PNAS
www.pnas.org

Supplementary Information for

Heme auxotrophy in abundant aquatic microbial lineages

Suhyun Kim[1], Ilnam Kang[2]*, Jin-Won Lee[3], Che Ok Jeon[4], Stephen J. Giovannoni[5], and Jang-Cheon Cho[1,6]*

Corresponding authors:
*Ilnam Kang: Center for Molecular and Cell Biology, Inha University, Inharo 100, Incheon 22212, Republic of Korea. Tel: +82-32-876-5541, ikang@inha.ac.kr
*Jang-Cheon Cho: Department of Biological Sciences, Inha University, Inharo 100, Incheon 22212, Republic of Korea. Tel: +82-32-860-7711, chojc@inha.ac.kr

This PDF file includes:

Other supplementary materials for this manuscript include the following:

Dataset S1

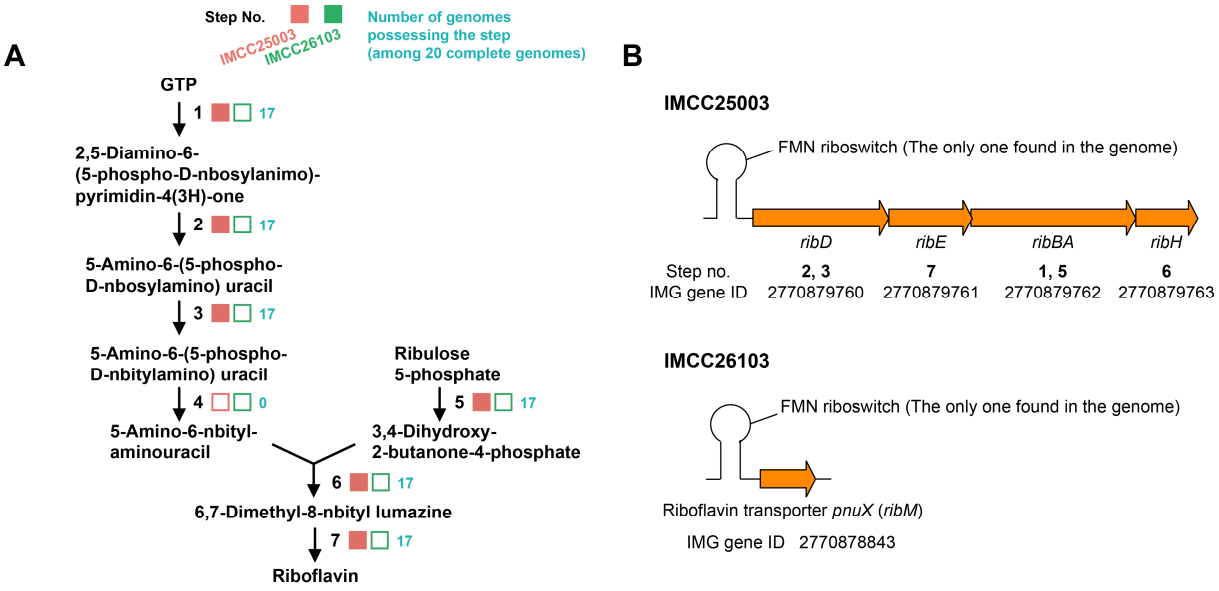

Fig. S1. Riboflavin biosynthetic pathway in the acI genomes. (*A*) The presence (filled) and absence (empty) of each step in the two acI genomes are indicated with colored boxes (orange, IMCC25003; green, IMCC26103). The numbers of the acI genomes (among 20) possessing each step were written in blue. Note that the putative enzymes mediating step four, the only missing step in IMCC25003, have only recently been identified[1,2], and therefore many genomes of riboflavin-prototrophic microorganisms are still predicted to lack this step according to KEGG-based annotation. (*B*) The genes for riboflavin biosynthesis (IMCC25003) and transport (IMCC26103) are located just downstream of the only FMN riboswitch predicted in each genome.

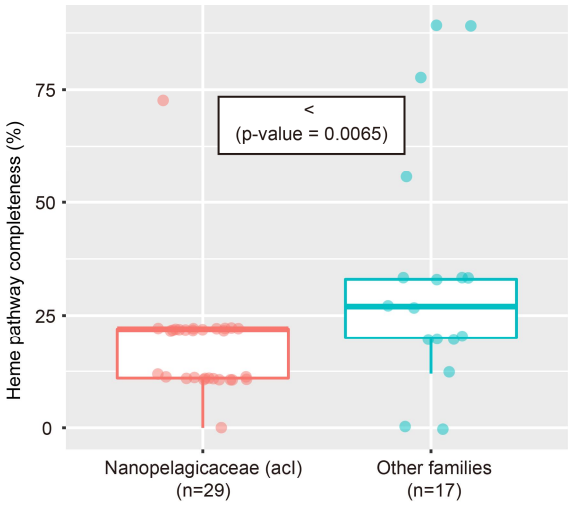

Fig. S2. Distribution of the completeness of heme biosynthesis pathway among the acI lineage (*Nanopelagicaceae*) and other families in the order *Nanopelagicales*. Inset shows the result of one-sided Mann-Whitney *U* test.

Fig. S3. Distribution of the heme biosynthetic pathway completeness in diverse bacterial phyla. **left**, An approximate maximum-likelihood tree of representative genomes for bacterial species clusters constructed by FastTree2 using a concatenated alignment of 120 conserved marker proteins. The genomes were grouped by phyla, and the phylum *Patescibacteria* was set as an outgroup. Phyla with less than 10 genomes were excluded from tree building. **middle,** Ridgeline density plots showing the distribution of the heme biosynthetic pathway completeness in bacterial phyla. The colors under the ridgelines indicate the pathway completeness according to the color gradient on the right. **right,** Average completeness of six pathway modules of bacterial phyla. The color intensity indicates the completeness according to the color gradient on the right. Refer to Fig. 1 for more details on the pathway modules.

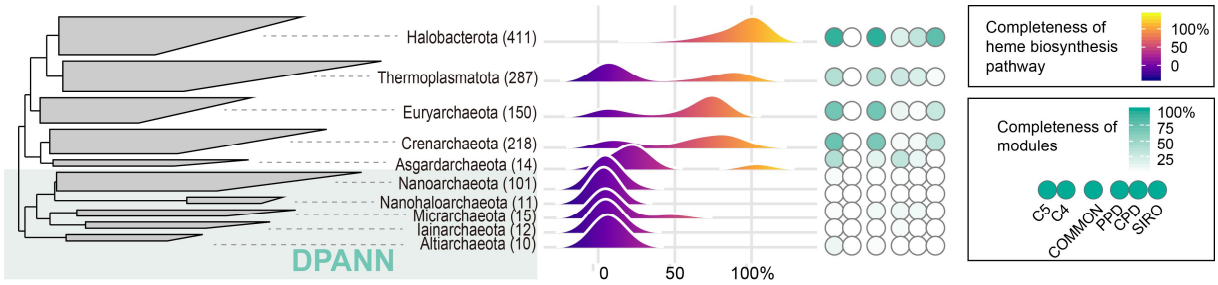

Fig. S4. Distribution of the heme biosynthetic pathway completeness in diverse archaeal phyla. **left**, An approximate maximum-likelihood tree of representative genomes for archaeal species clusters constructed by FastTree2 using a concatenated alignment of 122 conserved marker proteins. The genomes were grouped by phyla, and the superphylum DPANN (from *Nanoarchaeota* to *Altiarchaeota*) was set as an outgroup. Phyla with less than 10 genomes were excluded from tree building. **middle,** Ridgeline density plots showing the distribution of the heme biosynthetic pathway completeness in archaeal phyla. The colors under the ridgelines indicate pathway completeness according to the color gradient on the right. **right,** Average completeness of six pathway modules of archaeal phyla. The color intensity indicates the pathway completeness according to the color gradient on the right. Refer to Fig. 1 for more details on the pathway modules.

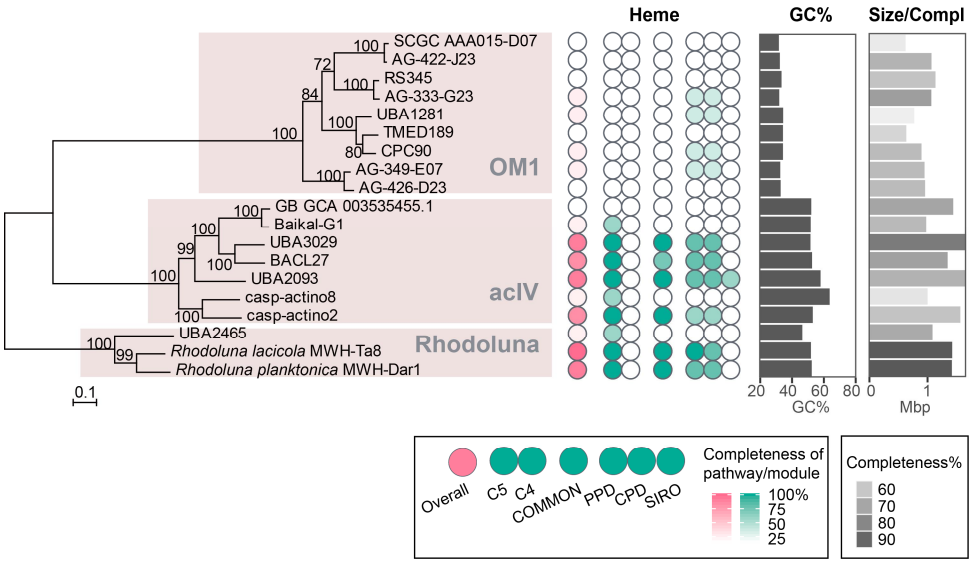

Fig. S5. Heme biosynthetic pathway completeness of representative genomes for species clusters belonging to OM1 (*Candidatus* Actinomarinales; o__TMED189 in GTDB), acIV, and the genus *Rhodoluna* of *Actinobacteriota*. **left**, Maximum-likelihood tree constructed by RAxML using a concatenated alignment of conserved marker proteins. The genus *Rhodoluna* was set as an outgroup. **middle,** Overall heme biosynthetic pathway completeness (indicated in pink) and the completeness of six pathway modules (indicated in green) within the genomes. The color intensity indicates the pathway completeness according to the color scales in the bottom legend. **right,** The GC contents and genome sizes are illustrated with bar graphs. Genome completeness is indicated as the darkness of the genome size bars according to the gray scale at the bottom. The data were downloaded from the GTDB (Release 89).

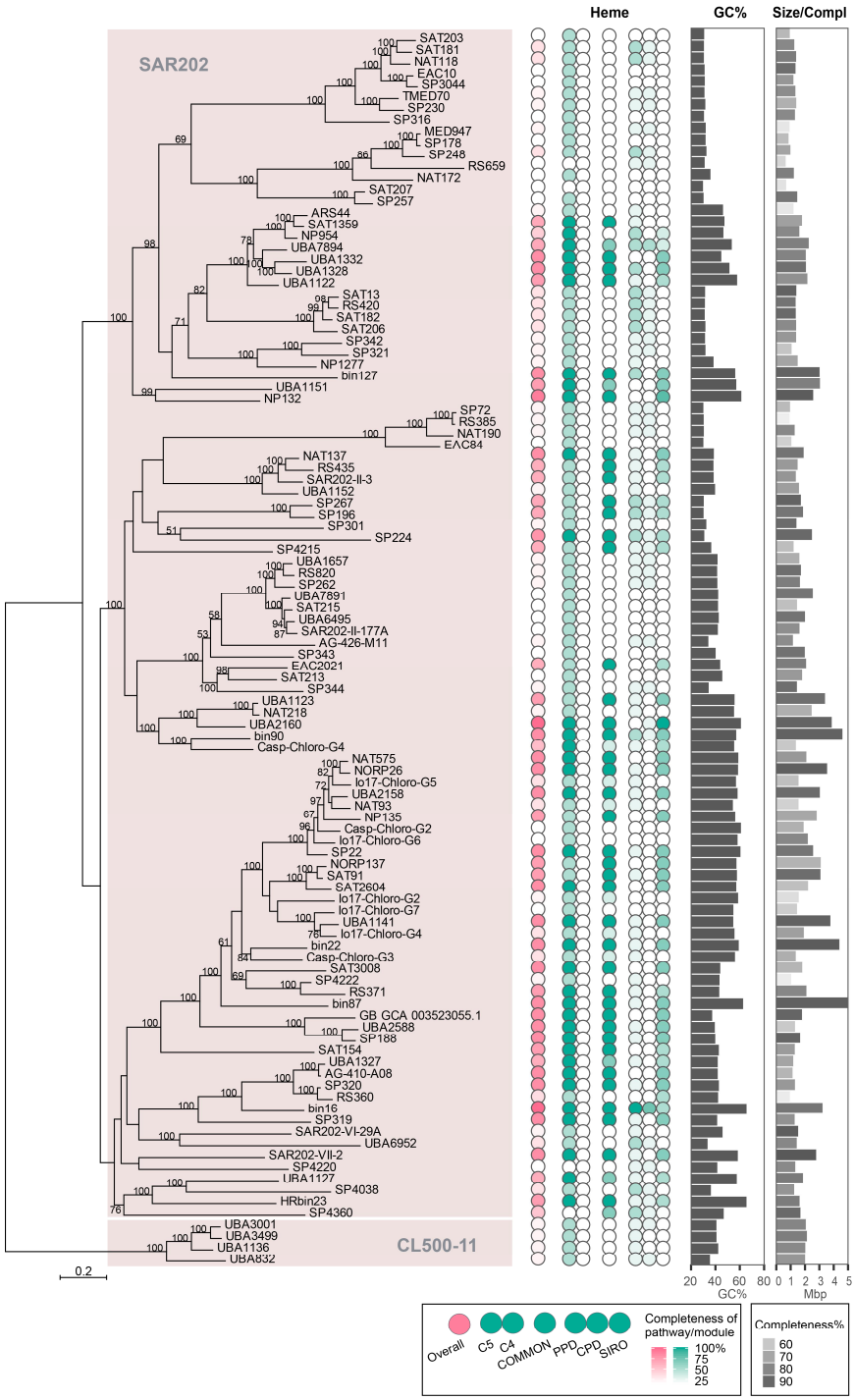

Fig. S6. Heme biosynthetic pathway completeness of representative genomes for species clusters belonging to the SAR202 group of *Chloroflexota*. The CL500-11 group was set as an outgroup in the tree on the left. Refer to the Fig. S4 legend for a detailed explanation.

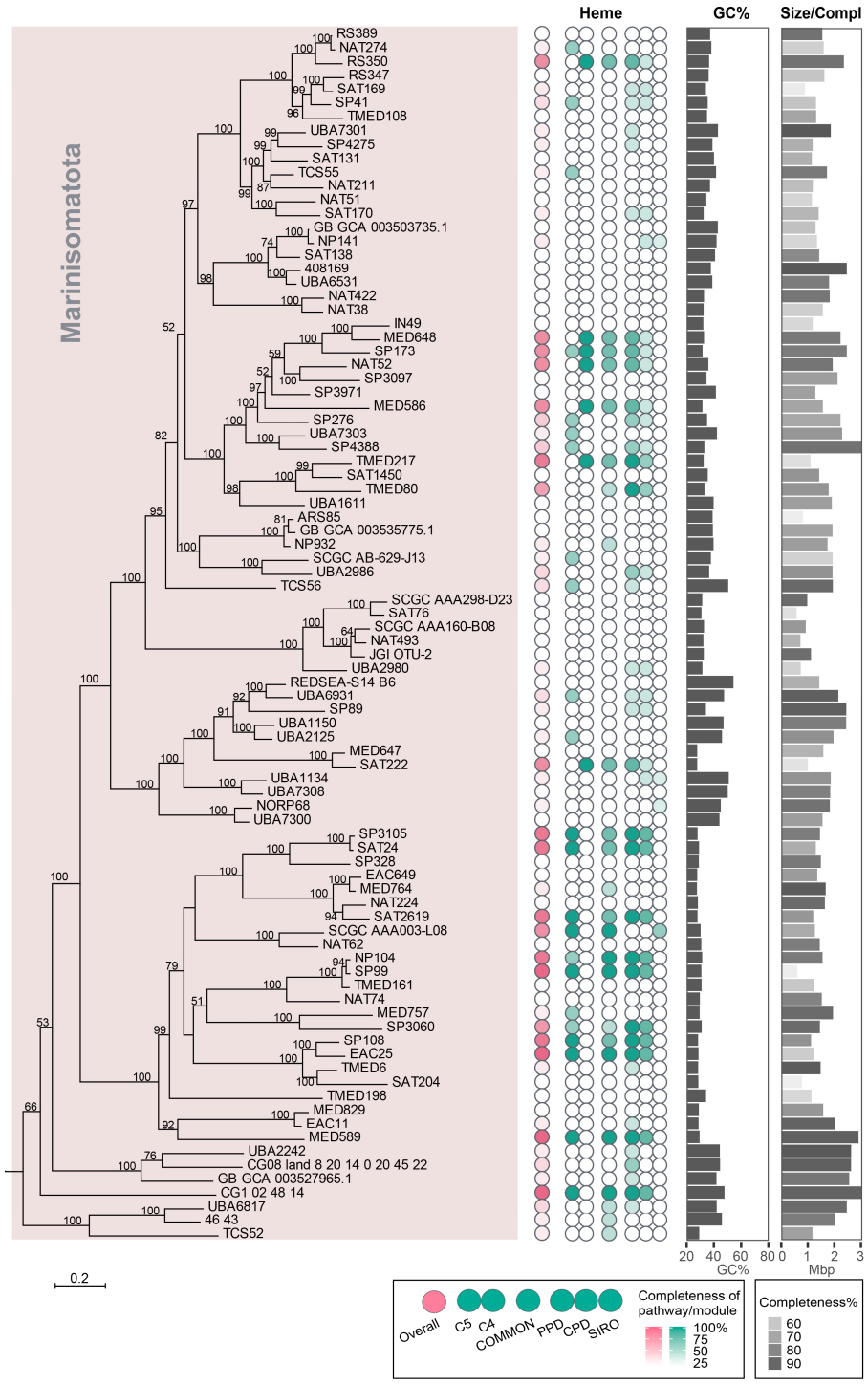

Fig. S7. Heme biosynthetic pathway completeness of representative genomes for species clusters belonging to the *Marinisomatota*. *Flavobacterium aquatile* LMG 4008 (RefSeq assembly ID: GCF_000757385.1) was set as an outgroup in the tree on the left. Refer to the Fig. S4 legend for a detailed explanation.

Fig. S8. Heme biosynthetic pathway completeness of representative genomes for species clusters belonging to the *Nitrososphaeria* and *Asgardarchaeota*. *Asgardarchaeota* was set as an outgroup in the tree on the left. Refer to the Fig. S4 legend for a detailed explanation.

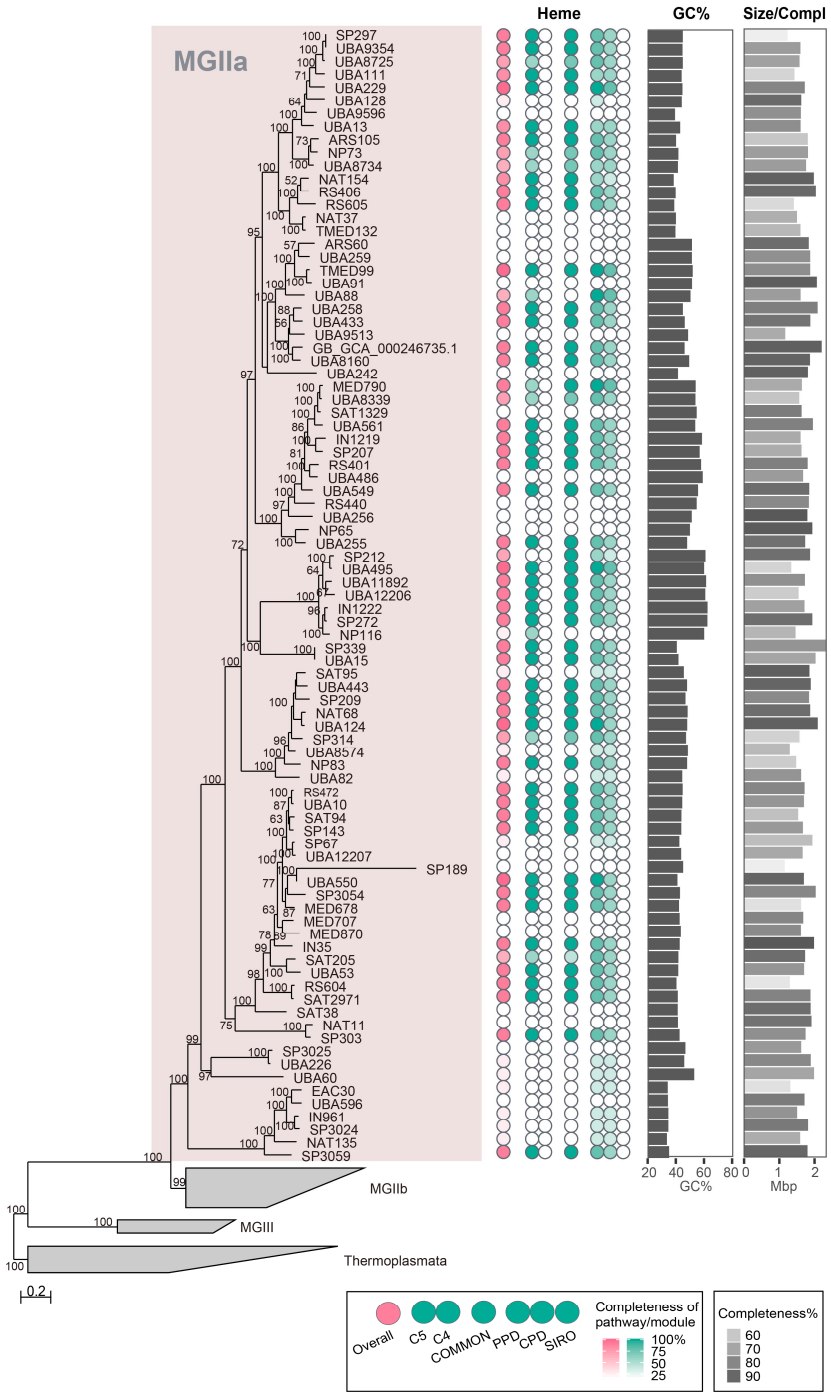

Fig. S9. Heme biosynthetic pathway completeness of representative genomes for species clusters belonging to MGIIa of *Archaea*. *Thermoplasmata* was set as an outgroup in the tree on the left. Refer to the Fig. S4 legend for a detailed explanation.

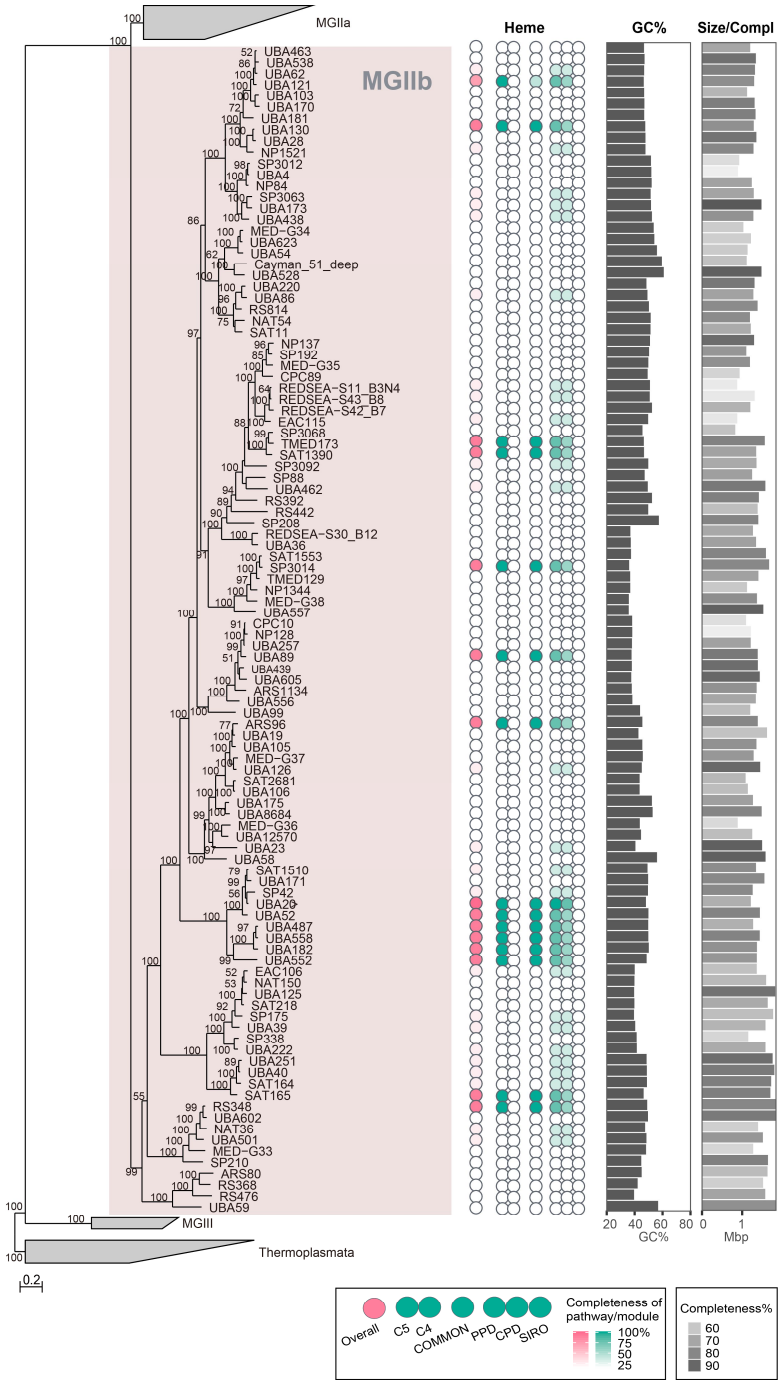

Fig. S10. Heme biosynthetic pathway completeness of representative genomes for species clusters belonging to MGIIb of *Archaea*. *Thermoplasmata* was set as an outgroup in the tree on the left. Refer to the Fig. S4 legend for a detailed explanation.

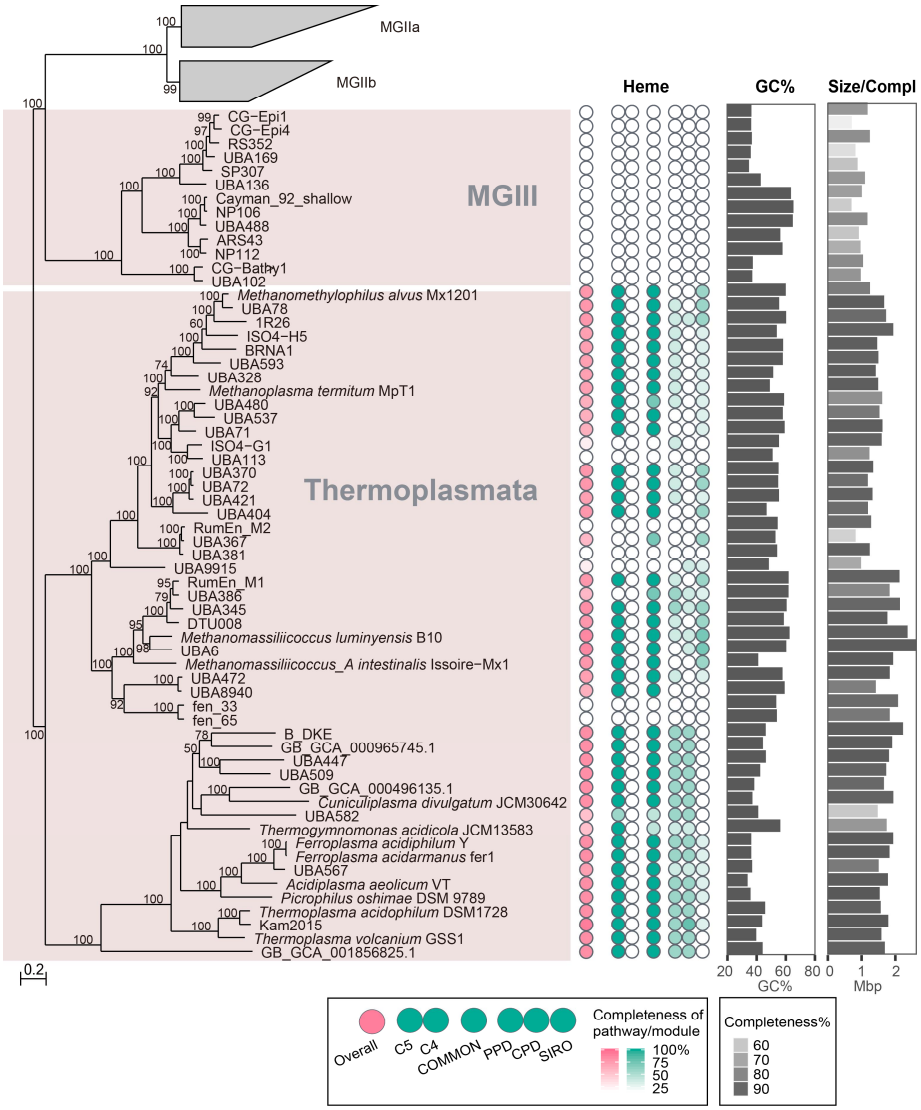

Fig. S11. Heme biosynthetic pathway completeness of representative genomes for species clusters belonging to MGIII and *Thermoplasmata* of *Archaea*. *Thermoplasmata* was set as an outgroup in the tree on the left. Refer to the Fig. S4 legend for a detailed explanation.

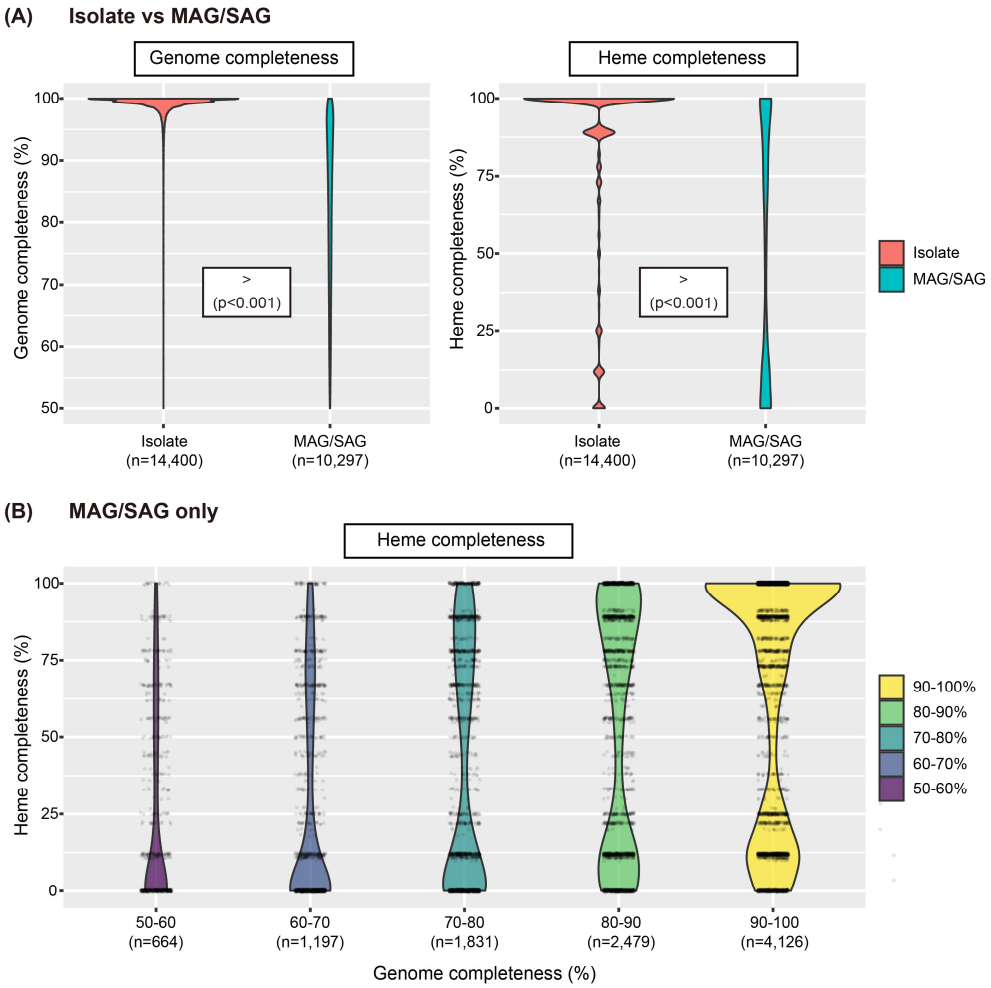

Fig. S12. The completeness of genome and heme biosynthetic pathway in the representative genomes for species clusters of GTDB. (*A*) Distribution of genome and heme completeness among the isolate genomes and MAG/SAGs. Insets show the results of one-sided Mann−Whitney *U* test (*B*) The distribution of heme completeness according to genome completeness among MAG/SAGs. Genomes were divided into five groups according to genome completeness and the distribution of heme completeness in each group was displayed using violin plots. A total of 9 genomes (6 isolates and 3 SAGs) with completeness of <50% were excluded from the analyses for visualization purpose. In all panels, the areas of plots are scaled proportionally to the number of genomes.

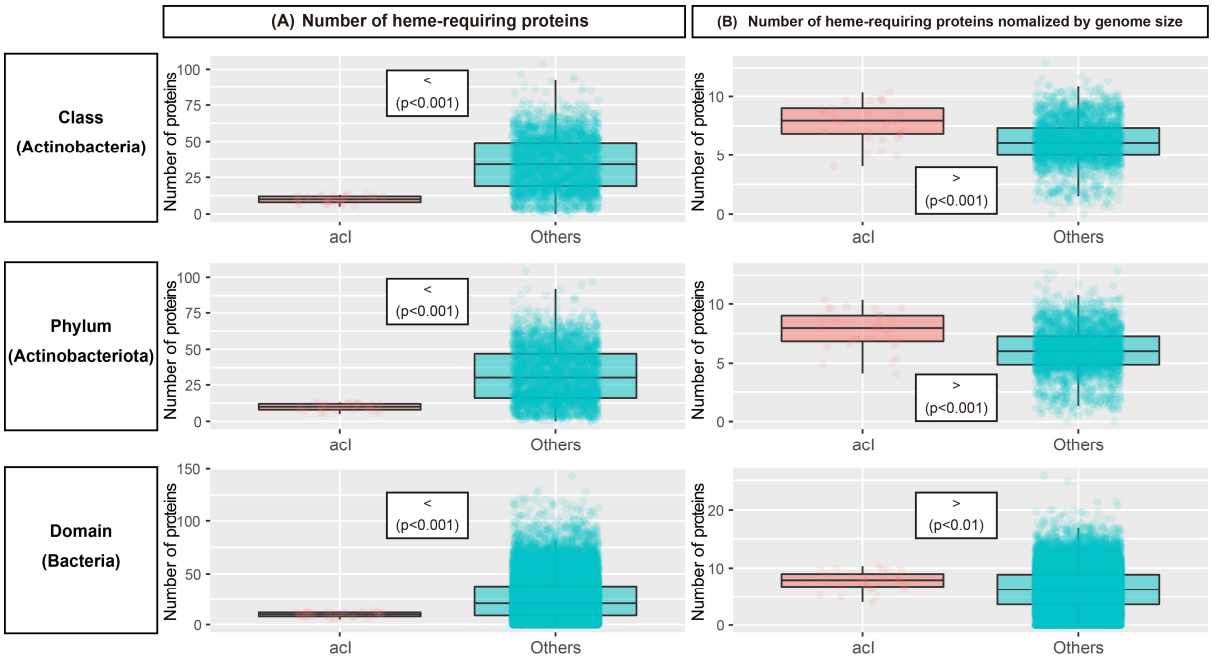

Fig. S13. Distribution of the number of heme-requiring proteins in the genomes of the acI lineage (*Nanopelagicaceae*) and other members of the same class, phyla, or domain. In the three panels on the left, the number of proteins was not normalized by genome sizes, whereas it was normalized in the three panels on the right. Dots represent the genomes. Insets shows the results of one-sided Mann–Whitney *U* test.

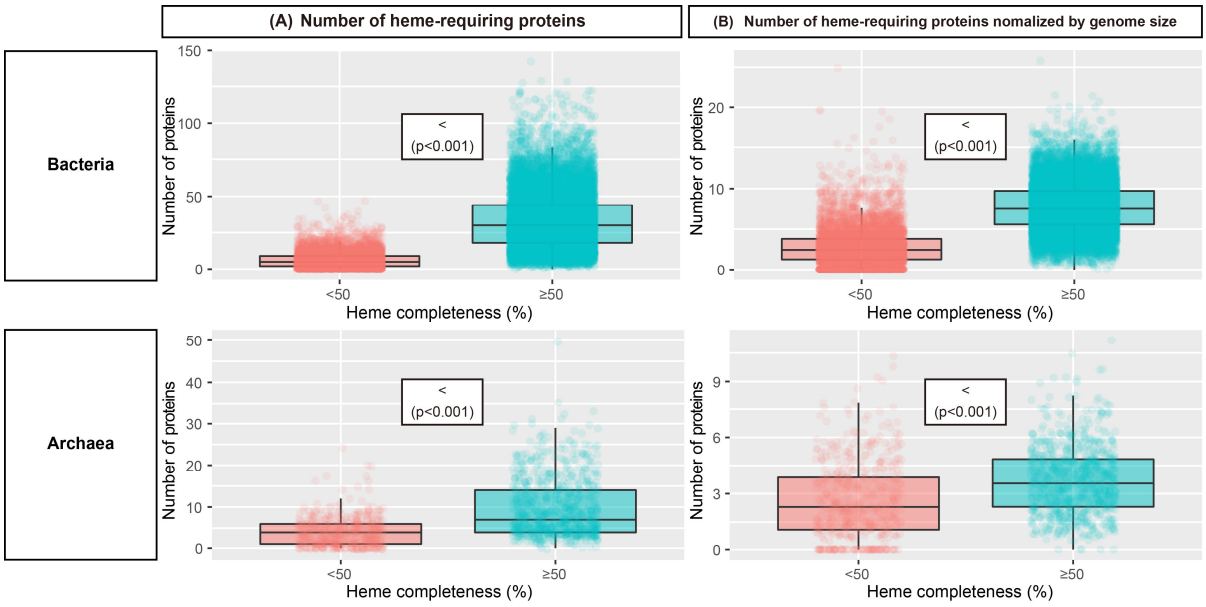

Fig. S14. Distribution of the number of heme-requiring proteins in bacterial (upper panels) and archaeal (lower panels) genomes according to the completeness of heme biosynthetic pathway. Bacterial and archaeal genomes were divided into two groups according to heme completeness with a threshold value of 50%, and the distribution of the number of heme-requiring proteins in each group was displayed using boxplots. In the panels on the left, the number of proteins was not normalized by genome sizes, whereas it was normalized in the panels on the right. Dots represent the genomes. Insets shows the results of one-sided Mann–Whitney *U* test.

Table S1. List of enzyme names and corresponding KO IDs for the steps defined in heme biosynthetic pathway presented in Fig. 1

Step	Enzyme	KO ("|" → OR)
1	Glutamyl-tRNA reductase (HemA)	K02492
2	Glutamate-1-semialdehyde 2,1-aminomutase (HemL)	K01845
3	5-aminolevulinate synthase (AlaS)	K00643
4	5-aminolevulinic acid dehydratase (HemB)	K01698
5	Porphobilinogen deaminase (HemC)	K01749
6	Uroporphyrinogen III synthase (HemD)	K01719 | K13542 | K13543
7	Uroporphyrinogen III decarboxylase (HemE)	K01599
8	Coproporphyrinogen III oxidase (HemF/N)	K00228 | K02495
9	Protoporphyrinogen oxidase (HemG/J/Y)	K00230 | K00231 | K08973
10	Protoporphyrin ferrochelatase (HemH)	K01772
11	Protoporphyrinogen oxidase (HemY)	K00231
12	Ferrochelatase (HemH)	K01772
13	Iron-coproporphyrin oxidative decarboxylase (HemQ, AhbD)	K00435 | K22227
14	Uroporphyrin-III C-methyltransferase (MET1, CysG, CobA, HemX, CobA-HemD, HemDX)	K00589 | K02302 | K02303 | K02496 | K13542 | K13543
15	Precorrin-2 dehydrogenase (CysG, MET8)	K02302 | K02304
16	Sirohydrochlorin ferrochelatase (CysG, MET8, SirB)	K02302 | K02304 | K03794
17	Siroheme decarboxylase (AhbAB)	K22225
18	Fe-coproporphyrin III synthase (AhbC)	K22226

Table S2. Composition of the medium used in this study

Components	Compound(s)	Final concentration
Ammonium	NH_4Cl	10 μM
Phosphate	KH_2PO_4	10 μM
Trace metals	$FeCl_3 \cdot 6H_2O$	117 nM
	$MnCl_2 \cdot 4H_2O$	9 nM
	$ZnSO_4 \cdot 7H_2O$	800 pM
	$CoCl_2 \cdot 6H_2O$	500 pM
	$Na_2MoO_4 \cdot 2H_2O$	300 pM
	Na_2SeO_3	1 nM
	$NiCl_2 \cdot 6H_2O$	1 nM
Vitamin mixture	Thiamine·HCl	59 nM
	Niacin	81 nM
	Ca-Pantothenate	84 nM
	Pyridoxine	59 nM
	Biotin	409 pM
	Folic acid	453 pM
	Vitamin B12	70 pM
	Myo-inositol	555 nM
	p-Aminobenzoic Acid	7 nM
Carbon mixture	Pyruvate	50 μM
	D-Glucose	5 μM
	N-Acetyl-D-glucosamine	5 μM
	D-Ribose	5 μM
	Methyl alcohol	5 μM
20 proteinogenic amino acid mixture	Each amino acid	100 nM, each

Table S3. Distribution of heme uptake systems HemTUV and DppABCDF in 20 complete acI genomes

Tribe	Genome	HemTUV	DppABCDF
A1	'*Candidatus* Planktophila rubra' IMCC25003	HemTUV	DppABCDF
A1	'*Candidatus* Planktophila dulcis' MMS-21-155		DppABCDF
A1	'*Candidatus* Planktophila dulcis' MMS-IA-53		DppABCDF
A1	'*Candidatus* Planktophila dulcis' MMS-IIA-65		DppABCDF
A1	'*Candidatus* Planktophila sulfonica' MMS-IA-56	HemTUV	DppABCDF
A1	'*Candidatus* Planktophila versatilis' MMS-IA-105	HemTUV	
A1	'*Candidatus* Planktophila versatilis' MMS-IA-79	HemTUV	
A1	'*Candidatus* Planktophila versatilis' MMS-IIB-142	HemTUV	
A1	'*Candidatus* Planktophila versatilis' MMS-IIB-76	HemTUV	
A2	'*Candidatus* Planktophila limnetica' MMS-VB-114	HemTUV	2 DppABCDF
A4	'*Candidatus* Planktophila aquatilis' IMCC26103		2 DppABCDF
A4	'*Candidatus* Planktophila lacus' MMS-21-148	HemTUV	DppABCDF
A4	'*Candidatus* Planktophila lacus' MMS-IIB-106	HemTUV	DppABCDF
A4	'*Candidatus* Planktophila lacus' MMS-IIB-60	HemTUV	DppABCDF
A7	Actinobacteria bacterium IMCC19121		DppABCDF
A7	'*Candidatus* Planktophila vernalis' MMS-IIA-15		DppABCD
B1	'*Candidatus* Nanopelagicus limnes' MMS-21-122		DppABCDF
B1	'*Candidatus* Nanopelagicus limnes' MMS-21-160		DppABCDF
B1	'*Candidatus* Nanopelagicus limnes' MMS-IIB-91		DppABCDF
C1	Actinobacteria bacterium IMCC26077	HemTUV	DppABCDF

Table S4. List of the microbial group names used in this study and their GTDB taxonomy

Group name	Alternative	GTDB taxonomy	Reference*
Limnohabitans		d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__Limnohabitans	(3)
Polynucleobacter		d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__Polynucleobacter	(3)
OM43		d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Methylophilaceae;g__BACL14	(4)
Methylopumilus	LD28	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Methylophilaceae;g__Methylopumilus	(4)
Halieaceae	OM60/NOR5	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Halieaceae	(5)
Spongiibacteraceae	BD1-7	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Spongiibacteraceae	(5)
Porticoccaceae	SAR92	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Porticoccaceae	(5)
Alteromonas		d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacterales;f__Alteromonadaceae;g__Alteromonas	(6, 7)
Thioglobus	SUP05	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Thiomicrospirales;f__Thioglobaceae;g__Thioglobus	(8)
SAR86		d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__SAR86	(9)
Pelagibacterales	SAR11, LD12	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Pelagibacterales	(10)
HIMB59	AEGEAN-169	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__HIMB59	(10)
Puniceispirillales	SAR116	d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Puniceispirillales	(9)
Rhodobacteraceae		d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodobacterales;f__Rhodobacteraceae	(6, 7, 11)
SAR324	Marine group B	d__Bacteria;p__SAR324	(12, 13)
Flavobacteriaceae		d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Flavobacteriales;f__Flavobacteriaceae	(7, 14)
Marinimicrobia	SAR406, Marine group A	d__Bacteria;p__Marinisomatota	(9, 15)
OM1		d__Bacteria;p__Actinobacteriota;c__Acidimicrobiia;o__TMED189	(16, 17)
acIV		d__Bacteria;p__Actinobacteriota;c__Acidimicrobiia;o__Microtrichales;f__Ilumatobacteraceae;g__UBA3006 d__Bacteria;p__Actinobacteriota;c__Acidimicrobiia;o__Microtrichales;f__Ilumatobacteraceae;g__BACL27 d__Bacteria;p__Actinobacteriota;c__Acidimicrobiia;o__Microtrichales;f__Ilumatobacteraceae;g__UBA2093 d__Bacteria;p__Actinobacteriota;c__Acidimicrobiia;o__Microtrichales;f__Ilumatobacteraceae;g__Casp-actino8	(3)
Rhodoluna		d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Actinomycetales;f__Microbacteriaceae;g__Rhodoluna	(3)
acI		d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Nanopelagicales;f__Nanopelagicaceae	(3)
SAR202		d__Bacteria;p__Chloroflexota;c__Dehalococcoidia;o__UBA1151 d__Bacteria;p__Chloroflexota;c__Dehalococcoidia;o__SAR202 d__Bacteria;p__Chloroflexota;c__Dehalococcoidia;o__UBA3495 d__Bacteria;p__Chloroflexota;c__Dehalococcoidia;o__UBA2985 d__Bacteria;p__Chloroflexota;c__Dehalococcoidia;o__GCA-2717565 d__Bacteria;p__Chloroflexota;c__Dehalococcoidia;o__UBA2963 d__Bacteria;p__Chloroflexota;c__Dehalococcoidia;o__UBA6952 d__Bacteria;p__Chloroflexota;c__Dehalococcoidia;o__UBA1127 d__Bacteria;p__Chloroflexota;c__Dehalococcoidia;o__SAR202-VII-2	(18)
CL500-11		d__Bacteria;p__Chloroflexota;c__Anaerolineae;o__Anaerolineales;f__UBA11657	(19)

Patescibacteria	Candidate phyla radiation	p__Patescibacteria	(20)
Asgardarchaeota	Asgard archaea	p__Asgardarchaeota	(21, 22)
Marine Group IIa		d__Archaea;p__Thermoplasmatota;c__Poseidoniia;o__Poseidoniales;f__Poseidoniaceae	(23, 24)
Marine Group IIb		d__Archaea;p__Thermoplasmatota;c__Poseidoniia;o__Poseidoniales;f__Thalassoarchaeaceae	(23, 24)
Marine Group III		d__Archaea;p__Thermoplasmatota;c__Poseidoniia;o__MGIII	(23, 24)
Thaumarchaeota	Marine Group I	d__Archaea;p__Crenarchaeota;c__Nitrososphaeria	(23, 24)

* References supporting the widespread distribution of each of the microbial groups listed in this table in aquatic habitats.

Table S5. Completeness and contamination of the representative genomes for species clusters of GTDB (R89)

	Contamination (%)	Completeness (%)					
		≥90	80–90	70–80	60–70	50–60	<50
Isolates (14,406)	<5	13,911	163	73	35	23	6
	5–10	162	4	1	0	0	0
	≥10	24	3	1	0	0	0
MAG/SAGs (10,300)	<5	3,852	2,337	1,813	1,197	661	3
	5–10	267	142	18	0	3	0
	≥10	7	0	0	0	0	0
All (24,706)	<5	17,763	2,500	1,886	1,232	684	9
	5–10	429	146	19	0	3	0
	≥10	31	3	1	0	0	0

* High-quality (completeness ≥90% and contamination <5%) and medium-quality (completeness ≥50% and contamination <10%) genomes are indicated in yellow and green background, respectively.

Dataset S1 (separate file). List of genomes analyzed in this study and their features including GTDB ID (GenBank Assembly accession number), GTDB taxonomy, genome characteristics, and heme biosynthetic pathway completeness.

SI References

1. Haase, I. et al. Enzymes from the haloacid dehalogenase (HAD) superfamily catalyse the elusive dephosphorylation step of riboflavin biosynthesis. *Chembiochem* **14**, 2272-2275 (2013).
2. Sarge, S. et al. Catalysis of an essential step in vitamin B2 biosynthesis by a consortium of broad spectrum hydrolases. *Chembiochem* **16**, 2466-2469 (2015).
3. R. J. Newton, S. E. Jones, A. Eiler, K. D. McMahon & S. Bertilsson. A guide to the natural history of freshwater lake bacteria. *Microbiol. Mol. Biol. Rev.* **75**, 14-49 (2011).
4. M. M. Salcher, D. Schaefle, M. Kaspar, S. M. Neuenschwander & R. Ghai. Evolution in action: habitat transition from sediment to the pelagial leads to genome streamlining in *Methylophilaceae*. *ISME J.* **13**, 2764-2777 (2019).
5. S. Spring, C. Scheuner, M. Goker & H. P. Klenk. A taxonomic framework for emerging groups of ecologically important marine gammaproteobacteria based on the reconstruction of evolutionary relationships using genome-scale data. *Front. Microbiol.* **6**, 281 (2015).

6. S. J. Giovannoni, J. C. Thrash & B. Temperton. Implications of streamlining theory for microbial ecology. *ISME J.* **8**, 1553-1565 (2014).
7. A. Buchan, G. R. LeCleir, C. A. Gulvik & J. M. Gonzalez. Master recyclers: features and functions of bacteria associated with phytoplankton blooms. *Nat. Rev. Microbiol.* **12**, 686-698 (2014).
8. V. Shah, B. X. Chang & R. M. Morris. Cultivation of a chemoautotroph from the SUP05 clade of marine bacteria that produces nitrite and consumes ammonium. *ISME J.* **11**, 263-271 (2017).
9. B. K. Swan et al. Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc. Natl. Acad. Sci. USA* **110**, 11463-11468 (2013).
10. J. Grote et al. Streamlining and core genome conservation among highly divergent members of the SAR11 clade. *mBio* **3**, e00252-12 (2012).
11. M. Simon et al. Phylogenomics of *Rhodobacteraceae* reveals evolutionary adaptation to marine and non-marine habitats. *ISME J.* **11**, 1483-1499 (2017).
12. C. S. Sheik, S. Jain & G. J. Dick. Metabolic flexibility of enigmatic SAR324 revealed through metagenomics and metatranscriptomics. *Environ. Microbiol.* **16**, 304-317 (2014).
13. B. K. Swan et al. Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science* **333**, 1296-1300 (2011).
14. H. Zhang et al. Repeated evolutionary transitions of *Flavobacteria* from marine to non-marine habitats. *Environ. Microbiol.* **21**, 648-666 (2019).
15. E. W. Getz, S. S. Tithi, L. Q. Zhang & F. O. Aylward. Parallel evolution of genome streamlining and cellular bioenergetics across the marine radiation of a bacterial phylum. *mBio* **9**, e01089-18 (2018).
16. R. M. Morris et al. Temporal and spatial response of bacterioplankton lineages to annual convective overturn at the Bermuda Atlantic time-series study site. *Limnol. Oceanogr.* **50**, 1687-1696 (2005).
17. C. M. Mizuno, F. Rodriguez-Valera & R. Ghai. Genomes of planktonic *Acidimicrobiales*: widening horizons for marine *Actinobacteria* by metagenomics. *mBio* **6**, e02083-14 (2015).
18. J. H. W. Saw et al. Pangenomics analysis reveals diversification of enzyme families and niche specialization in globally abundant SAR202 bacteria. *mBio* **11**, e02975-19 (2020).
19. Y. Okazaki, M. M. Saicher, C. Callieri & S. Nakano. The broad habitat spectrum of the CL500-11 lineage (phylum *Chloroflexi*), a dominant bacterioplankton in oxygenated hypolimnia of deep freshwater lakes. *Front. Microbiol.* **9**, 2891 (2018).

20. C. T. Brown et al. Unusual biology across a group comprising more than 15% of domain *Bacteria*. *Nature* **523**, 208-U173 (2015).

21. Y. Liu et al. Expanded diversity of Asgard archaea and their relationships with eukaryotes. *Nature* https://doi.org/10.1038/s41586-021-03494-3 (2021).

22. F. MacLeod, G. S. Kindler, H. L. Wong, R. Chen & B. P. Burns. Asgard archaea: diversity, function, and evolutionary implications in a range of microbiomes. *AIMS Microbiol.* **5**, 48-61 (2019).

23. P. S. Adam, G. Borrel, C. Brochier-Armanet & S. Gribaldo. The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. *ISME J.* **11**, 2407-2425 (2017).

24. R. Massana, E. F. DeLong & C. Pedros-Alio. A few cosmopolitan phylotypes dominate planktonic archaeal assemblages in widely different oceanic provinces. *Appl. Environ. Microbiol.* **66**, 1777-1787 (2000).