[48], and all archaeal MAGs in ggKbase
460 (ggKbase.berkeley.edu) datasets as of January 13, 2020 using the hmmsearch function of the
461 HMMER package with an HMM score threshold of 100. These genomes were phylogenetically
462 classified using the GTDB tool kit[49] (GTDB-Tk v1.3.0) classify workflow with default parameters.
463 All genomes were placed within the RBG-16-68-12 order (hereafter Ca. Angelarcheales) of the
464 GTDB taxonomy (**Supplementary Table 1**).

465 To produce a full genome reference set for our analyses we added an additional 31
466 genomes that did not have any hit to our custom amoA/pmoA HMM but fell within the Ca.
467 Angelarcheales order (**Supplementary Table 1**). These genomes came both from Angelo Coast
468 Range Reserve assemblies (n = 15 genomes) and from the Ca. Angelarcheales order in GTDB
469 (n = 16 genomes). We also added 719 reference genomes from GTDB derived from the archaeal
470 phyla Thermoplasmatota (n = 338 genomes) and Thermoproteota (n = 371). Thermoproteota
471 genomes were included for comparison of functional differences within and between phyla of
472 known (Nitrososphaerales) and putative (Ca. Angelarcheales) ammonia oxidizing archaea.

473 The full genomes set was de-replicated at the species level (Average Nucleotide Identity
474 ≥ 95 %) using dRep v3.0.1[50] with the following parameters: -p 16 -comp 10 -ms 10000 -sa 0.95. The
475 best genome from each species cluster was chosen as a representative genome by dRep.
476 Species representatives were required to have ≥ 60 % completeness and ≤ 10 % contamination
477 as estimated by checkM[51]. If no genome within a species cluster met these criteria the cluster
478 was discarded. All genome information can be found in **Supplementary Tables 2 and 4**.
479

480 **Amo/Pmo protein identification, alignment, analysis, and phylogenetic reconstruction**

481 The amoA/pmoA protein was used as an anchor sequence to manually annotate putative
482 amoB/pmoB, amoC/pmoC, and amoX proteins present on each contig across the Ca.
483 Angelarcheales genomes (**Supplementary Table 2**). Manual annotation used a combination of
484 protein order relative to the amoA/pmoA sequence, predicted protein length (compared to known
485 amo/pmo proteins), and best blast hits vs. the NCBI nr database.

486 To identify amoB/pmoB and amoC/pmoC in Ca. Angelarcheales genomes where no
487 amoA/pmoA was identified, and to develop a method to rapidly identify all novel amo/pmo
488 complex proteins in future work we aligned each set of amo/pmo proteins following manual
489 annotation using MAFFT v7.471 (--maxiterate 1000 --localpair) and constructed HMMs for each using
490 hmmbuild in the HMMER v3.3.1 package with default parameters. These were scored against all
491 proteins in the full set of 228 redundant Ca. Angelarcheales genomes (**Supplementary Table 1**)
492 using the hmmsearch function of the HMMER v3.3.1 package with an HMM score threshold of
493 100.

494 Alignments of putative amo/pmo sequences with known references were constructed by
495 merging amo/pmo sequences from Ca. Angelarcheales were with reference sequences from
496 Khadka, et al.[1] and those from an additional set of 22 ammonia oxidizing archaeal reference
497 genomes (**Supplementary Table 3**). This resulted in protein sets for amoA/pmoA, amoB/pmoB,

498 and amoC/pmoC that contained 112, 114, and 110 sequences respectively. Sequence sets were
499 aligned using MAFFT v7.471 with the following parameters: --maxiterate 1000 --localpair --reorder --thread
500 12. Mean amino acid identity between Ca. Angelarcheales, bacterial, and archeal reference
501 sequences were calculated in Geneious Prime v2020.2.4 from a pairwise sequence identity
502 matrix. Conserved residues for methane and ammonia monooxygenases were referenced from
503 Wang, et al.[29] and identified in each alignment through manual inspection.

504 Maximum likelihood phylogenetic trees were constructed for each individual amo/pmo
505 subunit alignment using IQ-TREE v1.6.12[52] with the following options: -m MFP -bb 1000 -alrt 1000 -nt
506 12. The empirically selected evolutionary rate model for all amo/pmo sequence sets was
507 LG+F+I+G4 based on Bayesian information criteria (BIC). Branch support was estimated using
508 ultrafast bootstrapping with 1000 bootstrap replicates. Individual amo/pmo protein trees can be
509 found in **Supplementary Figure 3**. For the combined amoABC/pmoABC tree, sequences from
510 the same organism were concatenated in Geneious Prime v2020.2.4 and only retained if at least
511 2 of the proteins were present, resulting in 112 total concatenated sequences in the alignment. A
512 Maximum likelihood phylogenetic tree was constructed for concatenated sequences using IQ-
513 TREE v1.6.12 with the following options: -m MFP -bb 1000 -alrt 1000 -nt 12. The empirically selected
514 evolutionary rate model for the concatenated amoABC/pmoABC tree was LG+F+I+G4 based on
515 BIC. Branch support was estimated using ultrafast bootstrapping with 1000 bootstrap replicates.
516

517 **Assessment of AmoA/PmoA primer complementarity**

518 Primer sequences used to amplify and quantify both archaeal (GenAOAF: 5'-ATA GAG
519 CCT CAA GTA GGA AAG TTC TA-3' ; GenAOAR: 5'-CCA AGC GGC CAT CCA GCT GTA TGT
520 CC-3') and bacterial (amoA-1Fmod: 5'-CTG GGG TTT CTA CTG GTG GTC-3' ; GenAOBR: 5'-
521 GCA GTG ATC ATC CAG TTG CG-3') amoA/pmoA sequences using PCR from environmental
522 samples were obtained from Meinhardt, et al.[28]. Sequence complementarity was assessed using

523 the map primers function of Geneious Prime v2020.2.4. Primers were tested against nucleotide
524 sequences of amoA genes from Nitrososphaerales genomes and amoA/pmoA genes from Ca.
525 Angelarcheales genomes as detailed in **Supplementary Table 3**. A match required both the
526 forward and reverse primers to bind to the sequence while allowing up to 7 mismatched bases in
527 each primer. Full data obtained for the number of mismatches and estimated product sizes for
528 each primer pair and template sequence are available in **Supplementary Table 3**.
529
530 **Genome taxonomy and phylogenetic reconstruction**
531 Initial taxonomic placement for the 645 non-redundant genomes used in this study was
532 performed using the GTDB-Tk[49] classify workflow with the following parameters: classify_wf -x fasta
533 --cpus 48. All GTDB-Tk based taxonomic classification is available in **Supplementary Table 4**. A
534 concatenated marker gene phylogenetic tree for all genomes classified within the archaeal
535 phylum Thermoplasmatota was constructed by combining the 34 Ca. Angelarcheales genomes
536 with 302 de-replicated reference genomes from this phylum spanning all known orders. The
537 *Archaeoglobus fulgidus* (GCF_000008665.1) genome was also included to be used as an
538 outgroup for tree rooting. GToTree v1.5.22[30] was used to identify and extract a set of 76
539 phylogenetically informative single copy archaeal marker genes (SCGs), defined in Lee, et al.[30],
540 from each genome using the following parameters: -H Archaea.hmm -j 8 -d. Genomes where < 50%
541 (38 genes) of the targeted marker genes could not be identified were removed from the analysis
542 (n = 10 genomes) retaining a total of 327 genomes in the final tree. SCG sequence sets were
543 then individually aligned with Muscle v3.8.31[53] and alignments were trimmed with Trimal v1.4[54].
544 All alignments were then concatenated, and an SCG alignment partition table was produced by
545 GToTree so evolutionary substitution rate models could be estimated for each SCG independently
546 during phylogenetic tree construction. A maximum likelihood phylogenetic tree was constructed
547 with IQ-TREE v1.6.12[52] with the following options: -spp Partitions.txt -m MFP -bb 1000 -alrt 1000 -nt 48.

548 Evolutionary rate models were empirically estimated for each marker gene independently and
549 selected based on BIC. Branch support was estimated using ultrafast bootstrapping with 1000
550 bootstrap replicates. Phylogenetic trees were rooted using *Archaeoglobus fulgidus* as an
551 outgroup, annotated, and displayed using iTOL v6. For the full tree see **Supplementary Figure**
552 **4**.
553
554 **RpL6 marker abundance and association analysis**
555 There are 6 study sites where at least 1 Ca. Angelarcheales genome was reconstructed
556 from shotgun metagenome data. We used all 185 shotgun metagenomic samples collected from
557 these sites, regardless of whether a Ca. Angelarcheales genome was recovered from a sample,
558 to estimate the relative abundance of the Ca. Angelarcheales order across these locations
559 (**Supplementary Table 5)**. Relative abundance of all bacterial and archaeal order level taxonomic
560 groups was quantified using marker gene taxonomic placement and quantification of the
561 phylogenetically informative SCG ribosomal protein L6 (rpL6). RpL6 marker gene profiling and
562 quantification was performed using GraftM v0.13.1[55]. Briefly, an rpL6 graftM database was
563 constructed using the ribosomal L6 protein sequences from all archaeal and bacterial genomes
564 (provided by GTDB using TIGR03653 and TIGR03654, respectively) in GTDB v95. A GraftM
565 package was then created using this set of sequences and the rpL6 Pfam HMM (PF00347.24)
566 sensitive for both bacterial and archaeal variants using the command: graftM create --sequences L6.faa
567 --taxonomy taxonomy.tsv --hmm PF00347.hmm. GraftM was then used to call genes, identify rpL6
568 sequences, phylogenetically place sequences, and quantify read counts of rpL6 sequences
569 identified in each sample using the following options: graftM graft --forward [forward reads] --reverse
570 [reverse reads] --graftm_package rpL6_gpkg --threads 48. Raw counts in each sample were aggregated to
571 the taxonomic rank of order using a custom R script
572 (https://github.com/SDmetagenomics/AMO_Archaea_2021). Counts of rpL6 sequences that

573 could only be resolved to taxonomic ranks higher than order level were retained in their original
574 form. All aggregated count data is available in **Supplementary Table 6**. Relative abundances of
575 the Ca. Angelarcheales and Nitrososphaerales orders in each sample were calculated as the
576 fraction of total reads in a sample that were associated with these order ranks. Plotting was
577 performed in R using the ggplot2 package[56].

578 Association analysis between the abundances of each taxonomic group across samples
579 (**Supplementary Table 7**) was conducted using the propr package[57] in R. Proportionality was
580 chosen as an association measure over correlation as it is better suited for compositional data
581 such as sequencing read counts[58]. Briefly, taxonomic groups were filtered such that only those
582 with ≥ 5 counts in at least 33 % of samples (n = 61 samples) were retained. This filtering was
583 performed to remove taxonomic groups with extremely low counts from association comparisons
584 as to avoid spurious associations. Following filtering 90.5 % of the original count data was retained
585 in each sample on average. Subsequently the rho proportionality metric was calculated for
586 centered log-ratio normalized counts between all pairs of taxonomic groups using the following
587 function: propr(count_matrix, metric = "rho", p = 1000). The magnitude of proportionality that represented
588 a significant association (FDR < 0.001) was calculated through a permutation based procedure
589 implemented using the updateCutoffs function within the propr package as follows:
590 updateCutoffs(L6_rho, cutoff = seq(from = 0.05, to = 0.40,by = 0.01), ncores = 12). It was determined that all rho
591 values ≥ 0.25 were statistically significant with a FDR < 0.001. Additionally, only positive rho
592 association values were considered in the final analysis, as negative associations in
593 compositional data are considered less reliable, even after compositional correction measures
594 are applied[57]. All significant proportionality values are available in **Supplementary Table 7**. The
595 abundance values in the plotted pairwise comparisons between Ca. Angelarcheales,
596 Nitrososphaerales, and Nitrospirales use the centered log-ratio of counts in each sample as
597 calculated by the propr package. All plotting was performed using the ggplot2 package in R.

598

599 **Genome annotation**

600 For the 645 genomes passing completeness and contamination quality criteria, annotation
601 was performed as follows. Genes and protein sequences were predicted using prodigal v2.3.6[59]
602 using the following options: prodigal -i [genome contigs] -a [proteins sequences out] -m -p meta. KEGG KO
603 annotations were predicted using KofamScan[60] using HMM models from release r02_18_2020
604 with the following options: exec_annotation -p [hmm profiles] -k [hmm cutoffs] --cpu 48 --tmp-dir [temp dir] -o
605 [output folder] [protein file]. As multiple KEGG HMMs can match to the same protein with scores
606 exceeding their score cutoff thresholds, the HMM with the lowest E-value had its annotation
607 transferred to the protein. If a protein did not match any KEGG HMM above the HMM cutoff
608 threshold then the lowest E-value annotation was transferred. An E-value cutoff of < 1e-10 was
609 applied, above which no annotations were transferred, and genes were not assigned to a KO.
610 Archeal COG (arCOGs) annotations were predicted for all proteins using HMMs from EggNOG
611 v5[61] using the following options: hmmsearch --tblout [arCOG hit table] -E 0.0001 --cpu 10 All_arCOG.hmm
612 All_Proteins.faa. We searched for XoxF/mxaF-like pqq-binding methanol dehydrogenases in
613 genomes using the hmmsearch function of the HMMER package and a custom HMM from
614 Anantharaman, K. et al.[26] with an HMM score threshold cutoff of ≥ 166 using the following options:
615 hmmsearch --tblout [xoxF_mxaF hits] -T 166 --cpu 12 methanol_dehydrogenase_pqq_xoxF_mxaF.hmm
616 All_Proteins.faa. Protease counts in each genome were determined via the METABOLIC pipeline
617 (https://github.com/AnantharamanLab/METABOLIC) implemented using the following options:
618 perl METABOLIC-G.pl -t 48 -m-cutoff 0.75 -in [input protein files] -kofam-db full -o [output annotations]. Custom R
619 code for parsing and plotting METABOLIC outputs used for peptidase quantification can be found
620 in our github repository. Pfam domain annotations were predicted by searching all proteins
621 against the PfamA database release r32[62] using the hmmsearch function of the HMMER package
622 with the following options: hmmsearch --domtblout [pfam domain table] --cut_ga --cpu 10 Pfam-A.hmm

623 All_Proteins.faa. Overlapping pfam domain matches to the same protein were resolved, and domain
624 boundaries were established, using the cath-resolve-hits function of the cath-tools package with
625 default parameters (https://github.com/UCLOrengoGroup/cath-tools). All annotations were
626 aggregated into a final table using a custom R script which is available in our github repository.
627 Also the complete gene level annotation table for all 645 genomes is available at FigShare
628 (https://figshare.com/projects/AMO_Archaea/112599).

629 For functional annotations of genes and pathways across all Ca. Angelarcheales
630 genomes, a subset of 190 specific target functions were analyzed (**Fig. 2b and Supplementary**
631 **Figure 6a)**. Pathways and protein complexes were subsequently grouped into 77 "gene groups"
632 that represent the capability to perform a pathway or metabolic function. Each gene group had
633 specific criteria that were required (i.e. a critical protein needed to be present) for a positive
634 detection. All targets, gene group organization, and criteria for detection are available in
635 **Supplementary Table 8**. For annotation and quantification of amino acid and peptide transport
636 systems across all 645 genomes in our analysis a set of 38 KEGG KOs representing individual
637 transporter proteins or transporter subunits were used as the search criteria (**Supplementary**
638 **Figure 6c-d and Supplementary Table 14**)**.** For functional annotation of the Angelarcheales-1
639 genome and metabolic map reconstruction (**Fig. 5**), a subset of 233 specific target functions were
640 analyzed that highlighted central carbon metabolism, the degradation of amino acids, and electron
641 transport/energy generation (**Supplementary Table 12).** Pathways and protein complexes were
642 subsequently grouped into 115 "reaction groups" that represent the capability to perform a
643 pathway or metabolic function. Each reaction group had specific criteria that were required (i.e. a
644 critical protein needed to be present) for a positive detection. All targets, reaction group
645 organization, genes associated with reactions, and criteria for detection are available in
646 **Supplementary Table 12.**
647

648 **Protein clustering**

649 Clustering of all proteins predicted in the 645 genomes passing completeness and
650 contamination quality criteria was accomplished as described in Méheust, et al.[33]. Code for this
651 pipeline is available at: https://github.com/raphael-upmc/proteinClusteringPipeline. Briefly, all
652 1,103,913 proteins were first clustered into subfamilies (subfams) using the subfams.py script of
653 the pipeline (which implements mmseqs2[63] clustering) using the following options: subfamilies.py --
654 output-directory [clustering dir] --cpu 48 --coverage 0.8 All_Proteins.faa. All proteins within a subfam must align
655 with bidirectional coverage of at least 80 % (--cov-mode 0 in mmseqs2), and alignments must
656 have an E-value < 1e-4. Proteins which did not cluster into a subfam of at least 2 proteins were
657 discarded from this analysis leaving a total of 964,644 (87.4 %) proteins with subfamily
658 assignments. A total of 76,216 protein subfam clusters were formed. Subsequently proteins within
659 each subfam were aligned, HMMs were constructed from these alignments, and all-v-all HMM
660 scoring was conducted using the hhblits.py script (which implements the hhblits function within
661 the hhsuite v3.0[64] software package) using default options. For HMM-HMM scoring any local
662 alignments between HMMs of different subfams had to have an hhblits probability score of ≥ 95
663 % to be retained for downstream analysis. Finally, family groupings of subfams (fams) were
664 formed by applying the Markov clustering algorithm (MCL)[65] to the network of all HMM-HMM
665 connections using the runningMclClustering.py script of the pipeline using the following options:
666 runningMclClustering.py --force --min-size 2 --cpu 4 --fasta config.json. This resulted in the formation of 19,828
667 protein fam clusters.

668

669 **BCP identification, quantification, and phylogenetics**

670 Cupredoxin-like blue copper proteins (BCPs) were initially identified in our dataset by
671 selecting all proteins that carried one of the following Pfam domains: Copper-bind, COX2,
672 COX_ARM, Cu-oxidase, Cu-oxidase_2, Cu-oxidase_3, Cu_bind_like, Cupredoxin_1, CzcE, DP-

673 EP, Ephrin, hGDE_N, PAD_N, PixA, SoxE. These domains are all members of the Pfam
674 CU_oxidase clan (CL0026). Due to the large sequence divergence in cupredoxin-like domains
675 (as evidenced by the number of pfam models required to appropriately capture their diversity), we
676 posited that many divergent domain sequences would be missed by direct Pfam annotation.
677 Alternatively, more sensitive local HMM-HMM comparison at the subfam level would cluster the
678 majority of BCP domain containing proteins into a single fam cluster. Thus, we quantified the
679 number of BCP domain containing proteins present in all fams. Given that 90.2% of proteins with
680 annotated BCP domains were members of fam00018, we used fam00018 to represent BCP
681 domain containing proteins in our dataset. Comparison of fam00018 protein counts in genomes
682 was carried out independently for the archaeal phyla Thermoplasmatota and Thermoproteota
683 between all archeal orders within these phyla that contained ≥ 3 genomes. Global significant
684 differences across all orders within a phylum was first tested using the Kruskal-Wallis rank sum
685 test implemented as the kruskal.test() function in R ($\alpha \leq 0.05$). Pairwise significant differences
686 between orders within a phylum was then tested using the Wilcoxon rank sum test implemented
687 as the pairwise.wilcox.test() function in R. P-values from tests were corrected for multiple
688 comparisons using false discovery rate (FDR) with a value of FDR ≤ 0.05 being considered
689 significant. All aggregation, quantification, statistical testing, and plotting of BCP protein count
690 data was performed using a custom R script available in our github repository.

691 Manual subfamily level annotation and domain architecture analysis was performed on
692 100 fam00018 subfams that contained at least 5 proteins (85.2% of all fam00018 proteins). The
693 proteins in each subfam were re-aligned with MAFFT v7.471 using the following options: mafft --
694 maxiterate 1000 --localpair --reorder --thread 12 [proteins in] > [alignment out]. HMMs were built for each
695 alignment using hhsuite v3.0, and HMM-HMM scoring was performed against the PfamA
696 database release r32 using hhsearch with the following options: hhsearch -i [input hmm] -o [output table]
697 -d [pfam hmm database] -p 50 -E 0.001 -z 1 -Z 32000 -b 0 -B 0 -n 1 -cpu 18. Domain matches with a probability

698 score of ≥ 95 % were retained, and Pfam domains overlapping the same region on target subfam
699 HMMs were resolved with the cath-resolve-hits function of the cath-tools package using default
700 parameters. Each of the 100 subfams were then manually annotated and placed into 1 of 6 broad
701 classification groups based on the subfam level Pfam domain architectures and the types of
702 KEGG and arCOG annotations assigned to individual proteins within the subfam. Annotation
703 groups were defined as follows: COX2 (Contains COX2 domain and KEGG or arCOG annotations
704 indicate > 40% of proteins in cluster are terminal oxidase subunits), 1D_Cu_Small (one Pfam
705 BCP domain, mean protein length < 250 amino acids), 1D_Cu_Med (one Pfam BCP domain,
706 mean protein length 250 - 400 amino acids), 1D_Cu_Large (one Pfam BCP domain, mean protein
707 length > 400 amino acids), 2D_Cu (two Pfam BCP domains), 3D_Cu (three Pfam BCP domains).
708 Due to its presence in Angelarcheales-1 and its proximity to the coxAB locus subfam00588 was
709 also manually annotated as above despite having < 5 proteins. Assignment of copper site types
710 (e.g. Type 1 Copper) was conducted by manual inspection of subfamily alignments and
711 referenced from Gräff, et al.[35]. All subfams within fam00018 and their associated annotations are
712 available in **Supplementary Table 9**.

713 Phylogenetic tree construction for manually annotated subfams with > 1 BCP domain was
714 undertaken as follows: All proteins from subfams in the 2D_Cu and 3D_Cu manual annotation
715 groups (n = 349 proteins) were combined with reference laccase and nitrite reductase (nirK)
716 sequences from Decleyre, et al.[66], Kobayashi, et al.[36], and Nakamura, et al.[67] (n = 90 proteins).
717 Proteins were aligned with MAFFT v7.471 using the following options: mafft --maxiterate 1000 --
718 genafpair --reorder --thread 12 [proteins in] > [alignment out]. A maximum likelihood phylogenetic tree was
719 constructed using IQ-TREE v1.6.12 with the following options: -m MFP -bb 1000 -alrt 1000 -nt 12. The
720 empirically selected evolutionary rate model for the tree was WAG+F+R7 based on BIC. Branch
721 support was estimated using ultrafast bootstrapping with 1000 bootstrap replicates. Tree was
722 annotated, and displayed using iTOL v6. Tree clades were manually defined based on positioning

723 of reference sequences within clusters and copper site types present in sequences defined in
724 Nakamura, et al.[67].
725
726 **Gene co-occurrence analysis and locus plotting**
727 Contigs containing genes for the amoCAXB cluster, the coxAB oxygen utilizing terminal
728 oxidase cluster, and the complex III like cytochrome b gene cluster were identified in Ca.
729 Angelarcheales genomes as follows: amoCAXB contigs were identified as any contig encoding
730 an amo/pmo subunit (as described above); coxAB containing contigs were identified as any contig
731 encoding a gene that matched to the KEGG HMM for K02274 (coxA); and complex III like gene
732 cluster containing contigs were identified as any counting encoding a gene that matched the
733 arCOG HMM for arCOG01721 (cytochrome b of bc complex). All loci were extracted from our
734 master annotation table using a custom R script available in our github repository, and loci were
735 displayed using the gggenes package in R (https://github.com/wilkox/gggenes). For ease of
736 viewing amoCAXB and coxAB containing contigs were truncated to display 10 genes on either
737 side of the gene cluster of interest. Complex III containing contigs were truncated to display 22
738 genes on either side of the gene cluster as to allow inclusion of other proximal respiratory
739 complexes.
740
741 **KEGG Ortholog (KO) enrichment analysis**
742 Detection of enriched KO terms between archaeal orders was carried out using a custom
743 R script available in our github repository. Archaeal orders from the phyla Thermoplasmatota and
744 Thermoproteota that contained < 3 genomes were excluded from the analysis. For the remainder
745 of genomes, a table was generated giving the count of all observed KOs (n = 5,929 total unique
746 KO terms) in every genome. KOs that occurred less than 10 times across all genomes were
747 filtered from the analysis. The filtered KO count table along with the taxonomic order level

748 groupings of genomes were used as the input for indicator species analysis implemented as the
749 multipatt() function in the R indicspecies package[68] using the following options: multipatt(x = [KO count
750 matrix], cluster = [genome taxonomic assignment], func = "IndVal.g", max.order = 2, restcomb = c(1:25, 262), control =
751 how(nperm = 9999)). We analyzed the enrichment of KO term frequency in all orders individually as
752 well as a grouping consisting of the combined genomes from orders Ca. Angelarcheales and
753 Nitrosospherales. This allows for the identification of enriched functions that are shared by these
754 two groups relative to all other lineages and to each other individually. Statistical significance of
755 enriched KO frequency was estimated by permutation using 9,999 permuted groupings of the
756 genomes. P-values were corrected for multiple testing using FDR with a value of FDR ≤ 0.05
757 being considered significant. Additionally only KOs with an indicator value ≥ 0.4 were retained for
758 downstream analysis. KOs enriched in Nitrosospherales (Gp 14), Ca. Angelarcheales (Gp 17), or
759 shared by both orders (Gp 26) were displayed as a heatmap using the superheat package in R.
760 For a list of all significant KOs and their functions see **Supplementary Table 10**. Functional
761 category assignments for KOs were derived from KEGG orthology group hierarchies and
762 manually curated, and quantified in R. The full list of KO to functional category assignments is
763 available in our github repository.
764

765 **DATA AVAILABILITY**

766 Genomic data, including assembled genomes and raw sequencing reads, are available
767 under the following NCBI BioProject accession numbers: PRJNA449266 and PRJNA288027. Ca.
768 Angelarcheales genomes from the project can also be found at ggKbase
769 (https://ggkbase.berkeley.edu/AMO_Archaea). Large datasets including an ORF-to-Genome
770 table, all predicted proteins, and the full gene level annotation table are available at figshare
771 (https://figshare.com/projects/AMO_Archaea/112599).
772

CODE AVAILABILITY

Code use for the analysis in this report are available at the following github repository: https://github.com/SDmetagenomics/AMO_Archaea_2021

REPORTING SUMMARY

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

ACKNOWLEDGEMENTS

The authors thank Ka Ki (Lily) Law and Shufei Lei for data support and Alex Thomas for helpful discussion, and the Joint Genome Institute for a subset of the sequencing. We acknowledge Dong, et al. (2019) whose public data we used in this study. This work was facilitated by samples previously collected at the University of California Natural Reserve system Angelo Coast Range Reserve (10.21973/N3R94R). Funding was provided by the Office of Science, Office of Biological and Environmental Research, of the US Department of Energy (grant DOE-SC10010566), and by m-CAFEs (Microbial Community Analysis & Functional Evaluation in Soils), a project led by Lawrence Berkeley National Laboratory supported by the U.S. Department of Energy, Office of Science, Office of Biological & Environmental Research under contract number DE-AC02-05CH11231.

AUTHOR CONTRIBUTIONS

SD, AL, and KHW collected samples. SD, AL and JFB designed the study. SD, AL, AS, and JFB assembled and curated genomes. SD, AL, ACC, and JFB performed genome annotation, bioinformatic and statistical analysis. SD and ACC developed code. SD, PBMC, and JFB

interpreted metabolic annotation data and curated metabolic pathways. SD, AL, ACC, and JFB wrote the manuscript. All others read and approved the content.

COMPETING INTERESTS

The authors declare no competing interests

CORRESPONDING AUTHORS

Correspondence should be addressed to Spencer Diamond (sdiamond@berkeley.edu) and Jillian F. Banfield (jbanfield@berkeley.edu).

REFERENCES

1. Khadka, R. *et al.* Evolutionary History of Copper Membrane Monooxygenases. *Frontiers in microbiology* 9, W132-13 (2018).

2. Coleman, N. V. et al. Hydrocarbon monooxygenase in Mycobacterium: recombinant expression of a member of the ammonia monooxygenase superfamily. *The ISME Journal* 6, 171–182 (2011).

3. Lehtovirta-Morley, L. E. Ammonia oxidation: *Ecology*, physiology, biochemistry and why they must all come together. *FEMS Microbiology Letters* 365, 28 (2018).

4. Stahl, D. A. & Torre, J. R. de la. Physiology and Diversity of Ammonia-Oxidizing Archaea. *Annual Review of Microbiology* 66, 83–101 (2012).

5. Mancinelli, R. L. The Regulation of Methane Oxidation in Soil. *Annu Rev Microbiol* 49, 581–605 (1995).

6. Wu, L. et al. A Critical Review on Nitrous Oxide Production by Ammonia-Oxidizing Archaea. *Environmental Science & Technology* 54, 9175–9190 (2020).

7. Monteiro, M., Séneca, J. & Magalhães, C. The history of aerobic ammonia oxidizers: from the

822 first discoveries to today. *Journal of Microbiology* 52, 537–547 (2014).
823 8. Alves, R. J. E., Minh, B. Q., Urich, T., Haeseler, A. von & Schleper, C. Unifying the global
824 phylogeny and environmental distribution of ammonia-oxidising archaea based on amoA genes.
825 *Nature communications* 9, 66–17 (2018).
826 9. Knief, C. Diversity and Habitat Preferences of Cultivated and Uncultivated Aerobic
827 Methanotrophic Bacteria Evaluated Based on pmoA as Molecular Marker. Frontiers in
828 microbiology 6, 457–38 (2015).
829 10. Hatzenpichler, R. Diversity, Physiology, and Niche Differentiation of Ammonia-Oxidizing
830 Archaea. *Applied and Environmental Microbiology* 78, 7501–7510 (2012).
831 11. Leininger, S. et al. Archaea predominate among ammonia-oxidizing prokaryotes in soils.
832 Nature 442, 806–809 (2006).
833 12. Liu, S. et al. Ammonia-oxidizing archaea have better adaptability in oxygenated/hypoxic
834 alternant conditions compared to ammonia-oxidizing bacteria. *Applied Microbiology and*
835 *Biotechnology* 99, 8587–8596 (2015).
836 13. Tolar, B. B. et al. Integrated structural biology and molecular ecology of N-cycling enzymes
837 from ammonia-oxidizing archaea. *Environmental Microbiology Reports* 9, 484–491 (2017).
838 14. Bédard, C. & Knowles, R. Physiology, biochemistry, and specific inhibitors of CH4, NH4+,
839 and CO oxidation by methanotrophs and nitrifiers. *Microbiol Rev* 53, 68–84 (1989).
840 15. Burrows, K. J., Cornish, A., Scott, D. & Higgins, I. J. Substrate Specificities of the Soluble
841 and Particulate Methane Mono-oxygenases of Methylosinus trichosporium OB3b. *Microbiology+*
842 130, 3327–3333 (1984).
843 16. Qin, W. et al. Alternative strategies of nutrient acquisition and energy conservation map to
844 the biogeography of marine ammonia-oxidizing archaea. The ISME Journal 409, 507–15 (2020).
845 17. Lancaster, K. M., Caranto, J. D., Majer, S. H. & Smith, M. A. Alternative Bioenergy: Updates
846 to and Challenges in Nitrification Metalloenzymology. *Joule* 2, 421–441 (2018).

18. Vajrala, N. et al. Hydroxylamine as an intermediate in ammonia oxidation by globally abundant marine archaea. *Proceedings of the National Academy of Sciences* 110, 1006–1011 (2013).

19. Hosseinzadeh, P. et al. A Purple Cupredoxin from Nitrosopumilus maritimus Containing a Mononuclear Type 1 Copper Center with an Open Binding Site. *J Am Chem Soc* 138, 6324–6327 (2016).

20. Stein, L. Y. Insights into the physiology of ammonia-oxidizing microorganisms. *Current Opinion in Chemical Biology* 49, 9–15 (2019).

21. Carini, P., Dupont, C. L. & Santoro, A. E. Patterns of thaumarchaeal gene expression in culture and diverse marine environments. Environmental Microbiology 20, 2112–2124 (2018).

22. Pester, M. et al. amoA-based consensus phylogeny of ammonia-oxidizing archaea and deep sequencing of amoA genes from soils of four different geographic regions. Environmental Microbiology 14, 525–539 (2012).

23. Diamond, S. et al. Mediterranean grassland soil C-N compound turnover is dependent on rainfall and depth, and is mediated by genomically divergent microorganisms. *Nature Microbiology* 4, 1356–1367 (2019).

24. Sharrar, A. M. et al. Bacterial Secondary Metabolite Biosynthetic Potential in Soil Varies with Phylum, Depth, and Vegetation Type. *Mbio* 11, e00416-20 (2020).

25. Lavy, A. et al. Microbial communities across a hillslope-riparian transect shaped by proximity to the stream, groundwater table, and weathered bedrock. *Ecol Evol* 9, 6869–6900 (2019).

26. Anantharaman, K. et al. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nature communications 7, 13219 (2016).

27. Dong, X. et al. Metabolic potential of uncultured bacteria and archaea associated with petroleum seepage in deep-sea sediments. *Nat Commun* 10, 1816 (2019).

28. Meinhardt, K. A. et al. Evaluation of revised polymerase chain reaction primers for more

872 inclusive quantification of ammonia-oxidizing archaea and bacteria. *Env Microbiol Rep* 7, 354–
873 363 (2015).
874 29. Wang, V. C. C. et al. Alkane Oxidation: Methane Monooxygenases, Related Enzymes, and
875 Their Biomimetics. *Chemical Reviews* 117, 8574–8621 (2017).
876 30. Lee, M. D. GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics* 35, btz188
877 (2019).
878 31. Lawton, T. J., Ham, J., Sun, T. & Rosenzweig, A. C. Structural conservation of the B subunit
879 in the ammonia monooxygenase/particulate methane monooxygenase superfamily. *Proteins* 82,
880 2263–2267 (2014).
881 32. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a
882 metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* 35, 725–
883 731 (2017).
884 33. Méheust, R., Burstein, D., Castelle, C. J. & Banfield, J. F. The distinction of CPR bacteria
885 from other bacteria based on protein family content. Nat Commun 10, 4173 (2019).
886 34. Komori, H. & Higuchi, Y. Structure and molecular evolution of multicopper blue proteins.
887 *Biomol Concepts* 1, 31–40 (2010).
888 35. Gräff, M., Buchholz, P. C. F., Roes-Hill, M. L. & Pleiss, J. Multicopper oxidases: modular
889 structure, sequence space, and evolutionary relationships. *Proteins Struct Funct Bioinform* 88,
890 1329–1339 (2020).
891 36. Kobayashi, S. et al. Nitric Oxide Production from Nitrite Reduction and Hydroxylamine
892 Oxidation by Copper-containing Dissimilatory Nitrite Reductase (NirK) from the Aerobic
893 Ammonia-oxidizing Archaeon, Nitrososphaera viennensis. *Microbes Environment* 33, ME18058
894 (2018).
895 37. Reyes, C. et al. Genome wide transcriptomic analysis of the soil ammonia oxidizing
896 archaeon Nitrososphaera viennensis upon exposure to copper limitation. The ISME Journal 1–

897 16 (2020) doi:10.1038/s41396-020-0715-2.
898 38. Kitzinger, K. et al. Cyanate and urea are substrates for nitrification by Thaumarchaeota in
899 the marine environment. Nature Microbiology 1–12 (2019) doi:10.1038/s41564-018-0316-2.
900 39. Bräsen, C., Esser, D., Rauch, B. & Siebers, B. Carbohydrate Metabolism in Archaea:
901 Current Insights into Unusual Enzymes and Pathways and Their Regulation. *Microbiol Mol Biol*
902 *R* 78, 89–175 (2014).
903 40. Tjaden, B., Plagens, A., Dörr, C., Siebers, B. & Hensel, R. Phosphoenolpyruvate synthetase
904 and pyruvate, phosphate dikinase of Thermoproteus tenax: key pieces in the puzzle of archaeal
905 carbohydrate metabolism. *Mol Microbiol* 60, 287–298 (2006).
906 41. Chadwick, G. L., Hemp, J., Fischer, W. W. & Orphan, V. J. Convergent evolution of unusual
907 complex I homologs with increased proton pumping capacity: energetic and ecological
908 implications. *Isme J* 12, 2668–2680 (2018).
909 42. Welte, C. & Deppenmeier, U. Membrane-Bound Electron Transport in Methanosaeta
910 thermophila. *J Bacteriol* 193, 2868–2870 (2011).
911 43. Bartossek, R., Nicol, G. W., Lanzen, A., Klenk, H.-P. & Schleper, C. Homologues of nitrite
912 reductases in ammonia-oxidizing archaea: diversity and genomic context. Environmental
913 Microbiology 12, 1075–1088 (2010).
914 44. Butterfield, C. N. et al. Proteogenomic analyses indicate bacterial methylotrophy and
915 archaeal heterotrophy are prevalent below the grass root zone. *PeerJ* 4, e2687-28 (2016).
916 45. Yabusaki, S. B. et al. Water Table Dynamics and Biogeochemical Cycling in a Shallow,
917 Variably-Saturated Floodplain. *Environ Sci Technol* 51, 3307–3317 (2017).
918 46. Ward, L. M., Johnston, D. T. & Shih, P. M. Phanerozoic radiation of ammonia oxidizing
919 bacteria. *Sci Rep-uk* 11, 2070 (2021).
920 47. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Computational Biology* 7, e1002195-
921 16 (2011).

48. Parks, D. H. et al. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nature Biotechnology* 1–12 (2020) doi:10.1038/s41587-020-0501-8.

49. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics 1–3 (2019) doi:10.1093/bioinformatics/btz848.

50. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. 11, 2864–2868 (2017).

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

52. Nguyen, L.-T., Schmidt, H. A., Haeseler, A. von & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular biology and evolution* 32, 268–274 (2014).

53. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* 32, 1792–1797 (2004).

54. 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).

55. Boyd, J. A., Woodcroft, B. J. & Tyson, G. W. GraftM: a tool for scalable, phylogenetically informed classification of genes within metagenomes. *Nucleic Acids Res* 46, gky174- (2018).

56. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer Science & Business Media, 2009). doi:10.1007/978-0-387-98141-3.

57. Quinn, T. P., Richardson, M. F., Lovell, D. & Crowley, T. M. propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis. Sci Rep-uk 7, 16252 (2017).

947 58. Quinn, T. P. et al. A field guide for the compositional analysis of any-omics data.
948 *Gigascience* 8, (2019).
949 59. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site
950 identification. *BMC Bioinformatics* 11, 119–11 (2010).
951 60. Aramaki, T. et al. KofamKOALA: KEGG ortholog assignment based on profile HMM and
952 adaptive score threshold. Bioinformatics 36, 2251–2252 (2019).
953 61. Huerta-Cepas, J. et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically
954 annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res
955 47, gky1085 (2018).
956 62. Mistry, J. et al. Pfam: The protein families database in 2021. Nucleic Acids Res 49,
957 gkaa913- (2020).
958 63. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the
959 analysis of massive data sets. Nat Biotechnol 35, 1026–1028 (2017).
960 64. Steinegger, M. et al. HH-suite3 for fast remote homology detection and deep protein
961 annotation. *Bmc Bioinformatics* 20, 473 (2019).
962 65. Enright, A. J., Dongen, S. V. & Ouzounis, C. A. An efficient algorithm for large-scale
963 detection of protein families. Nucleic Acids Res 30, 1575–1584 (2002).
964 66. Helen, D., Kim, H., Tytgat, B. & Anne, W. Highly diverse nirK genes comprise two major
965 clades that harbour ammonium- producing denitrifiers. *BMC Genomics* 17, 1–13 (2016).
966 67. Nakamura, K. & Go, N. Function and molecular evolution of multicopper blue proteins.
967 *Cellular and Molecular Life Sciences* 62, 2050–2066 (2005).
968 68. Cáceres, M. D. & Legendre, P. Associations between species and groups of sites: indices
969 and statistical inference. Ecology 90, 3566–3574 (2009).