

“ARMAN” archaea depend on association with euryarchaeal host in culture and in situ

Golyshina, Olga; Toshchakov, Stepan V.; Makarova, Kira S.; Gavrillov, Sergei N.; Korzhenkov, Alexei A.; La Cono, Violetta ; Arcadi, Erika; Nechitaylo, Taras Y.; Ferrer, Manuel; Kublanov, Ilya V.; Wolf, Yuri I.; Yakimov, Michail M.; Golyshin, Peter

**Nature Communications**

DOI:  
[10.1038/s41467-017-00104-7](https://doi.org/10.1038/s41467-017-00104-7)

Published: 01/01/2017

Peer reviewed version

[Cyswllt i'r cyhoeddiad / Link to publication](#)

*Dyfyniad o'r fersiwn a gyhoeddwyd / Citation for published version (APA):*  
Golyshina, O., Toshchakov, S. V., Makarova, K. S., Gavrillov, S. N., Korzhenkov, A. A., La Cono, V., Arcadi, E., Nechitaylo, T. Y., Ferrer, M., Kublanov, I. V., Wolf, Y. I., Yakimov, M. M., & Golyshin, P. (2017). “ARMAN” archaea depend on association with euryarchaeal host in culture and in situ. *Nature Communications*, 8, Article 60. <https://doi.org/10.1038/s41467-017-00104-7>

#### **Hawliau Cyffredinol / General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

#### **Take down policy**

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

1 “ARMAN” archaea depend on association with euryarchaeal host in culture and *in situ*.

2

3 Olga V. Golyshina<sup>1</sup>, Stepan V. Toshchakov<sup>2</sup>, Kira S. Makarova<sup>3</sup>, Sergey N. Gavrilov<sup>4</sup>, Aleksei  
4 A. Korzhenkov<sup>2</sup>, Violetta La Cono<sup>5</sup>, Erika Arcadi<sup>5</sup>, Taras Y. Nechitaylo<sup>6</sup>, Manuel Ferrer<sup>7</sup>, Ilya  
5 V. Kublanov<sup>2,4</sup>, Yuri I. Wolf<sup>3</sup>, Michail M. Yakimov<sup>2,5</sup> & Peter N. Golyshin<sup>1</sup>

6 <sup>1</sup>School of Biological Sciences, Bangor University, Deiniol Rd, Bangor, LL57 2UW, UK

7 <sup>2</sup>Immanuel Kant Baltic Federal University, Kaliningrad, 236040, Russia

8 <sup>3</sup>National Center for Biotechnology Information, National Library of Medicine – National  
9 Institutes of Health, Bethesda, MD, 20894, USA

10 <sup>4</sup>Winogradsky Institute of Microbiology, Research Center for Biotechnology Russian  
11 Academy of Sciences, Prospect 60-Letiya Oktyabrya 7/2, Moscow, 117312, Russia

12 <sup>5</sup>Institute for Coastal Marine Environment, CNR, Spianata S. Raineri 86, 98122 Messina,  
13 Italy

14 <sup>6</sup>Insect Symbiosis Research Group, Max Planck Institute for Chemical Ecology, Hans-  
15 Knöll-Strasse 8, Jena, 07745, Germany

16 <sup>7</sup>Institute of Catalysis CSIC, Campus Cantoblanco, 28049 Madrid, Spain

17

18 Correspondence and requests for materials should be addressed to O.V.G. (email:

19 o.golyshina@bangor.ac.uk).

20

21

22

23

24

25 **Abstract**

26 Intriguing, yet uncultured ‘ARMAN’-like archaea are metabolically dependent on other  
27 members of the microbial community. It remains uncertain though which hosts they rely upon,  
28 and, because of the lack of complete genomes, to what extent. Here, we report the co-culturing  
29 of ARMAN-2-related organism, Mia14, with *Cuniculiplasma divulgatum* PM4 during the  
30 isolation of this strain from acidic streamer in Parys Mountain (Isle of Anglesey, UK). Mia14  
31 is highly enriched in the binary culture (ca. 10% genomic reads) and its ungapped 0.95 Mbp  
32 genome points at severe voids in central metabolic pathways, indicating dependence on the  
33 host, *C. divulgatum* PM4. Analysis of *C. divulgatum* isolates from different sites and shotgun  
34 sequence data of Parys Mountain samples suggests an extensive genetic exchange between  
35 Mia14 and hosts *in situ*. Within the subset of organisms with high-quality genomic assemblies  
36 representing ‘DPANN’ superphylum, the Mia14 lineage has had the largest gene flux, with  
37 dozens of genes gained that are implicated in the host interaction.

38

39

## 40 Introduction

41 Deep metagenomic analysis of environmental samples from acidic environments across our  
42 planet has demonstrated the existence of previously neglected uncultured archaea that are only  
43 very distantly related to recognised phyla<sup>1</sup>. Initially detected at Iron Mountain (California,  
44 USA), these archaeal lineages were subsequently confirmed to occur in various acid mine  
45 drainage (AMD) systems<sup>2</sup>. This enigmatic group of archaea (the so-called ‘Archaeal  
46 Richmond Mine Acidophilic Nano-organisms’, or ‘ARMAN’ was initially found in the  
47 fraction of cells filtered through 0.22 µm membrane filters<sup>1</sup>. Metagenomic assemblies  
48 suggested average genome sizes of these organisms to be relatively small for free-living  
49 organisms (approx. 1 Mbp)<sup>1</sup>. An interesting observation documented by electron microscopy  
50 was that some cells of a small size (<500 nm) interact through pili-like structures with larger  
51 cells that lacked cell walls. Comolli and colleagues<sup>3</sup> suggested the ‘ARMAN’ organisms were  
52 the ‘small’ cells, whereas cell wall-deficient larger cells were attributed to some members of  
53 the order *Thermoplasmatales*, a group of organisms known to be widely represented in acid  
54 mine drainage systems<sup>4</sup>. Emerging findings from metagenomic datasets of ARMAN-like  
55 archaea and especially their ubiquity suggest that this group plays important roles in the  
56 environment, although the exact roles have yet to be established<sup>2</sup>. The phylogenomic  
57 placement of archaea from this group still represents a matter for discussion<sup>5, 6, 7</sup>.

58 The known example of small-sized cultured archaea is represented by *Nanoarchaeum*  
59 *equitans*, currently the only validly described member of the phylum *Nanoarchaeota*. Cells  
60 are about 500 nm (or smaller) in diameter and exhibit a typical archaeal ultrastructure<sup>8</sup>.  
61 *Nanoarchaeum equitans* exists only in association with the host, *Ignicoccus hospitalis*, which  
62 supplies certain organic compounds (lipids and amino acids), growth factors and likely ATP  
63 to *N. equitans*<sup>9</sup>. Other nanoarchaeota-related examples include an Nst1 archaeon forming an  
64 association with its host, the *Sulfolobales*-related organism<sup>10</sup>, and “*Candidatus Nanopusillus*  
65 *acidilobi*”, thriving in a partnership with *Acidilobus* spp.<sup>11</sup>. These nanoarchaeota are  
66 hyperthermophilic marine and terrestrial organisms with extremely compact genomes that  
67 likely are not of an ancestral nature, but rather probably resulted from massive gene loss<sup>6</sup>.  
68 *Nanoarchaeota*-related organisms (including those known only by metagenomics-resolved  
69 genomes) are phylogenetically clustered within the ‘DPANN’ candidate superphylum  
70 (abbreviated after candidate divisions “*Diapherotrites*”, “*Parvarchaeota*”,

71 “*Aenigmarchaeota*”, “*Nanohalarchaeota*” and the only validly described phylum  
72 *Nanoarchaeota*)<sup>12</sup>. Recently, a number of uncultured ‘DPANN’ archaea with almost complete  
73 genomes were predicted by Castelle and co-authors<sup>13</sup> to be symbiotic and/or to have a lifestyle  
74 based on fermentation. To summarise, all experimentally validated examples of interactions  
75 between co-cultured small (or ‘nanosized’) archaea and their partners are limited to  
76 *Crenarchaea* being the hosts. All of them (except *Ignicoccus* sp.) are acidophiles, while so far  
77 no associations have been co-cultured or characterised for *Euryarchaeota*, except those from  
78 the recent report on a four-member consortium containing a fungus, two strains of  
79 *Thermoplasmatales* and ARMAN-1-related organism with, due to the complexity of this  
80 enrichment culture, only a partially sequenced genome<sup>14</sup>.

81 Here we report the co-cultivation and analysis of the ungapped genome of an ARMAN-  
82 like organism, the “*Candidatus Mancarchaeum acidiphilum*” Mia14, which was enriched in  
83 the laboratory binary culture with *Cuniculiplasma divulgatum* PM4, a recently described  
84 representative of the family *Cuniculiplasmataceae* within *Thermoplasmata*<sup>15</sup>. After additional  
85 sampling campaigns and *de novo* metagenome sequencing of the microbial community of the  
86 acidic streamer of Mynydd Parys/Parys Mountain, we revealed possible *in situ* interactions of  
87 these organisms with other microbial community members. Furthermore, we analysed the  
88 voids in its metabolic pathways (and thus dependencies on potential hosts) and mapped its  
89 phylogenetic position. Finally, using data on arCOGs gains and losses, we reconstructed its  
90 evolutionary trajectory starting from the last archaeal common ancestor (LACA), which  
91 pointed at Mia14 having the greatest known extent of gene fluxes within the “DPANN”  
92 superphylum.

93

## 94 **Results**

95 **Coexistence of Mia14 with *Cuniculiplasma divulgatum* PM4.** We have previously isolated  
96 and described two strains of a new archaeal family, genus and species, named *Cuniculiplasma*  
97 *divulgatum* (order *Thermoplasmatales*), from acidic streamers at Parys Mountain (UK) and  
98 Cantareras mine (Spain)<sup>15</sup>. Both strains S5 and PM4 were characterised as acidophilic  
99 organoheterotrophes with mesophilic optima for growth and as facultative anaerobes<sup>15</sup>. The  
100 genomes of these isolates were remarkably similar to one another (>98% average nucleotide  
101 identity (ANI)<sup>16</sup> and to that of the genomic assembly ‘G-plasma’ from Iron Mountain

102 (USA))<sup>17</sup>. During the isolation, *C. divulgatum* strain PM4 was co-cultured for two years with  
103 another archaeon designated Mia14 with a proportion of genomic reads, PM4:Mia14 of  
104 approx. 10:1. The initially poor growth of the PM4 component was significantly improved by  
105 the addition of complex organic compounds, such as beef extract and trypton (0.1% w/vol).  
106 However, the enhanced growth of *C. divulgatum* and an increased frequency of re-inoculations  
107 had a dramatic effect on the growth of Mia14, which was eliminated from the culture and, after  
108 approximately 2.5 years of regular (every 20-22 days) passages into the fresh medium, was  
109 not detectable by PCR with specific primers. Another possible explanation is that the faster  
110 growth of *C. divulgatum* strain PM4 was the result of the elimination of Mia14, which may  
111 have negatively affected the growth of PM4 in earlier cultivation stages. Whatever the case,  
112 we could not maintain Mia14 for longer than 2.5 years. However, as Mia14 was highly  
113 enriched in the initial enrichment cultures with *C. divulgatum* strain PM4, we obtained enough  
114 coverage of its genomic reads (approximately 40 fold) to assemble a single chromosome. After  
115 the loss of Mia14 from the enrichment culture, we performed additional sampling of the acidic  
116 streamer (from the same site in Parys Mountain where the isolate PM4 was derived from), and  
117 detected Mia14 initially by PCR using specific primers, then by the *de novo* sequencing of  
118 environmental DNA, and ultimately, by CARD-FISH.

119 Analysis of metagenomic contigs showed that the most abundant group (up to 57%)  
120 was *Thermoplasmatales*-related archaea. Small-genome archaeal lineages (“*Candidatus*  
121 *Parvarchaeota*” and “*Ca. Micrarchaeota*”) were also detected at 0.31% and 3.84%, respectively  
122 (Fig. 1).

123 Interestingly, both median contig coverage and coverage-based abundance calculation  
124 indicate that the amount of *Cuniculiplasma* cells in the Parys Mountain acidic streamer is  
125 nearly equal to the amount of “*Candidatus Mancarchaeum*” cells (Figs 1 and 2). Also, analysis  
126 of read coverage vs. GC content of metagenomic contigs reveals that “*Ca. Mancarchaeum*”  
127 and *C. divulgatum*-related contigs form a very compact cluster similar both in coverage and  
128 GC content (Fig. 2). Notably, we also observed another *Thermoplasmatales*, “*Ca.*  
129 *Micrarchaeota*” contig cluster, in the Parys Mountain metagenome, suggesting that there are  
130 several stable two-member microbial associates in the community at this site.

131 **Fluorescence microscopy shows interaction of Mia14 and the host**

132 Microbial cells from enrichment cultures set up with the environmental sample from 2014  
133 were either hybridised with probe EUB338(I-III) mix or probe ARCH915 to target Bacteria or  
134 Archaea, respectively. The following CARD–FISH analysis revealed dense populations of  
135 Archaea and the almost complete absence of bacterial cells. Pleomorphic morphologies of cells  
136 of various size, typical for *Cuniculiplasma/Thermoplasmatales*<sup>16</sup>, were confirmed with  
137 hybridizations with *Cuniculiplasma*-specific probe Clpm-1100R. Besides bright signals, the  
138 CARD–FISH microphotography retrieved numerous debris-like structural forms, which likely  
139 could be referred to as either dying or metabolically dormant cells. This observation is typical  
140 for both natural samples and initial enrichments, where the cells of different metabolic states  
141 coexist. Noteworthily, parallel hybridizations with “*Ca. Mancarchaeum*”-specific probe  
142 ARM-MIA1469R and *Thermoplasmata*-specific probe Thpmt680R showed quite similar  
143 images (Fig. 3), suggesting that the organisms live in a tight association. Cross-hybridization  
144 of ARM-MIA1469R probe with pure *Cuniculiplasma* culture was controlled at specific  
145 hybridization conditions and no positive signals were retrieved. Side-by-side comparisons of  
146 “*Ca. Mancarchaeum*” versus *Cuniculiplasma* cells revealed that the former are slightly smaller  
147 in size and only a minor fraction of cells do not overlap in each frame. Detailed view of some  
148 double-hybridized cell formations revealed single coccoid-shaped *Cuniculiplasma* cells were  
149 surrounded by ARM-MIA1469R probe-labeled organisms (Fig. 3 (c1)).

150 **Phylogenetic position of Mia14 and related organisms** Based on 16S rRNA gene sequence,  
151 Mia14 was found to be only distantly related to organisms with established taxonomic status.  
152 Less than 75 % SSU rRNA gene sequence identities with the thaumarchaeon *Nitrososphaera*  
153 *viennensis* and euryarchaeon *Methanosaeta consilii* are observed. Among candidate status  
154 holders (Supplementary Figure 1), the nearest relative was inferred to be ARMAN-2  
155 (“*Candidatus* Micrarchaeum acidiphilum”<sup>1,5</sup>) originally detected in acidic environments and  
156 sharing 92% 16S rRNA sequence identity with Mia14. Other similar sequences (92% sequence  
157 identity) belong to PCR-amplified and cloned SSU rRNA genes from fumarolic thermal and  
158 acidic green biofilms, Mexico, Michoacan, Los Azufres (KJ907762). Interestingly, both above  
159 sequences and the sequence of Mia14 possess introns in their 16S rRNA genes. In addition,  
160 sequences with a lower sequence identity and coverage (91%, 58%) were detected in a PCR-  
161 amplified SSU rRNA clone from Rio Tinto (FN865418)<sup>18</sup>, acidic hot springs (JF280243; 91%,  
162 58%)<sup>19</sup> and a number of other acid mine drainage and volcanic environments. Furthermore,

163 similar sequences have been retrieved from southern Appalachian peatlands (PF82012)<sup>20</sup> and  
164 wetlands in Finland (AM905392, AM905420)<sup>21</sup>. The two latter sites were oligotrophic, with  
165 temperatures in the range from 0 to 15 °C and slightly acidic pH (4-5.6 and 3.9-4.3,  
166 respectively). Along with wetland clones, similar signatures (BioProject PRJNA279923) have  
167 been found in metagenomic data from another oligotrophic environment, the pH-neutral  
168 groundwater from Fennoscandian terrestrial deep biosphere<sup>22</sup>. All these records suggest a wide  
169 distribution of organisms similar to Mia14 and ARMAN-2 in natural settings with various pH  
170 characteristics, not necessarily tied to acidic environments.

171 The placement of Mia14 on the phylogenetic tree constructed with concatenated  
172 ribosomal proteins is presented in Fig. 4. In agreement with previous observations<sup>12</sup>, the  
173 position of Mia14 within the 'DPANN' superphylum is strongly supported.

#### 174 **Genome statistics**

175 The genome of Mia14 is a single, circular chromosome with 952,257 bp, with the molar G +C  
176 % of 39.36% (Fig. 5). The coding density in the genome is of 1.032 genes per kbp (968 bases  
177 per gene). About ~150-200 hypothetical proteins were present. The genome encodes 45  
178 tRNAs. Three introns were detected across the chromosome. All these traits are typical for  
179 small archaeal genomes, e.g. in *Nanoarchaeum equitans* (491 kbp)<sup>23</sup>, "*Candidatus*  
180 *Nanobsidianus stetterii*", Nst1 belonging to the phylum *Nanoarchaeota* (592 kbp)<sup>10</sup>, ARMAN-  
181 2 (~1 Mbp)<sup>5</sup> and other host-associated or symbiotic microorganisms.

#### 182 **Lateral gene transfer between *Thermoplasmatales* and Mia14**

183 Comparative analysis of *in silico* proteomes of Mia14 and strains S5 and PM4 with  
184 ProteinOrtho<sup>24</sup> revealed several clusters of orthologous genes shared between Mia14 and *C.*  
185 *divulgatum* S5, but absent in *C. divulgatum* PM4 (Supplementary Data 1, Supplementary Fig.  
186 2). These genes encode several membrane-associated proteins (MIA14\_0876, \_0886, \_0893  
187 and \_0478), two SAM-dependent methyltransferases (MIA14\_0883 and \_0885), sulfocyanin  
188 (MIA14\_0884) and peroxiredoxin (MIA14\_0479). It should be noted that the majority of these  
189 proteins have homologues in PM4, but are more distant to those from both Mia14 and S5 and  
190 have different gene context. Few Mia14 genes from these clusters have no homologues in  
191 PM4.

192 Analysis with IslandViewer<sup>325</sup> showed that altogether five genomic islands (GIs) are  
193 present; the largest GI contains 41 genes and spans 36.5 kbp (Fig. 5). A closer inspection of

194 this island reveals that the integration occurred in the gene for zinc-binding pyruvate-formate  
195 lyase-activating enzyme (MIA14\_0850), splitting it in two parts: MIA14\_0850 and  
196 MIA14\_0891, with the latter located in the immediate vicinity of 23S rRNA gene. About 50%  
197 of genes within this GI could not be assigned to known arCOGs and represent small proteins  
198 that often contain transmembrane segments, which is typical for archaeal ‘dark matter’. In  
199 turn, the genes assigned to arCOGs (i.e., MIA14\_0898, the DNA invertase Pin homolog,  
200 MIA14\_0894 (similar to those from other *Thermoplasmatales*), ParA family chromosome  
201 partitioning ATPase and MIA14\_0890, integrase of XerD family) were shown to be strongly  
202 associated with “dark matter” islands in archaeal genomes and could be specifically attributed  
203 to integrated mobile elements<sup>26</sup>.

204         Among GI-associated genes, we also found cation transport ATPase /copper-  
205 transporting P-type ATPase (MIA14\_0877), which may have significance for the fitness of  
206 this organism in the harsh conditions of Parys Mountain AMD. Phylogenetic analysis of this  
207 ATPase showed that its close homologues are widely distributed among acidophilic  
208 *Thermoplasmatales*. At the same time, the copper-transporting ATPase of ARMAN-2 seems  
209 only quite distantly related to MIA14\_0877 (Fig. 6, Supplementary Table 1). Gene  
210 neighborhood of MIA14\_0877 included an Lrp-AsnC family transcriptional regulator and a  
211 copper chaperone, resembling functional copper fitness islands described for “*Ferroplasma*  
212 *acidarmanus*”<sup>27</sup>. This gene cluster was found to be conserved in *Cuniculiplasma*-related  
213 archaea. Furthermore, *C. divulgatum* S5 genome possessed two copies of this copper-fitness  
214 island (Fig. 6). Interestingly, one of the *C. divulgatum* S5 copper fitness islands was adjacent  
215 to genetic loci for SHOCT family and DUF 302 family proteins as in the Mia14 copper gene  
216 cluster, while another *C. divulgatum* S5 copper gene island had a high level of gene synteny  
217 with *C. divulgatum* PM4 (Figs 5 and 6). The above observation supports the lateral gene  
218 transfer from ancestral *Cuniculiplasma*-related lineage(s) to Mia14. In that case, it is more  
219 likely that LCA of *Cuniculiplasma* had two copies of this gene cluster, one of which was lost  
220 during the evolution of *C. divulgatum* PM4 and ‘G-plasma’.

221         Detailed analysis of *de novo* metagenome sequencing data from Parys Mountain  
222 samples shows that gene clusters similar to the abovementioned copper fitness island of Mia14  
223 are widely present in different metagenomic contigs (Supplementary Data 2). It indicates that

224 this highly mobile gene set is important for heavy metal resistance in microbial communities  
225 inhabiting acidic environments with high concentrations of dissolved metal ions.

226 Other smaller GIs of Mia14 (Fig. 5) contain defence systems (toxin/antitoxin and type  
227 III restriction-modification proteins), 2-oxoacid dehydrogenase multienzyme complexes, 2-  
228 oxoacid decarboxylase (E1) component subunits  $\alpha$  and  $\beta$ , glycosyltransferases and numerous  
229 hypothetical proteins. Interestingly, the laminin G-encoding gene locus is also situated on the  
230 GI (see the section 'Secretion systems').

231

### 232 **Carbohydrate metabolism**

233 The Mia14 genome has no genes for central carbohydrate metabolism pathways such as  
234 glycolysis and gluconeogenesis, pentose phosphate pathway or TCA cycle. A detailed manual  
235 inspection suggested that the genome encodes a complete set of enzymes for glucose oxidation  
236 via the non-phosphorylating Entner-Doudoroff pathway<sup>28</sup>: a glucose dehydrogenase  
237 (MIA14\_0575), D-gluconate dehydratase (MIA14\_0298), 2-dehydro-3-phosphogluconate  
238 aldolase (MIA14\_0299) and NAD-dependent D-glyceraldehyde dehydrogenase  
239 (MIA14\_0297). Surprisingly, no enzymes for further conversion of glycerate, e.g. to glycerate-  
240 2-phosphate or glycerate-3-phosphate were found. The pyruvate released during the action of  
241 2-dehydro-3-phosphogluconate aldolase, could be carboxylated to malate and oxaloacetate in  
242 the reaction catalysed by NAD-dependent malic enzyme (MIA14\_0243/EC 1.1.1.38).  
243 Characterised homolog of *scfA* from *E. coli* (or *maeA*<sup>29</sup>) was reversible despite the  
244 carboxylation reaction being 28 times slower than the forward reaction. Other enzymes found  
245 to catalyse pyruvate conversions are phosphoenolpyruvate synthase/pyruvate phosphate  
246 dikinase (MIA14\_0437, EC 2.7.9.2 and MIA14\_0462) and pyruvate kinase (MIA14\_0326, EC  
247 2.7.1.40). It is worth mentioning that ARMAN-2, one of the most closely related organisms  
248 to Mia14 among those with partially sequenced genomes, exhibits a relatively scarce repertoire  
249 of genes in comparison to sibling lineages ARMAN-4 and -5. Only a few genes for glycolysis  
250 in ARMAN-2 and a near-complete set of genes in ARMAN-4 and -5 were predicted. TCA,  
251 which is dysfunctional in Mia14, was reported to be complete or almost complete in all  
252 ARMAN cluster organisms mentioned above<sup>5</sup>. Furthermore, central metabolic pathways in  
253 Mia14 starkly contrast with AR10 assembly representing "*Ca. Diapherotrites*", but to some  
254 extent resemble those predicted for a more phylogenetically distant AR20 ("*Ca.*

255 Woesarchaeota”)<sup>13</sup>. Furthermore, the inspection of amino acid biosynthetic pathways in Mia14  
256 found them to be either incomplete or entirely missing. However, the total number of proteins  
257 in this functional category is higher in comparison to *N. equitans* and Nst1 (ref. <sup>10,23</sup>).

### 258 **Cofactors, vitamins, prosthetic groups and pigments**

259 No genes for coenzyme A, folate, lipoic acid, NAD and NADP cofactor, pyridoxin (Vitamin  
260 B6), heme and siroheme, thiamin biosynthesis and riboflavin, FMN and FAD metabolism were  
261 present in the entire genome of Mia14. The lack of functional pathways for cofactors and  
262 amino acids is quite characteristic for organisms with reduced genomes<sup>10,23</sup>.

### 263 **Protein metabolism**

264 Protein processing and modification-related genes (G3E family of P-loop GTPases, peptide  
265 methionine sulfoxide reductase and Rio family of protein kinases, amino- and carboxy-  
266 terminal intein-mediated trans-splice, and ribonucleotide reductase of class III (anaerobic),  
267 large subunit (EC 1.17.4.2) were missing in Mia14. Altogether, we have identified 35 large-  
268 and 26 small-subunits of ribosomal proteins; L37E (arCOG04126), S17e (arCOG01885), and  
269 S27e (arCOG04108) were absent. All three and, correspondingly, two former proteins were  
270 found in *N. equitans* and Nst1 (Supplementary Data 3).

### 271 **Secretion systems**

272 We have identified a number of genes affiliated with secretion processes in the genome of  
273 Mia14 (Supplementary Table 2). Two distinct type IV pili systems are present in the genome:  
274 one belongs to *Methanococci/Methanothermobacteria/Thermococci* group (MIA14\_0170 -  
275 \_0177) and another is more similar to a euryarchaeal group (MIA14\_0252- \_0260)<sup>30</sup>. No  
276 archaeum-related genes were found, in agreement with the loss of motility in most of  
277 ‘DPANN’ species. Only one FlaK-like prepilin peptidase (MIA14\_0570) was found. Key  
278 components of both systems are shared by many “*Ca. Micrarchaea*” species. Additionally, the  
279 genome encodes Sec translocon genes for preprotein translocase subunits SecYE  
280 (MIA14\_0832, \_0132) and Sec61beta (MIA14\_0736), SecDF (MIA14\_0121 and \_0122),  
281 signal peptide peptidase and signal recognition particle subunits and receptors. The presence  
282 of Sec-independent Tat pathway genes, suggests this system is operational for secretion of  
283 folded proteins.

284 The Mia14 surface layer deserves special attention. Besides a protection function in  
285 archaea, this compartment often regulates both cell adhesion and cell-cell interaction. We

286 identified at least eight different proteins that eventually account for the architecture of the  
287 surface layer. It contains strain-specific secreted proteins with polycystic kidney disease  
288 (PKD) superfamily fold and  $\beta$ -propeller repeat domains fused to CARDB (cell adhesion  
289 related domain found in bacteria)-like adhesion module. Proteins of the  $\beta$ -propeller fold are  
290 ubiquitous in nature and widely used as structural scaffolds for ligand binding and enzymatic  
291 activity. This fold comprises between four and twelve four-stranded  $\beta$ -meanders, the so-called  
292 blades that are arranged circularly around a central funnel-shaped pore.

293 Another observation is the expansion of genes encoding jellyroll fold LamG-like  
294 proteins in the Mia14 genome, which are only distantly similar to other LamG proteins from  
295 archaea with *Candidatus* status and from bacteria, but generally abundant in 'DPANN'  
296 superphylum species. This finding suggests an association of these proteins with laminin  
297 (glycoprotein)-containing extracellular matrix and their key role in host cell interactions. Some  
298 of them are linked to aforementioned type IV pili loci and are localised in GI (Fig. 5).

### 299 **Respiration**

300 The Mia14 genome encodes all typical subunits K, E, C, F, A, B, D, G, H and I of V/A Na<sup>+</sup>-  
301 and H<sup>+</sup>-transporting type ATP synthases, in this particular order (MIA14\_0355-0364),  
302 whereas the genome of *N. equitans* encodes only five subunits of ATP synthase<sup>23</sup>. The analysis  
303 of conserved motifs supported H<sup>+</sup>-translocating V-type ATPase<sup>31</sup>.

304 All genes coding for cytochrome *bd* quinol oxidase were identified. Subunits I and II  
305 are encoded by MIA14\_0653-0654. This type of oxidoreductase could generate proton motive  
306 force (PMF) by transmembrane charge separation, but do so without being a "proton pump".  
307 The main electron acceptor for them is oxygen, but cytochrome *bd* oxidoreductases are usually  
308 induced in response to low oxygen concentrations and serve for oxygen detoxification<sup>32</sup>. The  
309 role of the cytochrome *bd* oxidoreductase in Mia14 is puzzling, as the organism completely  
310 lacks any genes for biosynthesis of isoprenoid quinones, which are the only electron donors  
311 for this electrogenic enzyme complex. Moreover, no genes were identified coding for electron  
312 donating type I NADH dehydrogenase or succinate dehydrogenase or other known respiratory  
313 complexes (III and IV).

### 314 **Transporters**

315 ABC transporters, amino acid permeases, Major Facilitator Superfamily and others have been  
316 predicted in Mia14 (Supplementary Table 3) to notably outnumber those in nanoarchaeal  
317 genomes<sup>11</sup>.

318 **Evolutionary patterns.** Overall, compared to other 'DPANN' group members, the Mia14  
319 genome experienced an unusually high level of gene flux (Fig. 7). In addition to the 226  
320 genes that do not belong to known arCOGs (a large fraction of such genes was probably also  
321 acquired at the terminal branches of the 'DPANN' tree), Mia14 has lost determinants for  
322 over 400 arCOG families from the genome of its common ancestor with "*Ca.*  
323 *Ianarchaeum*"/AR10 lineage (46% of the ancestral set), but also gained over 130 arCOGs  
324 (21% of its current arCOG complement). Gains and losses of comparable scale exist within  
325 the 'DPANN' group tree (e.g. the loss of 49% of the ancestral genome in the lineage of  
326 AR17 or acquisition of 18% of the gene complement in the lineage of G17-L22-OTU1), but  
327 not on the same tree branch. Gene gains and losses seem to affect all functional groups  
328 equally with a notable exception of the "Cell motility" group where more gains than losses  
329 were predicted (Fig. 7). This functional group includes components of secretion systems,  
330 which might play a key role in interaction of Mia14 with its host. Moreover, as mentioned  
331 before many of the unique genes in Mia14 belong to GI, many of which encode membrane  
332 proteins and are associated with potential conjugative elements which might be involved in  
333 the extensive gene exchange between Mia14 and its host.

334 Analysing the trajectory of evolution of Mia14 from LACA through the prism of  
335 losses and gains of functional genes, a few interesting facts became apparent. Genes for the  
336 majority of enzymes of the TCA cycle were already lost during the transition from LACA to  
337 the 'DPANN' ancestor, together with many genes involved with amino acid, vitamin and  
338 cofactor biosynthesis along with the CRISPR-Cas system. Glycolysis and gluconeogenesis  
339 were present in all ascendants of Mia14 (i.e. in LACA, 'DPANN'- and Mia14-"*Ca.*  
340 *Nanohaloarchaeota*"-ancestors). However, many genes of these pathways were lost en route  
341 to the extant Mia14 species. Pathways for pyrimidine and purine biosynthesis and salvage  
342 were also lost at the very last (and long) step of evolution from Mia14/"*Ca.* *Ianarchaeum*"  
343 ancestor to the modern Mia14, with 414 genes lost and only 131 gained (Fig. 7A and  
344 Supplementary Data 4).

345 Analysis of the taxonomic affiliations of the ‘DPANN’ group species (Supplementary  
346 Data 5 and 6) shows that, in contrast with the other group members, the genome of Mia14 was  
347 and continues to be involved in extensive gene exchange with the *Thermoplasmata* lineages.  
348 Unsurprisingly, the most common source of the acquired genes is identified as *Cuniculiplasma*  
349 *divulgatum*, the Mia14 host.

350

### 351 **Etymology**

352 “*Candidatus Mancarchaeum acidiphilum*”

353 *Mancarchaeum* (*Manc.archaeum* M.L. mancus (adj.) crippled, maimed, referred to the  
354 absence of many pathways in the genome; N.L. neut. n. *archaeum* (from Gr. adj. *archaios* -ê  
355 -on, ancient), ancient one, archaeon; N.L. neut. n. Mancarchaeum, an archaeon with absence  
356 of many pathways in the genome.

357 a.ci.di’ phi.lum. M.L. neut. n. *acidum* from an acid; Gr. adj. *philos* from loving; M.L. neut.  
358 adj. *acidiphilum* means acid-loving.

359

### 360 **Discussion**

361 In the present work, the enrichment culture from Parys Mountain AMD system was set up to  
362 grow acidophilic members of the order *Thermoplasmatales*. The culture was eventually highly  
363 enriched in archaea from the genus *Cuniculiplasma*, and incidentally, with the significant (ca.  
364 10 % genomic reads or 20 % of total population) community component belonging to yet  
365 uncultured archaea distantly related with ARMAN-2. Due to its high numbers in the  
366 enrichment, we were able to produce the fully assembled genome of the “ARMAN”-related  
367 organism. Based on the genome annotation and experimental data (co-existence in an  
368 enrichment culture and fluorescence microscopy), we inferred that the metabolic needs of this  
369 sentinel of *Cuniculiplasma* spp. termed “*Ca. Mancarchaeum acidiphilum*” resemble to some  
370 extent those of other archaea co-occupying the environment (e.g. reliance on external  
371 proteinaceous compounds and amino acids). However, the incompleteness or absence of the  
372 central metabolic pathways (e.g. TCA, glycolysis, quinone biosynthesis, etc.) and reduced  
373 genome size support an obligate partner-dependent (or ‘ectoparasitic’) lifestyle. Our data (Fig.  
374 3) further suggest that sizes of Mia14 cells (and likely other ARMAN-related archaea) have a  
375 broad range, usually larger than the diameter of membrane filter pores (0.22  $\mu\text{m}$ ) used to enrich

376 for these organisms. The penetration of cells through the 0.22  $\mu\text{m}$  pores of membrane filters  
377 observed previously<sup>3</sup> may also be explained by the lack of rigid cell walls in these organisms.  
378 For example, the majority of 1-2  $\mu\text{m}$ , cell wall-deficient *Thermoplasmatales* may squeeze  
379 through pores of this diameter.

380 The occurrence of laterally transferred genes and genomic islands from  
381 *Cuniculiplasma* spp. in Mia14 highlights the relative connection between these organisms co-  
382 existing in one environment. It is furthermore likely that extracellular structures such as pili or  
383 pili-like organelles might be present in Mia14. One may also speculate on massive exchange  
384 of DNA through some cell pores or by using the Type IV pili system and numerous membrane  
385 proteins encoded within GIs, the likely conjugative elements.

386 Under our experimental conditions, the preferred partner of Mia14 was  
387 *Cuniculiplasma divulgatum* (previously known as ‘G-plasma’<sup>16</sup>), which is an abundant  
388 inhabitant in AMD. However, the distribution of archaea related to Mia14 (or to ARMAN-2  
389 cluster) in diverse, sometimes non-acidic environments, emphasizes their higher plasticity and  
390 ability to adapt to the broader range of environmental conditions. This broader distribution of  
391 “ARMAN”-related organisms in other environments also suggests that *Cuniculiplasma* spp.  
392 may not necessarily be the exclusive partner (host) for ARMAN-2-like organisms.

393 Mia14 is characterised by a very rudimentary metabolic capability. It is even devoid of  
394 minimal sets of enzymes required for biosynthesis of both types of nucleotides (purine and  
395 pyrimidine) and of 12 out of 20 amino acids (lysine, methionine, arginine, asparagine, alanine,  
396 aspartate, leucine, isoleucine, threonine, phenylalanine, tyrosine, and tryptophan).  
397 Biosynthetic pathways for vitamins and cofactors (B1, B2, Coenzyme A, Coenzyme PQQ, B6,  
398 B12, heme, methanopterin, and ubiquinone/menaquinone) are incomplete.

399 In Mia14, all glycolytic enzymes are missing. The majority of enzymes for the pentose-  
400 phosphate pathway and the entire TCA cycle are also absent. On the other hand, the non-  
401 phosphorylating Entner-Doudoroff (ED) pathway of glucose oxidation is present.  
402 Additionally, fatty acid metabolism and beta-oxidation, folate cycle, phospholipid  
403 biosynthesis, aminosugar metabolism, glycine and serine catabolism pathway, urea cycle and  
404 amino group metabolism, nicotinamide, pyruvate metabolism and interconversion of pyruvate  
405 and acetyl-CoA, trehalose biosynthesis, glycogen metabolism and biosynthesis, propionate  
406 metabolism, heme biosynthesis, pentose-phosphate pathway (non-oxidative phase) and LPS

407 synthesis are absent. Furthermore, we have not found any substrate-level phosphorylation  
408 pathways. The Mia14 respiratory chain is also absent; no Complex I (NADH:ubiquinone  
409 oxidoreductase)<sup>33</sup>, Complex II (succinate:quinone oxidoreductase)<sup>34</sup>, Complex III (either  
410 cytochrome *bc<sub>1</sub>* complex<sup>35</sup> or ACIII<sup>36</sup>) or Complex IV (heme-copper oxygen reductases)<sup>37</sup>  
411 proteins-coding genes were found in the genome. However, the presence of H<sup>+</sup>-translocating  
412 V-type ATP synthase in the organism suggests the activity of PMF-generating complexes. The  
413 only candidate complex for this role is the cytochrome *bd* quinol oxidase, which was found in  
414 the genome. The lack of appropriate endogenous electron donors for this complex in Mia14,  
415 which is deficient in isoprenoid quinone biosynthesis, could be compensated by exogenous  
416 quinones from the membrane of *Cuniculiplasma* sp., considering the assumption of mutualistic  
417 interactions between Mia14 and this organism. Indeed, the QH<sub>2</sub> oxidising cytochrome *b<sub>558</sub>* in  
418 the Mia14 cytochrome *bd* complex is localised on the surface of the cell membrane, as inferred  
419 from topology prediction and alignment of MIA14\_0653 amino acid sequence with its  
420 extensively characterised homolog from *E. coli*<sup>31</sup>. As both *Cuniculiplasma* species<sup>15</sup> lack cell  
421 walls and their cells are usually found in tight contact with Mia14 (Fig. 3), we can speculate  
422 that the latter organism utilises a broad diversity of *Cuniculiplasma* membrane quinones  
423 (either from living or dead cells) as electron donors for energy conservation. However, no  
424 genes of canonical heme biosynthesis, heme import pathways<sup>38</sup> or an alternative pathway for  
425 the formation of heme<sup>39</sup> have been found in the Mia14 genome.

426 Besides the possibility of a completely novel heme biosynthesis pathway in this  
427 archaeon, the only way for proper assembly of the cytochrome *bd* complex is the incorporation  
428 of exogenous hemes. Accumulation of exogenous hemes in the membrane, which is capable  
429 of complementing the growth of heme-deficient organisms, has been demonstrated for  
430 pathogenic bacteria<sup>40</sup>. Considering that hemes *b* and *d* bind covalently to apoproteins and that  
431 the heme-binding amino acids are localised close to the surface of the cell membrane in  
432 cytochrome *bd* complexes<sup>31</sup>, it seems possible for Mia14 to acquire exogenous hemes from  
433 *Cuniculiplasma* spp. to assemble its only PMF-generating complex. It should be noted that the  
434 complete set of genes for canonical or non-canonical heme biosynthesis pathways is also  
435 absent in *Cuniculiplasma* strains PM4 and S5, although these aerobically respiring organisms  
436 possess heme-containing enzymes of the electron transfer chain<sup>16</sup>. It therefore seems possible

437 that *Cuniculiplasma*, and probably Mia14, possess yet unknown mechanisms of heme  
438 biosynthesis.

439 In many archaea, the surface layer is the only cell envelope component providing all  
440 functions normally associated with a cell wall, i.e. acting as the protective barrier and  
441 maintaining the cell shape. However, in some cases the surface layer proteins may also help  
442 in cell-cell association<sup>41,42</sup>. The Mia14 surface layer likely possesses a very complex and  
443 unique architecture, consisting of at least eight strain-specific secreted surface proteins. It is  
444 noteworthy that only four of these surface proteins (MIA14\_0152, \_0331, \_0793 and \_0946)  
445 require almost 2.5% of the whole genome. We identified two domain types in surface layer  
446 proteins displaying the polycystic kidney disease (PKD) superfamily fold and beta-propeller  
447 Kelch and YVTN  $\beta$ -repeat domains fused to CARDB (cell adhesion related domain found in  
448 bacteria)-like adhesion module. Six of these surface layer proteins are predicted to be gained  
449 from various methanogenic and acidophilic euryarchaea and the members of 'TACK'  
450 superphylum. As previously hypothesised<sup>42</sup>, the expansion of proteins containing PKD and  
451 YVTN domains indicates their function in cell-cell interactions. Thus, we propose that the  
452 very rudimentary metabolic capability of Mia14 indicates a *Cuniculiplasma*-associated  
453 lifestyle and that numerous systems such as type IV pili, surface proteins and membrane  
454 channels provide an interface for the exchange of metabolites, energy, macromolecules  
455 including DNA between Mia14 and its host.

456

457

## 458 **Methods**

### 459 **Samples proceedings**

460 Samples from sediments and water of acidic streamer were taken for the establishment of  
461 enrichment cultures in March of 2011 from copper-containing sulfidic ores, Parys Mountain,  
462 Anglesey, North Wales, UK (53°23'13.6"N 4°20'58.6"W ). The enrichment cultures were  
463 supplemented with yeast extract and glucose each at concentrations of 0.1 % (w/vol) and  
464 grown at pH 1-1.2 and 37 °C in AB medium<sup>15</sup>.

465 DNA was extracted by G'NOME DNA Kit (MP Biomedicals). For the metagenomic  
466 study and second series of enrichment cultures set up for CARD-FISH experiments, sediments

467 and water were collected in July, 2014 from the same sampling spot as in March, 2011. The  
468 metagenomic DNA was isolated with DNA Power Isolation Kit for Soil (MoBio).

469 DNA concentrations in all cases were measured using Varian Cary Eclipse  
470 fluorescence spectrophotometer using Quant-iT DNA Assay Broad Range Kit (Life  
471 Technologies).

#### 472 **Genome sequencing and annotation**

473 The genomes were sequenced, assembled and annotated at Fidelity Systems, Inc.  
474 (Gaithersburg, MD), as previously reported<sup>16</sup>. Final assemblies provided ca. 564 and 561-fold  
475 coverages for strain S5 and PM4, respectively<sup>16</sup>, while Mia14 genome was covered 42-fold.  
476 Genomic islands (GIs) were inspected using Island Viewer 3 (ref. <sup>25</sup>) using two different  
477 algorithms: IslandPath-DIMOB, based on the analysis of mobile element-related genes and  
478 dinucleotide distribution biases<sup>43</sup>, and SIGI-HMM, exploiting biases of codon usage  
479 implementing a hidden Markov model approach<sup>44</sup>. In most cases, after manual inspection of  
480 taxonomic affiliation of best blast hits of predicted horizontally transferred proteins, both  
481 predictions were considered as genomic islands. Analysis of proteins shared between Mia14  
482 and *C. divulgatum* S5, but absent in *C. divulgatum* PM4, was performed by ProteinOrtho<sup>23</sup>  
483 V5.15 using default parameters (10<sup>-5</sup> blastp e-value, 50% minimal query coverage and 25%  
484 minimal percent identity).

485 Based on the genomic data, the specific primers for the detection of Mia14-related  
486 organisms in enrichment cultures were: 5' - 3' F Micr (GCTTGGCGAATAAGTGCTGGGC)  
487 and R Micr (ATCTTGCGACCGTACTCCCCAG).

488

#### 489 **Metagenome sequencing of Parys Mountain community**

490 For sequencing of the Parys Mountain acidic streamer metagenome, both paired-end and mate-  
491 paired DNA libraries were used. Paired end library was prepared from 400 ng of environmental  
492 DNA with NEBNext Ultra DNA library preparation kit (New England Biolabs, Ipswich, USA)  
493 according to manufacturer's instructions to obtain mean library size of 500 bp. Mate-paired  
494 libraries were prepared with Nextera™ Mate Pair Library Prep Kit (Illumina Inc., San Diego,  
495 CA, USA) using gel-free protocol supplied by manufacturer. Both libraries were sequenced  
496 with 2x250 bp reads with MiSeq™ Personal Sequencing System (Illumina Inc., San Diego,  
497 CA, USA). After sequencing, all reads were subjected to stringent quality filtering with CLC

498 Genomics Workbench 8.5 (Qiagen, Germany). After filtering, overlapping paired-end library  
499 reads were merged with SeqPrep tool (<https://github.com/jstjohn/SeqPrep>) resulting in  
500 4,110,617 single reads and 7,539,176 read pairs. Mate paired reads were treated with NextClip  
501 tool<sup>45</sup>, resulting in 663,171 read pairs with mean insert size of 2170 bp. Reads were assembled  
502 with metaSPADES<sup>46</sup>, resulting in metagenomic assembly of about 200 Mb of total length  
503 consisting of 93342 contigs with N50 of 3295.

504 For the binning for metagenomic contigs they were aligned against NCBI non-  
505 redundant protein database using DIAMOND in “blastx” mode<sup>47</sup> with e-value of  $10^{-6}$ . Results  
506 of the alignment were imported to MEGAN 6.4.22 (ref. <sup>48</sup>) with default settings adjusted as  
507 follows: min score – 80, top percent – 10, min support – 20. Binning by MEGAN was  
508 performed using default settings. After the initial automatic binning step, additional manual  
509 inspection was performed. In particular, contigs with ambiguous taxonomic affiliation,  
510 characterised by mixed blastx hits were either reassigned to a bin of a higher taxonomic level  
511 or moved to the ‘Unassigned’ bin. *Cuniculiplasma sp.* – related and Mia14-related contigs  
512 were identified manually using blasting their genome sequences with blastn against the local  
513 Parys Mountain metagenomic contigs nucleotide blast database.

514 For the calculation of the taxon abundance, all metagenomic reads were mapped to the  
515 contigs with Bowtie 2 (ref. <sup>49</sup>). Total length of all sequencing reads mapped to every particular  
516 bin was calculated with samtools<sup>50</sup>. Abundance was calculated as a ratio between total lengths  
517 of all sequencing reads to the average genome size of the corresponding taxon (based on NCBI  
518 genomes database). Relative abundance value which was used for the Fig. 1 was calculated as  
519 ratio of bin abundance to the sum of bin abundances.

## 520 **CARD-FISH**

521 Samples were fixed for 1 h at room temperature with pre-filtered formaldehyde (final  
522 concentration 2% vol/vol). Sample (diluted from  $10^{-1}$  to  $10^{-3}$ , according to cell concentrations)  
523 was filtered through 0.22  $\mu\text{m}$  ( $\varnothing$  25 mm) polycarbonate membranes (New Technologies Group  
524 Srl, NTG). Cell permeabilisation was performed by incubation for 1 h with lysozyme (10 mg  
525  $\text{ml}^{-1}$  in TE buffer pH 8.0) followed by incubation with achromopeptidase for 30 min (5 mg  
526  $\text{ml}^{-1}$ ), both at 37°C. Filters were cut into sections and cells were hybridized with universal  
527 horseradish peroxidase (HRP)-labelled oligonucleotide probes for *Eubacteria* (EUB338 I, II,  
528 III probe mix)<sup>51,52</sup> to check for bacterial presence and for *Archaea* (Arch915)<sup>53</sup>. Absence of

529 unspecific hybridization was controlled by implication of the nonspecific probe NON338. The  
530 CARD-FISH probes specific for members of order *Thermoplasmatales* (Thpmt-680R), of  
531 family *Cuniculiplasmataceae* (Clpm-1100R) and of “*Ca. Mancarchaeum acidiphilum*” Mia14  
532 (ARM-MIA1469R) were designed through this study. Detailed information about the probes  
533 is given in Supplementary Table 4. Intracellular peroxidase was inhibited by treatment with  
534 1% H<sub>2</sub>O<sub>2</sub> at room temperature for 20 min. For signal amplification tyramide-Alexa488 and -  
535 Alexa594 were used<sup>54</sup>. The filter sections were counter-stained with DAPI (2 µg ml<sup>-1</sup>) in a  
536 four-to-one ratio of Citifluor (Citifluor Ltd, Leicester, UK): Vectashield (Linaris GmbH,  
537 Wertheim-Bettingen, Germany). At least 200 DAPI-stained and Alexa-positive cells were  
538 counted in a minimum of 10 fields under an AXIOPLAN 2 Imaging microscope (Zeiss,  
539 Germany).

540 **Sequence analysis and evolutionary reconstructions.** Protein coding genes of Mia14 were  
541 assigned to archaeal Clusters of Orthologous Groups (arCOGs) as follows: PSSMs derived  
542 from arCOG alignments were used as PSI-BLAST queries in a search against a database of  
543 archaeal proteins with e-value cutoff of 10<sup>-4</sup>. Proteins (fragments) were assigned to arCOGs  
544 with the highest-scoring hits<sup>55</sup>. Also, sequences of the 56 ribosomal proteins universally  
545 conserved in archaea<sup>56</sup> from 285 organisms with completely or almost completely sequenced  
546 genomes were aligned using the MUSCLE program<sup>57</sup>. Alignments were concatenated; the  
547 phylogenetic tree was reconstructed using the FastTree program<sup>58</sup> with WAG evolutionary  
548 model and gamma-distributed site rates.

549 Manual curation of automatic functional predictions was performed according to the  
550 recent protocol<sup>59</sup>. In particular, the proteins of central carbohydrate pathways (Embden-  
551 Meyerhoff & Gluconeogenesis, Entner-Doudoroff, pentose-phosphate, TCA) including  
552 currently known archaeal modifications<sup>60</sup> were searched by BLAST (with a consciously low  
553 e-value cut-off = 1.0 to avoid loss of distantly related sequences) of sets, including several  
554 amino acid sequences of biochemically characterised (mainly, that with “Evidence at protein  
555 level” in Swissprot database) archaeal and bacterial proteins against the genome (tBLASTn)  
556 or in silico translated proteome (BLASTp) of Mia14. If no hits with all queries were found the  
557 protein was regarded as absent. If all / many of proteins of the pathway were absent the  
558 pathway was regarded as absent. If any BLAST hits were obtained, these sequences were

559 BLASTed against Uniprot and Swissprot (e-value threshold = 0.01) and resulted hits analysed.  
560 The co-localization of genes for a particular pathway was also taken into account.

561 arCOG phyletic patterns of the 15 'DPANN' group genomes were analysed using the  
562 COUNT program<sup>61</sup> as described previously<sup>62</sup>. A matrix with the numbers of orthologs in the  
563 given arCOG in the given organism and the tree of the corresponding genomes were used to  
564 estimate the parameters of a phylogenetic birth and death model with gamma-distributed gain,  
565 loss and duplication rates<sup>61</sup>. The solution produces posterior probabilities for the presence or  
566 absence of a gene in ancestral genomes as well as the probabilities of gene gains and losses on  
567 all tree branches, providing a comprehensive picture of these events in the evolutionary history  
568 of the 'DPANN' group. The reconstructed 'DPANN' group ancestor was compared to the  
569 previously reconstructed last common ancestor of all *Archaea*<sup>63</sup>. To identify actual arCOGs in  
570 three groups (likely present, lost, gained) probability of each event more or equal 50% has  
571 been chosen for each lineage of interest.

572 Taxonomic affiliations for proteins, encoded in the 10 out of 15 'DPANN' genomes (to  
573 the exclusion of the three genomes in the *Nanoarchaeota* archaeon SCGC AAA011-G17  
574 lineage and two genomes in the "*Candidatus Haloredivivus*" lineage that have close relatives  
575 within the 'DPANN' group) was assessed by running protein BLAST search other archaeal  
576 genomes. The database contained 704,591 proteins from 286 complete and nearly complete  
577 archaeal genomes, available at GenBank and the protein set encoded by Mia14. The top  
578 BLAST hit (e-value threshold of  $10^{-6}$ ) outside of the self genome was recorded as an  
579 approximate indication of the taxonomic affiliation of the protein.

580

581 **Data availability.** Sequence data determined in this study are available at NCBI under  
582 BioProject Accession PRJNA353339. Genome sequencing and assembly are deposited in the  
583 GenBank under Accession CP019964. Metagenomic reads and contigs were submitted to MG-  
584 RAST and can be provided from the corresponding author upon request.

585

586

587

588 **References**

- 589 1. Baker, B. J. *et al.* Lineages of acidophilic archaea revealed by community genomic  
590 analysis. *Science* **314**, 1933-1935 (2006).
- 591 2. Méndez-García, C. *et al.* Microbial diversity and metabolic networks in acid mine  
592 drainage habitats. *Front. Microbiol.* **6**, 475 (2015).
- 593 3. Comolli, L. R., Baker, B. J., Downing, K. H., Siegerist, C. E. & Banfield, J. F. Three-  
594 dimensional analysis of the structure and ecology of a novel, ultra-small archaeon.  
595 *ISME J.* **3**, 159-167 (2009).
- 596 4. Golyshina, O. V. Environmental, biogeographic, and biochemical patterns of archaea  
597 of the family *Ferroplasmaceae*. *Appl. Environ. Microbiol.* **77**, 5071-5078 (2011).
- 598 5. Baker, B. J. *et al.* Enigmatic, ultrasmall, uncultivated Archaea. *Proc. Natl Acad.*  
599 *Sci. USA* **107**, 8806-8811 (2010).
- 600 6. Petitjean, C., Deschamps, P., López-García, P. & Moreira, D. Rooting the domain  
601 archaea by phylogenomic analysis supports the foundation of the new kingdom  
602 Proteoarchaeota. *Genome Biol. Evol.* **7**, 191-204 (2014).
- 603 7. Forterre, P. The universal tree of life: an update. *Front. Microbiol.* **6**, 717. (2015).
- 604 8. Huber, H. *et al.* A new phylum of Archaea represented by a nanosized  
605 hyperthermophilic symbiont. *Nature* **417**, 63-67 (2002).
- 606 9. Morris, B. E., Henneberger, R., Huber, H. & Moissl-Eichinger C. Microbial syntrophy:  
607 interaction for the common good. *FEMS Microbiol. Rev.* **37**, 384-406 (2013).
- 608 10. Podar, M. *et al.* Insights into archaeal evolution and symbiosis from the genomes of a  
609 nanoarchaeon and its inferred crenarchaeal host from Obsidian Pool, Yellowstone  
610 National Park. *Biol. Direct.* **8**, 9 (2013).
- 611 11. Wurch, L. *et al.* Genomics-informed isolation and characterization of a symbiotic  
612 Nanoarchaeota system from a terrestrial geothermal environment. *Nat. Commun.* **7**,  
613 12115 (2016).
- 614 12. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark  
615 matter. *Nature* **499**, 431-437 (2013).
- 616 13. Castelle, C. J. *et al.* Genomic expansion of domain archaea highlights roles for  
617 organisms from new phyla in anaerobic carbon cycling. *Curr. Biol.* **25**, 690-701  
618 (2015).
- 619 14. Krause, S., Bremges, A., Muench, P. A., McHardy, A. C. & Gescher, J. Laboratory  
620 cultivation of acidophilic nanoorganisms. Physiological and bioinformatic dissection  
621 of a stable laboratory co-culture. *bioRxiv* 103150 (2017)  
622 doi:<https://doi.org/10.1101/103150>.
- 623 15. Golyshina, O. V. *et al.* The novel, extremely acidophilic, cell wall-deficient archaeon  
624 *Cuniculiplasma divulgatum* gen. nov., sp. nov. represents a new family of  
625 *Cuniculiplasmataceae* fam. nov., order *Thermoplasmatales*. *Int. J. Syst. Evol.*  
626 *Microbiol.* **66**, 332-340 (2016).

- 627 16. Golyshina, O. V. *et al.* Biology of archaea from a novel family *Cuniculiplasmataceae*  
628 (*Thermoplasmata*) ubiquitous in hyperacidic environments. *Sci. Rep.* **6**, 39034 (2016).
- 629 17. Yelton, A. P. *et al.* Comparative genomics in acid mine drainage biofilm communities  
630 reveals metabolic and structural differentiation of co-occurring archaea. *BMC*  
631 *Genomics* **14**, 485 (2013).
- 632 18. Amaral-Zettler, L. A. *et al.* Microbial community structure across the tree of life in  
633 the extreme Rio Tinto. *ISME J.* **5**, 42-50 (2011).
- 634 19. Bohorquez, L. C. *et al.* In-depth characterization via complementing culture-  
635 independent approaches of the microbial community in an acidic hot spring of the  
636 Colombian Andes. *Microb. Ecol.* **63**, 103-115 (2012). Erratum in: *Microb. Ecol.* **63**,  
637 238 (2012).
- 638 20. Hawkins, A. N., Johnson, K. W. & Bräuer, S. L. Southern Appalachian peatlands  
639 support high archaeal diversity. *Microb. Ecol.* **67**, 587-602 (2014).
- 640 21. Juottonen, H., Tuittila, E. S., Juutinen, S., Fritze, H. & Yrjälä, K. Seasonality of rDNA-  
641 and rRNA-derived archaeal communities and methanogenic potential in a boreal mire.  
642 *ISME J.* **2**, 1157-1168 (2008).
- 643 22. Wu, X. *et al.* Microbial metagenomes from three aquifers in the Fennoscandian shield  
644 terrestrial deep biosphere reveal metabolic partitioning among populations. *ISME J.*  
645 **10**, 1192-1203 (2016).
- 646 23. Waters, E. *et al.* The genome of *Nanoarchaeum equitans*: insights into early archaeal  
647 evolution and derived parasitism. *Proc. Natl. Acad. Sci. USA* **100**, 12984-12988  
648 (2003).
- 649 24. Lechner, M. *et al.* Proteinortho: detection of (co-)orthologs in large-scale analysis.  
650 *BMC Bioinformatics* **12**, 124 (2011).
- 651 25. Dhillon, B.K. *et al.* IslandViewer 3: more flexible, interactive genomic island  
652 discovery, visualization and analysis. *Nucleic Acids Res.* **43**, W104-W108 (2015).
- 653 26. Makarova, K. S. *et al.* Dark matter in archaeal genomes: a rich source of novel mobile  
654 elements, defence systems and secretory complexes. *Extremophiles* **5**, 877-893 (2014).
- 655 27. Baker-Austin, C., Dopson, M., Wexler, M., Sawers, R. G. & Bond, P. L. Molecular  
656 insight into extreme copper resistance in the extremophilic archaeon '*Ferroplasma*  
657 *acidarmanus*' Fer1. *Microbiology* **151**, 2637-2646 (2005).
- 658 28. Bräsen, C., Esser, D., Rauch, B. & Siebers, B. Carbohydrate metabolism in Archaea:  
659 current insights into unusual enzymes and pathways and their regulation. *Microbiol.*  
660 *Mol. Biol. Rev.* **78**, 89-175 (2014).
- 661 29. Bologna, F. P., Andreo, C. S. & Drincovich, M. F. *Escherichia coli* malic enzymes:  
662 two isoforms with substantial differences in kinetic properties, metabolic regulation,  
663 and structure. *J. Bacteriol.* **189**, 5937-5946 (2007).
- 664 30. Makarova, K. S., Koonin, E. V. & Albers, S. V. Diversity and Evolution of Type IV  
665 pili Systems in Archaea. *Front. Microbiol.* **7**, 667 (2016).

- 666 31. Mulkidjanian, A. Y., Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V.  
667 Evolutionary primacy of sodium bioenergetics. *Biol. Direct.* **3**, 13 (2008).
- 668 32. Borisov, V. B., Gennis, R. B., Hemp, J. & Verkhovsky M. I. The  
669 cytochrome *bd* respiratory oxygen reductases. *Biochim. Biophys. Acta* **1807**, 1398-  
670 1413 (2011).
- 671 33. Sazanov, L. A. A giant molecular proton pump: structure and mechanism of respiratory  
672 complex I. *Nat. Rev. Mol. Cell Biol.* **16**, 375–388 (2015).
- 673 34. Cecchini, G. Function and structure of complex II of the respiratory chain. *Annu. Rev.*  
674 *Biochem.* **72**, 77–109 (2003).
- 675 35. Dibrova, D. V., Cherepanov, D. A., Galperin, M. Y., Skulachev, V. P., &  
676 Mulkidjanian, A. Y. Evolution of cytochrome *bc* complexes: from membrane-  
677 anchored dehydrogenases of ancient bacteria to triggers of apoptosis in vertebrates.  
678 *Biochim. Biophys. Acta.* **1827**, 1407-1427. (2013).
- 679 36. Refojo, P. N., Teixeira, M. & Pereira, M. M. The alternative complex III: properties  
680 and possible mechanisms for electron transfer and energy conservation. *Biochim.*  
681 *Biophys. Acta.* **1817**, 1852-1859 (2012).
- 682 37. Ducluzeau, A.-L. *et al.* The evolution of respiratory O<sub>2</sub>/NO reductases: an out-of-the-  
683 phylogenetic-box perspective. *J. R. Soc. Interface* **11**, 196 (2014).
- 684 38. Choby, J. E. & Skaar, E. P. Heme synthesis and acquisition in bacterial pathogens. *J*  
685 *of Molecular Biology.* **428**, 3408-3428 (2016).
- 686 39. Bali, S., Palmer, D. J., Schroeder, S. & Warren N. J. Recent advances in the  
687 biosynthesis of modified tetrapyrroles: the discovery of an alternative pathway for the  
688 formation of heme and heme d<sub>1</sub>. *Cell. Mol. Life Sci.* **71**, 2837 (2014)
- 689 40. Skaar, E. P., Humayun, M., Bae, T., DeBord, K. L. & Schneewind, O. Iron-source  
690 preference of *Staphylococcus aureus* infections. *Science* **305**, 1626–1628 (2004).
- 691 41. Mayerhofer, L.E., Conway de Macario, E., Yao, R. & Macario, A.J. Structure,  
692 organization, and expression of genes coding for envelope components in the archaeon  
693 *Methanosarcina mazei* S-6. *Arch. of Microbiol.* **169**, 339-345 (1998).
- 694 42. Jing, H. *et al.* Archaeal surface layer proteins contain  $\beta$  propeller, PKD, and  $\beta$  helix  
695 domains and are related to metazoan cell surface proteins. *Structure* **10**, 1453-1464  
696 (2002).
- 697 43. Waack, S. *et al.* Score-based prediction of genomic islands in prokaryotic genomes  
698 using hidden Markov models. *BMC Bioinformatics* **7**, 142 (2006).
- 699 44. Hsiao, W., Wan, I., Jones, S. J. & Brinkman, F. S. L. IslandPath: Aiding detection of  
700 genomic islands in prokaryotes. *Bioinformatics* **19**, 418–420 (2003).
- 701 45. Leggett, R. M., Clavijo, B. J., Clissold, L., Clark, M. D. & Caccamo, M. NextClip: an  
702 analysis and read preparation tool for Nextera Long Mate Pair libraries. *Bioinformatics*  
703 **30**, 566-568 (2014).
- 704 46. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. metaSPAdes: a new versatile  
705 *de novo* metagenomics assembler. arXiv preprint arXiv:1604.03071 (2016).

- 706 47. Buchfink, B., Chao X., & Huson D. H. Fast and sensitive protein alignment using  
707 DIAMOND. *Nat. Methods* **12**, 59-60 (2015).
- 708 48. Huson, D. H. *et al.* MEGAN community edition-interactive exploration and analysis  
709 of large-scale microbiome sequencing data. *PLoS Comput. Biol.* **12**, e1004957 (2016).
- 710 49. Langmead, B. & Salzberg, S. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*  
711 **9**, 357-359 (2012).
- 712 50. Li, H. *et al.* 1000 Genome Project Data Processing Subgroup. The Sequence  
713 Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
- 714 51. Amann, R. I., Krumholz, L. & Stahl, D. A. Fluorescent-oligonucleotide probing of  
715 whole cells for determinative, phylogenetic, and environmental studies in  
716 microbiology. *J. Bacteriol.* **172**, 762-770 (1990).
- 717 52. Daims, H., Brühl, A., Amann, R., Schleifer, K. H. & Wagner, M. The domain-specific  
718 probe EUB338 is insufficient for the detection of all Bacteria: development and  
719 evaluation of a more comprehensive probe set. *Syst. Appl. Microbiol.* **3**, 434-444  
720 (1999).
- 721 53. Stahl, D. A. & Amann, R. I. Development and application of nucleic acid probes, in  
722 "Nucleic acids techniques in bacterial systematics (Eds. Stakebrandt E, Goodfellow,  
723 M Wiley, Chichester 1991), pp.205-242.
- 724 54. Teira, E., Reinthaler, T., Pernthaler, A., Pernthaler, J. & Herndl, G. J. Combining  
725 catalyzed reporter deposition-fluorescence in situ hybridization and  
726 microautoradiography to detect substrate utilization by bacteria and Archaea in the  
727 deep ocean. *Appl. Environ. Microbiol.* **7**, 4411-4414 (2004).
- 728 55. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Archaeal Clusters of Orthologous Genes  
729 (arCOGs): an update and application for analysis of shared features between  
730 *Thermococcales*, *Methanococcales*, and *Methanobacteriales*. *Life* (Basel) **5**, 818-840  
731 (2015).
- 732 56. Yutin, N., Puigbò, P., Koonin, E. V. & Wolf, Y. I. Phylogenomics of prokaryotic  
733 ribosomal proteins. *PLoS One* **7**, e36972 (2012).
- 734 57. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and  
735 space complexity. *BMC Bioinformatics* **5**, 113 (2004).
- 736 58. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-  
737 likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
- 738 59. Toshchakov, S. V., Kublanov I. V., Messina, E., Yakimov, M. M. & Golyshin, P. N.  
739 Genomic analysis of pure cultures and communities. Hydrocarbon and lipid  
740 microbiology protocols (Springer Protocols Handbooks. Humana Press.) ISBN:  
741 1949-2448 (2015).
- 742 60. Bräsen, C., Esser, D., Rauch, B. & Siebers, B. Carbohydrate metabolism in Archaea:  
743 current insights into unusual enzymes and pathways and their regulation. *Microbiol*  
744 *Mol Biol Rev.* **78**, 89-175 (2014).
- 745 61. Csurös, M. Count: evolutionary analysis of phylogenetic profiles with parsimony and

- 746 likelihood. *Bioinformatics* **26**, 1910-1912 (2010).  
747 62. Puigbò, P., Lobkovsky, A. E., Kristensen, D. M., Wolf, Y.I. & Koonin, E. V. Genomes  
748 in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biol.*  
749 **12**, 66 (2014).  
750 63. Wolf, Y.I., Makarova, K. S., Yutin, N. & Koonin, E. V. Updated clusters of  
751 orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of  
752 horizontal gene transfer. *Biol. Direct.* **7**, 46 (2012).

753

#### 754 **Acknowledgments**

755 The work of OVG and PNG was supported by the Royal Society UK- Russia Exchange Grant  
756 #IE 160224 and ERA Net IB2 Project MetaCat through UK Biotechnology and Biological  
757 Sciences Research Council (BBSRC) Grant BB/M029085/1. MMY is thankful to the Italian  
758 Ministry of University and Research under RITMARE Flagship Project (2012-2016). YIW  
759 and KSM are supported by intramural funds of the US Department of Health and Human  
760 Services (to the National Library of Medicine). The work of SVT and AAK was supported by  
761 Royal Society UK- RFBR grant 16-54-10072. We are indebted to Dr Paul Hill (SENRGy,  
762 Bangor University) for critical reading of the manuscript.

763

#### 764 **Author contributions**

765 OVG and PNG conceived the research. OVG, SVT, KSM, YIW, SNG, AAK, IVK, TYN, MF,  
766 MMY and PNG did the genome analysis. VLC and EA did the CARD-FISH experiment. OVG  
767 drafted the manuscript. SVT, KSM, YIW, SNG, AAK, IVK, MMY and PNG contributed to  
768 the manuscript writing.

769

770 **Figure legends**

771 **Figure 1. General structure of Parys Mountain acidic streamer community.** Abundance  
 772 values were calculated using median coverage of metagenomic contigs of each bin with  
 773 normalization to average genome size of bin representatives. Abundance values for *C.*  
 774 *divulgatum* and “*Ca. Mancarchaeum acidiphilum*” are highlighted with green. \* - excluding  
 775 *C. divulgatum*. \*\* - excluding “*Ca. Mancarchaeum acidiphilum*”.

777 **Figure 2. Distribution of Parys Mt. metagenomic contigs by coverage and GC content.**  
 778 Clusters of contigs related to *C. divulgatum* - “*Ca. Mancarchaeum*” and uncultivated  
 779 *Thermoplasmatales* – “*Ca. Micrarchaeota*” excluding “*Ca. Mancarchaeum acidiphilum*”  
 780 microbial consortia are shown with dotted-line ovals.

782 **Figure 3. Archaeal cells visualised by CARD-FISH.** (A) Hybridization with (a) probe  
 783 Thpmt680R (a) and (b) probe ARM-MIA1469R to target *Cuniculiplasma* spp. and “*Ca.*  
 784 *Mancarchaeum*”, respectively. (c) Side-by-side comparison of “*Ca. Mancarchaeum*” versus  
 785 *Cuniculiplasma*. “*Ca. Mancarchaeum*” cells (magenta) localised on green *Cuniculiplasma*  
 786 spp.). Panels (a1), (b1) and (c1) are the magnified images of yellow-boxed fields of panels (a),  
 787 (b) and (c), correspondingly. The image was corrected with Daltonize tool  
 788 (<https://github.com/joergdietrich/daltonize>) to improve perception of deuteranopic persons.

789

790 **Figure 4. Phylogenetic position of Mia14 within Archaea.** . An approximate Maximum  
 791 Likelihood tree based on the concatenated alignment of 56 ribosomal proteins universally  
 792 conserved in *Archaea*. In total, 285 genomes were analysed. Taxa are named according to the  
 793 NCBI taxonomy. Candidate phyla are shown in quotation marks. Lineages with cultured/co-  
 794 cultured representatives are highlighted in blue. NCBI Genome Assembly IDs are shown for  
 795 individual genomes. Scale bar reflects 0.1 substitutions per amino acid position.

796

797 **Figure 5. Genomic features and GIs in “*Ca. Mancarchaeum acidiphilum*”.** Rings from  
 798 outside to inside: genomic coordinates (grey colour); plus-strand CDS (blue); minus-strand  
 799 CDS (blue); genomic islands (green) and RNA (red); GC-content (orange); GC-skew  
 800 (green/magenta); blastn hits with e-value cutoff  $10^{-5}$  vs *C. divulgatum* PM4; blastn hits with e-  
 801 value cutoff  $10^{-5}$  vs *C. divulgatum* S5.

802

803 **Figure 6. Phylogeny of MIA14\_0877 and the neighbourhood of its gene.** Gene  
 804 neighbourhood of CDS for cation transport ATPase/copper-transporting P-type ATPase  
 805 MIA14\_0877 is shown on the right. Cation transport ATPases are shown in green,  
 806 TRASH/YHS-like protein, metallochaperones (arCOG04507) are shown in purple, Lrp-AsnC  
 807 family transcriptional regulators (arCOG01585) and conserved hypothetical proteins  
 808 (arCOG05383) are shown in cyan. Other protein coding genes are shown as dark grey  
 809 pentagons. Size of pentagons is proportional to size of corresponding proteins. The list of  
 810 proteins included in the analysis with protein IDs is provided in Supplementary Table 1.

811

812 **Figure 7. Gains and losses of arCOG families in the evolution of ‘DPANN’ group.**

813 (a) Reconstruction of gene loss and gain along the ‘DPANN’ subtree. Triplets of numbers  
814 indicate the estimates for the arCOG complement, arCOG gains and arCOG losses respectively  
815 for the selected extant or ancestral genomes and adjacent tree branches. Estimates for the  
816 terminal branches are shown next to the extant genome names. The number at the base of the  
817 tree indicates the arCOG complement with gains (+) and losses (-) estimated for the last  
818 archaeal common ancestor (LACA)<sup>49</sup>, ‘DPANN’ ancestor (blue rectangle), common ancestor  
819 of “*Ca. Nanohaloarchaea*”-“*Ca. Micrarchaea*” (grey rectangle), common ancestor of “*Ca.*  
820 *Micrarchaea*” (magenta rectangle) and Mia14 (yellow rectangle). Losses and gains of selected  
821 protein families in course of evolution at above time-points are indicated in textboxes of same  
822 colours, with gains indicated in textboxes located above and losses listed in boxes below (see  
823 Supplementary Data 4 for further details).

824 (b) Number of arCOGs predicted to be gained or lost in the course of evolution of Mia14  
825 lineage with respective COUNT probability >50% by arCOG functional categories. The  
826 functional classification of the arCOGs is shown for two 4 major groups: C-Q - metabolic  
827 genes; J-N - informational genes; V - defence genes; R-S poorly characterized or  
828 uncharacterized genes (for details see  
829 <ftp://ftp.ncbi.nih.gov/pub/wolf/COGs/arCOG/funclass.tab>).