Supplementary Information: Genomics of soil depth niche partitioning in the Thaumarchaeota family Gagatemarchaeaceae

Paul O. Sheridan[1,2], Yiyu Meng[1], Tom A. Williams[3], Cécile Gubry-Rangin[1,#]

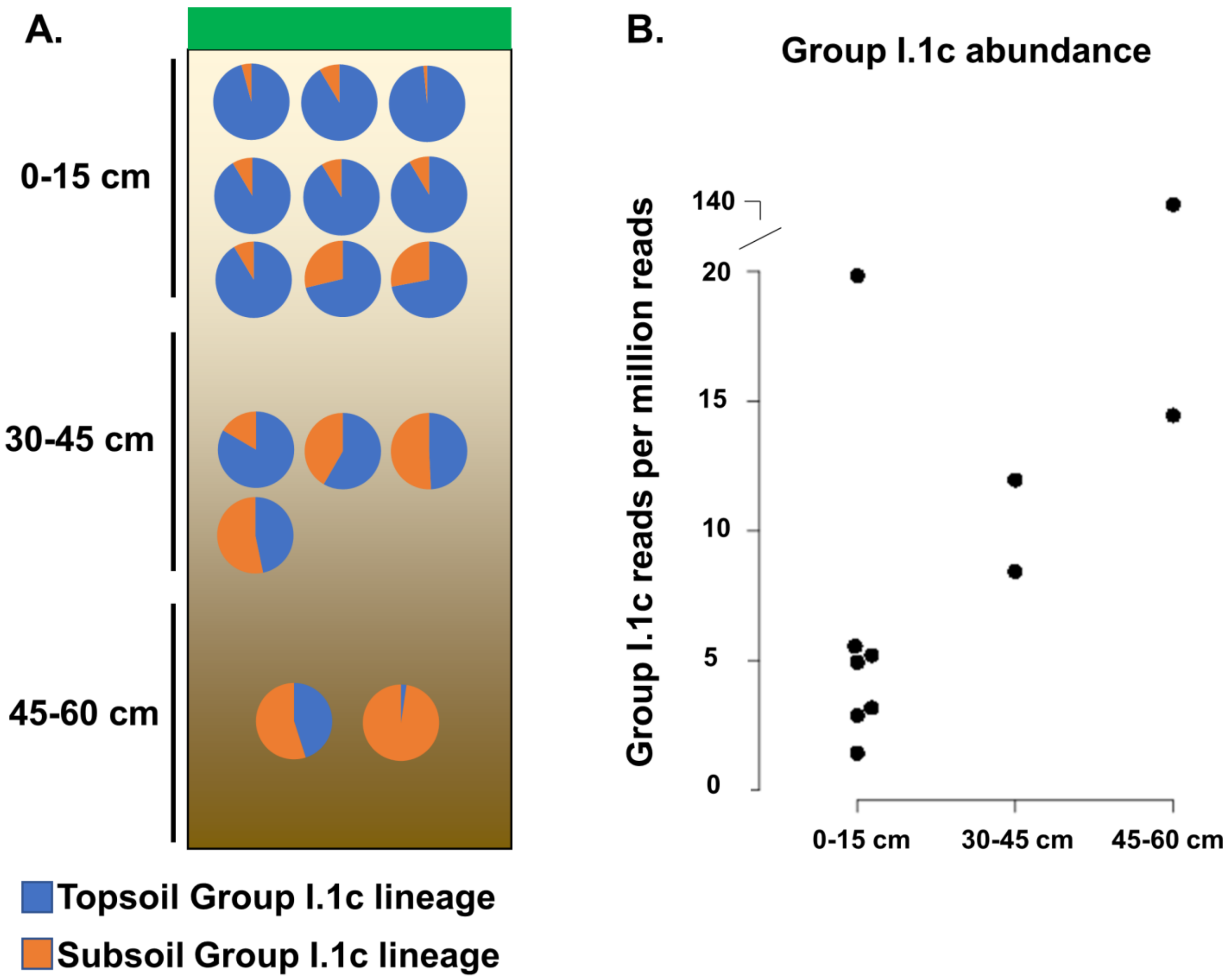
Supplementary Note 1: Classifications

Candidatus "*Gagatemarchaeum stordalenia*" (sp. nov., gen. nov). "Gagatem" refers to its prevalence in peat environments. The name "stordalenia" originates from the geographical origin (Stordalen Mire, Sweden) of the reference (type) genome, bog-1369. It encodes genes for aerobic respiration and likely uses organic substrates, such as carbohydrates, peptides and fatty acids for organoheterotrophic growth. It is currently not cultured and known only from environmental sequencing. Genomes of this genera possess a high GC-content (around 60%) and genome size of its members range from 2.2-3.4 Mb.

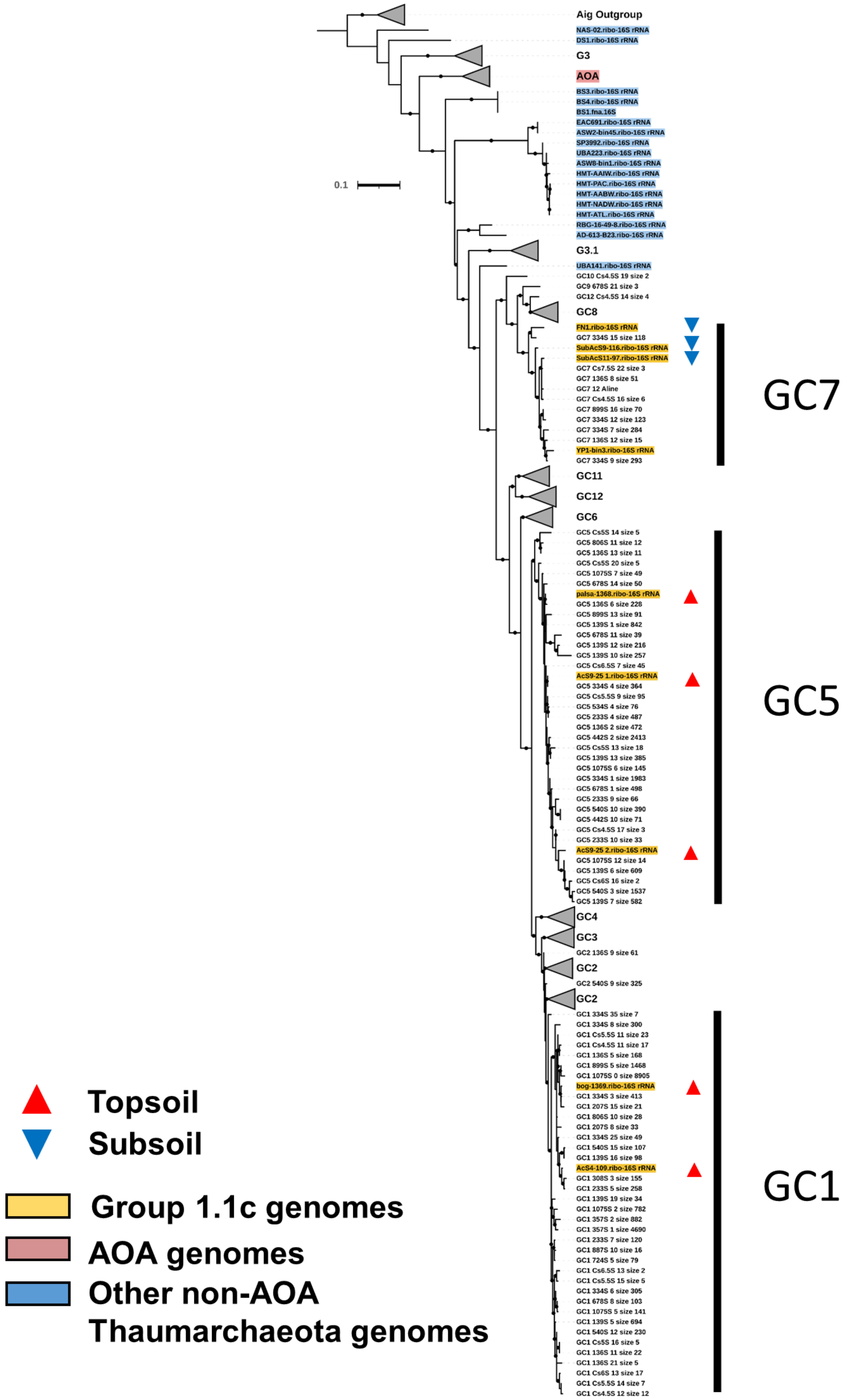
Candidatus "*Subgagatemarchaeum marcellia*" (sp. nov., gen. nov). "Subgagatem" refers to its prevalence in subsoils and peat environments. The name "'marcellia" originates from the geographical origin (Marcell Experimental Forest, MN, USA) of the reference (type) genome, Fn1. It likely uses organic substrates, such as carbohydrates, peptides and fatty acids for organoheterotrophic growth. While the type genome, Fn1, encodes the genes for microaerophilic respiration cytochrome bd ubiquinol oxidase, the majority of studied genomes from this genera encode the aerobic respiration heme-copper oxygen reductases. It is currently not cultured and only known from environmental sequencing. Genomes of this genera possess a high GC-content (around 58%) and genome size of its members range form 1.2-2.5 Mb.

Description of Gagatemarchaeaceae (fam. nov). This family was previously referred as the Group I.1c Thaumarchaeota. Description is the same as for the genus *Gagatemarchaeum*. Suff. -aceae, ending to denote family. Type genus *Gagatemarchaeum* gen. nov.

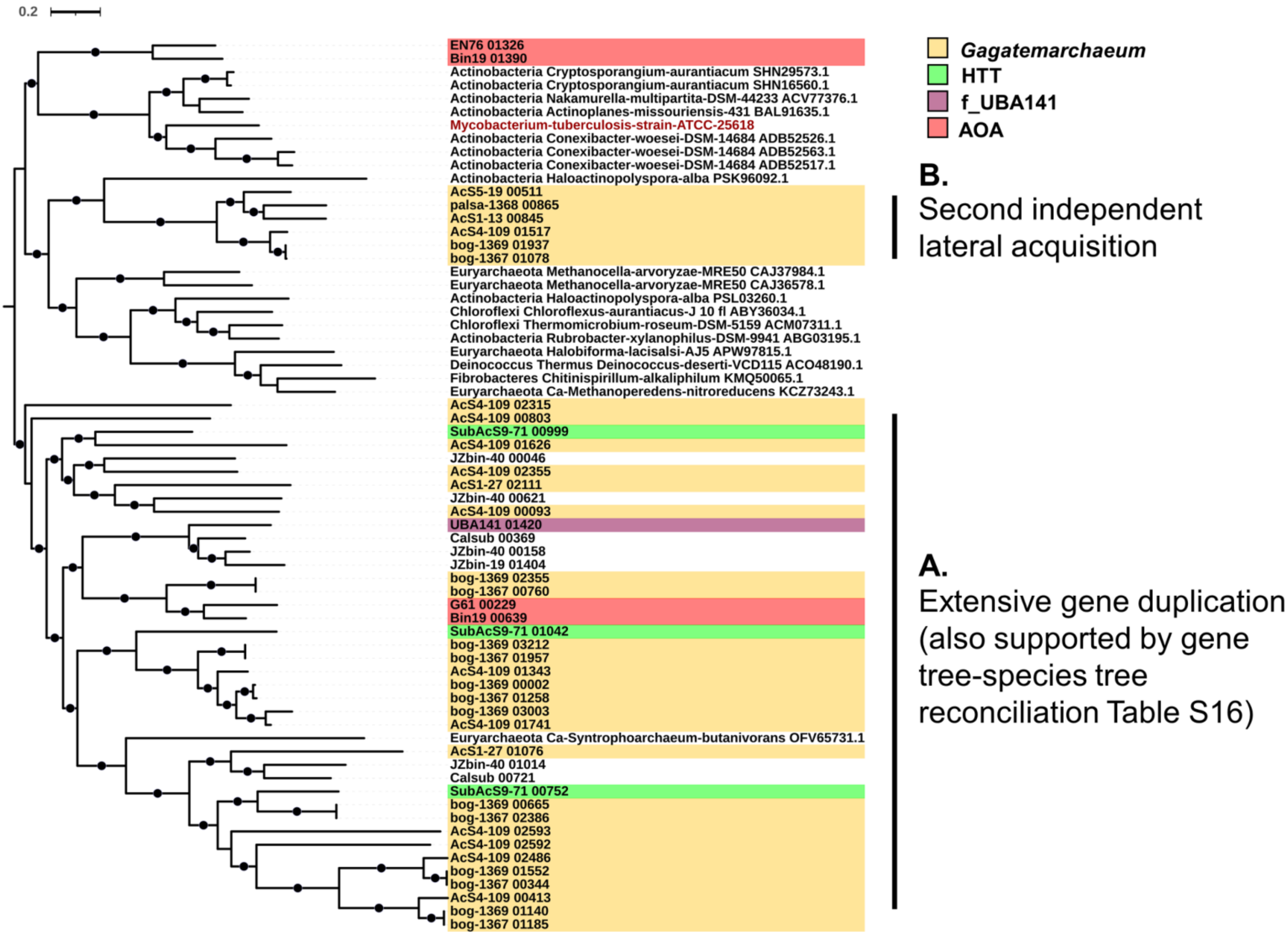
Supplementary Figures



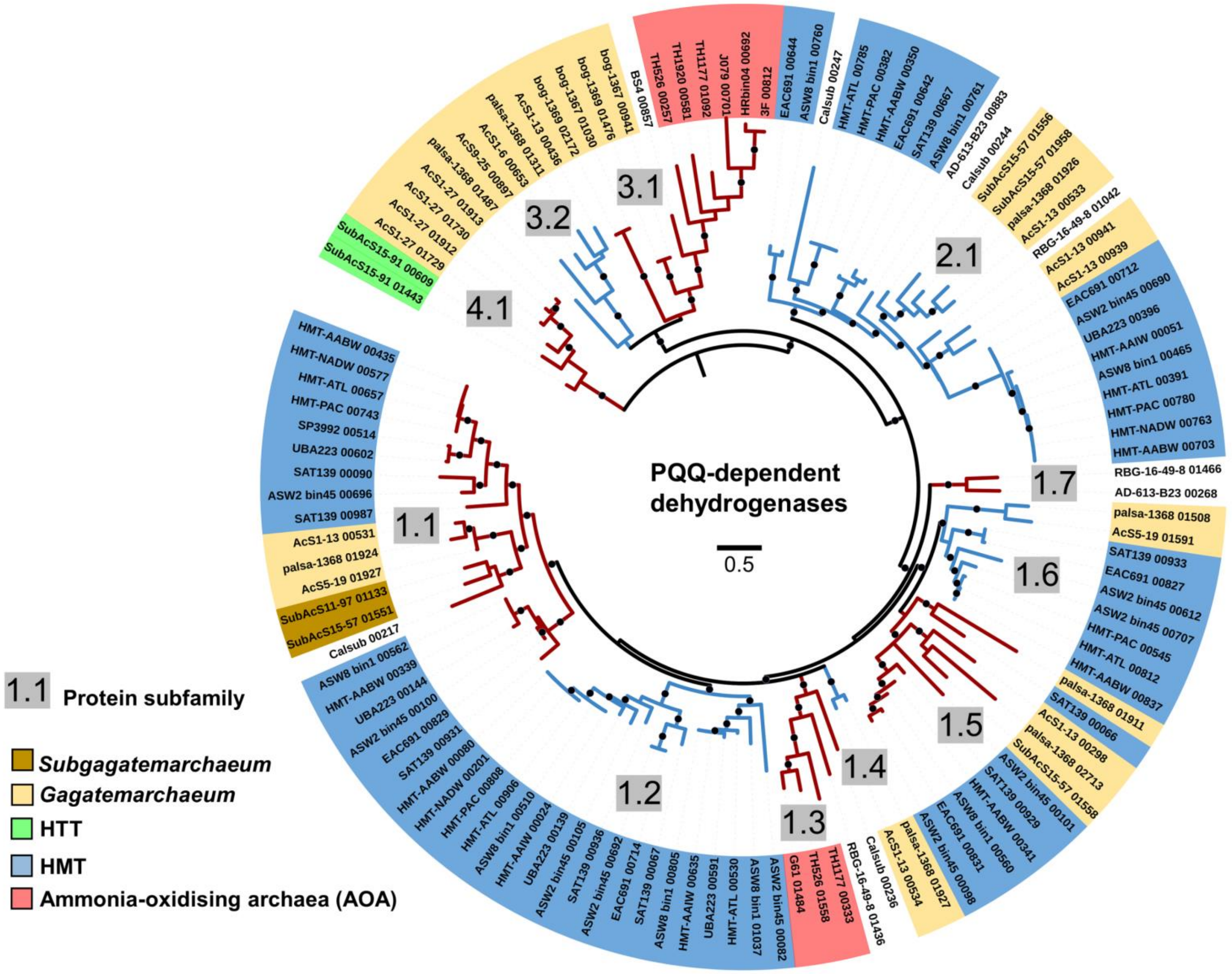
Supplementary Fig. 1. Depth distribution of soil Group I.1c lineages. (A) The Group I.1c community composition changes from Topsoil lineage dominated to Subsoil lineage dominated with increasing depth, and (B) the overall abundance of Group I.1c increases with depth. Metagenomic reads were recruited to genomes of the Topsoil and Subsoil lineages. Source data are provided as a Source Data file..



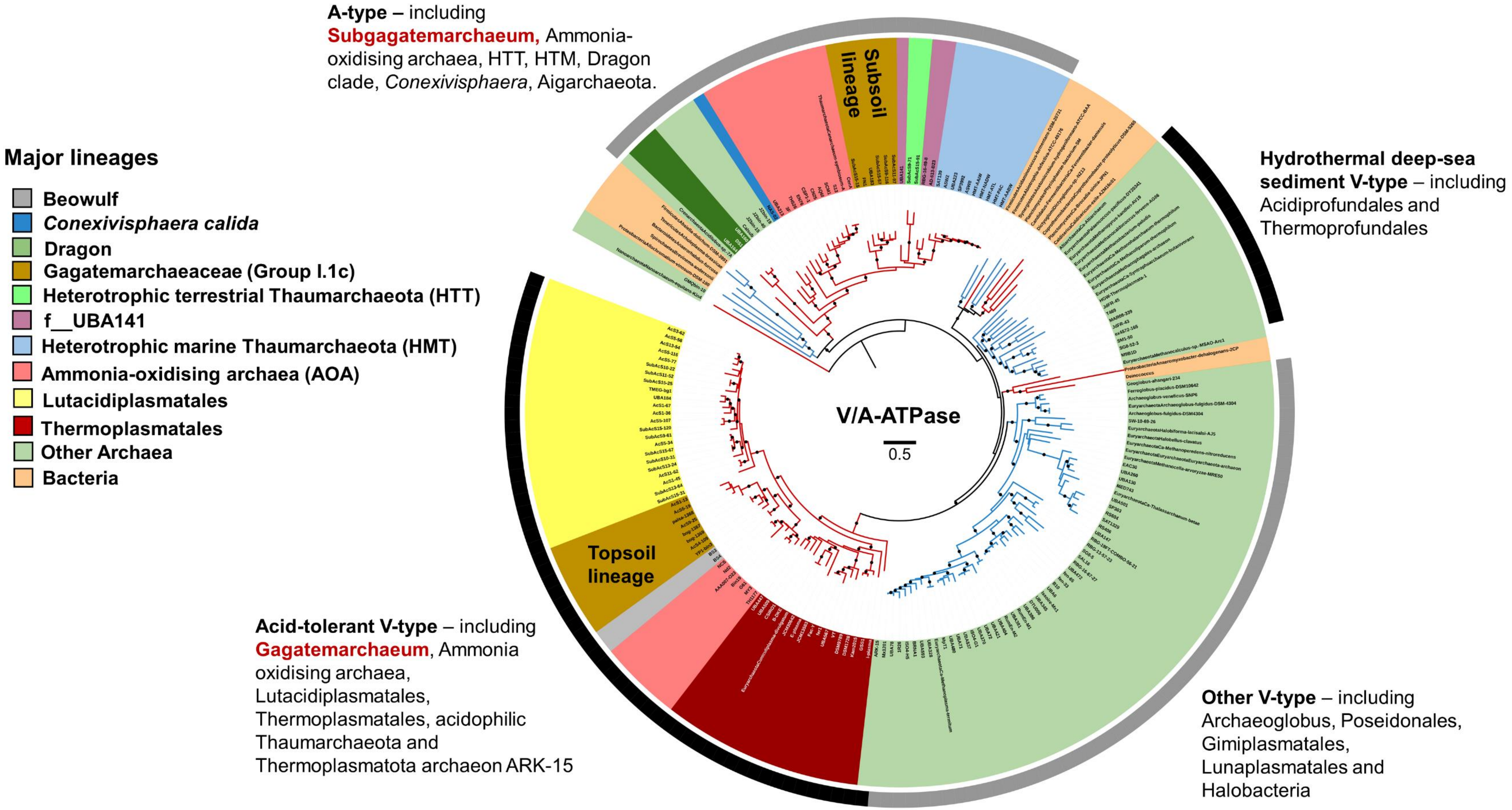
Supplementary Fig. 2. 16S rRNA gene tree of Group I.1c and related sequences. This maximum-likelihood tree was created using 16S rRNA genes extracted from the studied genomes and combined with the 16S rRNA database of soil Thaumarchaeota presented in Vico Oton et al 2016[1] (587 column alignment). Dots indicate branches with >70% of 1,000 UFBoot replicates.


0.2
EN76 01326
Bin19 01390
Actinobacteria Cryptosporangium-aurantiacum SHN29573.1
Actinobacteria Cryptosporangium-aurantiacum SHN16560.1
Actinobacteria Nakamurella-multipartita-DSM-44233 ACV77376.1
Actinobacteria Actinoplanes-missouriensis-431 BAL91635.1
Mycobacterium-tuberculosis-strain-ATCC-25618
Actinobacteria Conexibacter-woesei-DSM-14684 ADB52526.1
Actinobacteria Conexibacter-woesei-DSM-14684 ADB52563.1
Actinobacteria Conexibacter-woesei-DSM-14684 ADB52517.1
Actinobacteria Haloactinopolyspora-alba PSK96092.1
AcS5-19 00511
palsa-1368 00865
AcS1-13 00845
AcS4-109 01517
bog-1369 01937
bog-1367 01078
Euryarchaeota Methanocella-arvoryzae-MRE50 CAJ37984.1
Euryarchaeota Methanocella-arvoryzae-MRE50 CAJ36578.1
Actinobacteria Haloactinopolyspora-alba PSL03260.1
Chloroflexi Chloroflexus-aurantiacus-J 10 fl ABY36034.1
Chloroflexi Thermomicrobium-roseum-DSM-5159 ACM07311.1
Actinobacteria Rubrobacter-xylanophilus-DSM-9941 ABG03195.1
Euryarchaeota Halobiforma-lacisalsi-AJ5 APW97815.1
Deinococcus Thermus Deinococcus-deserti-VCD115 ACO48190.1
Fibrobacteres Chitinispirillum-alkaliphilum KMQ50065.1
Euryarchaeota Ca-Methanoperedens-nitroreducens KCZ73243.1
AcS4-109 02315
AcS4-109 00803
SubAcS9-71 00999
AcS4-109 01626
JZbin-40 00046
AcS4-109 02355
AcS1-27 02111
JZbin-40 00621
AcS4-109 00093
UBA141 01420
Calsub 00369
JZbin-40 00158
JZbin-19 01404
bog-1369 02355
bog-1367 00760
G61 00229
Bin19 00639
SubAcS9-71 01042
bog-1369 03212
bog-1367 01957
AcS4-109 01343
bog-1369 00002
bog-1367 01258
bog-1369 03003
AcS4-109 01741
Euryarchaeota Ca-Syntrophoarchaeum-butanivorans OFV65731.1
AcS1-27 01076
JZbin-40 01014
Calsub 00721
SubAcS9-71 00752
bog-1369 00665
bog-1367 02386
AcS4-109 02593
AcS4-109 02592
AcS4-109 02486
bog-1369 01552
bog-1367 00344
AcS4-109 00413
bog-1369 01140
bog-1367 01185
Gagatemarchaeum
HTT
f_UBA141
AOA
B.
Second independent lateral acquisition
A.
Extensive gene duplication (also supported by gene tree-species tree reconciliation Table S16)

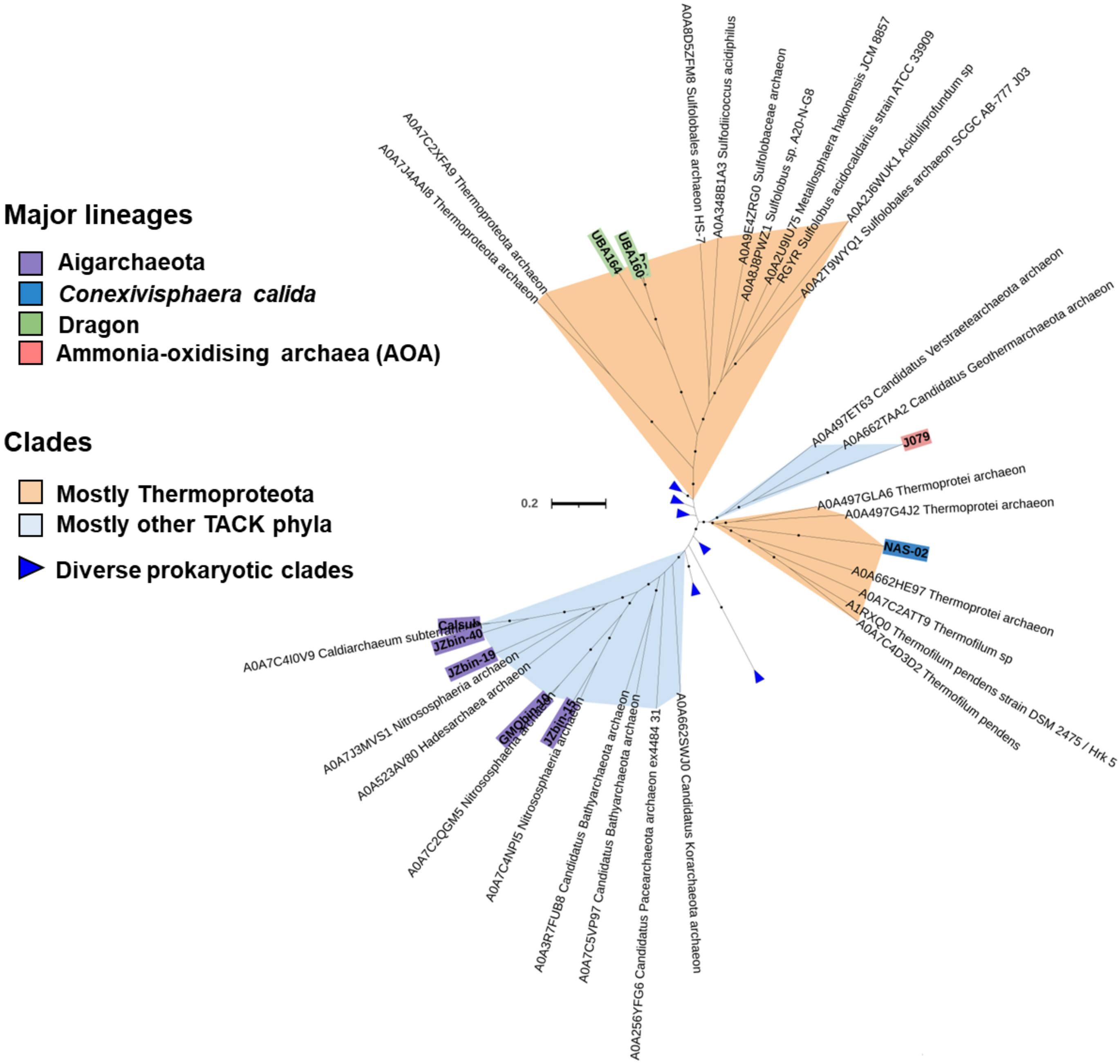
Supplementary Fig. 3. Phylogeny of F420-dependent glucose-6-phosphate dehydrogenase. This gene family was present in multiple copies in the topsoil Group I.1c genomes, owing to (A) extensive gene duplication throughout the evolution of *Gagatemarchaeum* and (B) a second independent lateral acquisition. The leaf in red is the experimentally validated gene. Dots indicate branches with >70% of 1,000 UFBoot replicates. The ML tree was created using LG+R4 and rooted with minimal ancestor deviation (MAD).



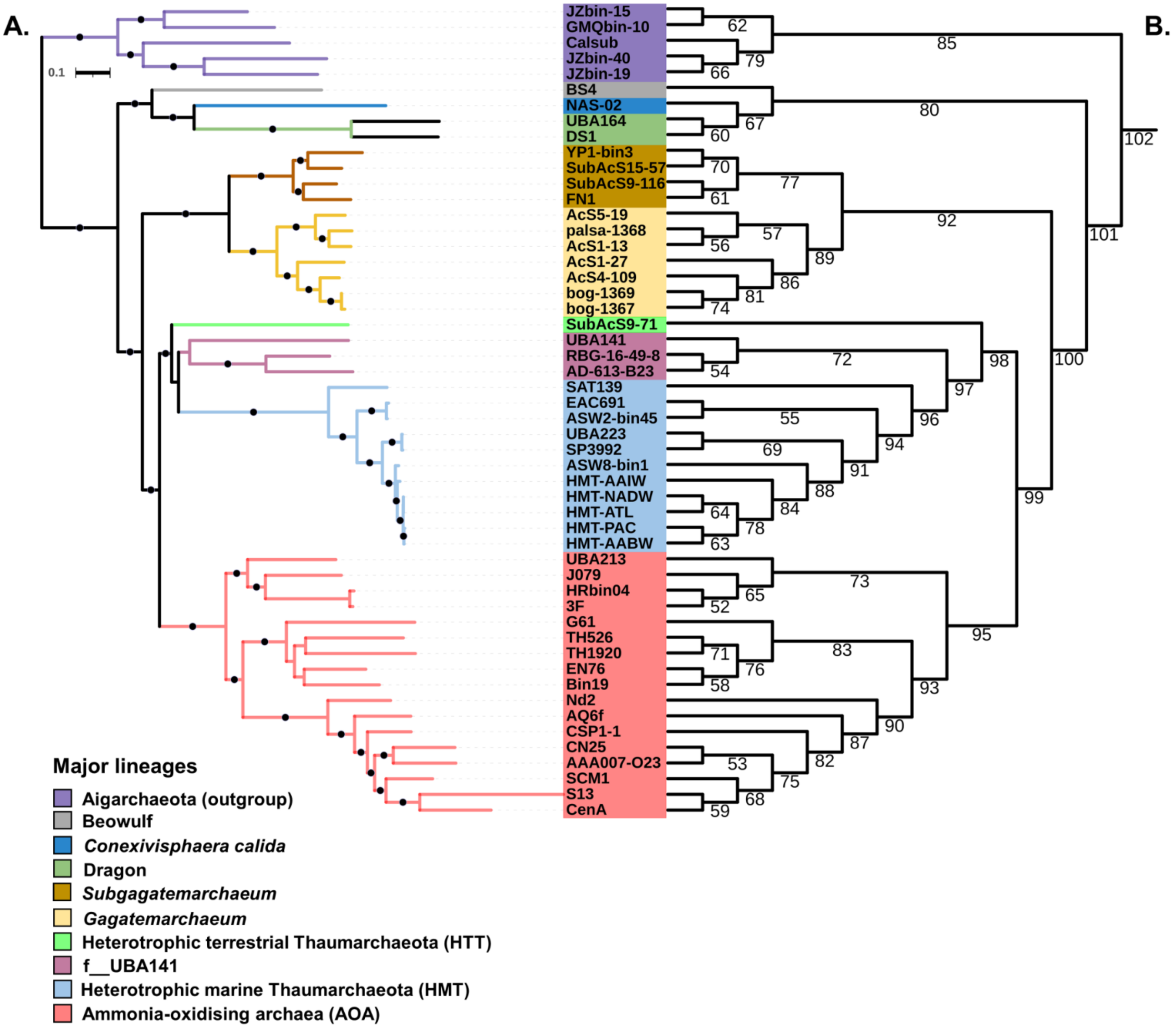
Supplementary Fig. 4. Phylogeny of Thaumarchaeota PQQ-dependent dehydrogenases. Dots indicate branches with >70% of 1,000 UFBoot replicates**.** The alternating blue and red branches indicate different subfamilies as determined by average pairwise distance between leaves. Tree was rooted using minimal ancestor deviation (MAD).


A-type – including
Subgagatemarchaeum, Ammonia-oxidising archaea, HTT, HTM, Dragon clade, Conexivisphaera, Aigarchaeota.
Subsoil lineage
Major lineages
Beowulf
Conexivisphaera calida
Dragon
Gagatemarchaeaceae (Group I.1c)
Heterotrophic terrestrial Thaumarchaeota (HTT)
f__UBA141
Heterotrophic marine Thaumarchaeota (HMT)
Ammonia-oxidising archaea (AOA)
Lutacidiplasmatales
Thermoplasmatales
Other Archaea
Bacteria
Hydrothermal deep-sea sediment V-type – including
Acidiprofundales and Thermoprofundales
V/A-ATPase
0.5
Topsoil lineage
Acid-tolerant V-type – including
Gagatemarchaeum, Ammonia oxidising archaea, Lutacidiplasmatales, Thermoplasmatales, acidophilic Thaumarchaeota and Thermoplasmatota archaeon ARK-15
Other V-type – including
Archaeoglobus, Poseidonales, Gimiplasmatales, Lunaplasmatales and Halobacteria

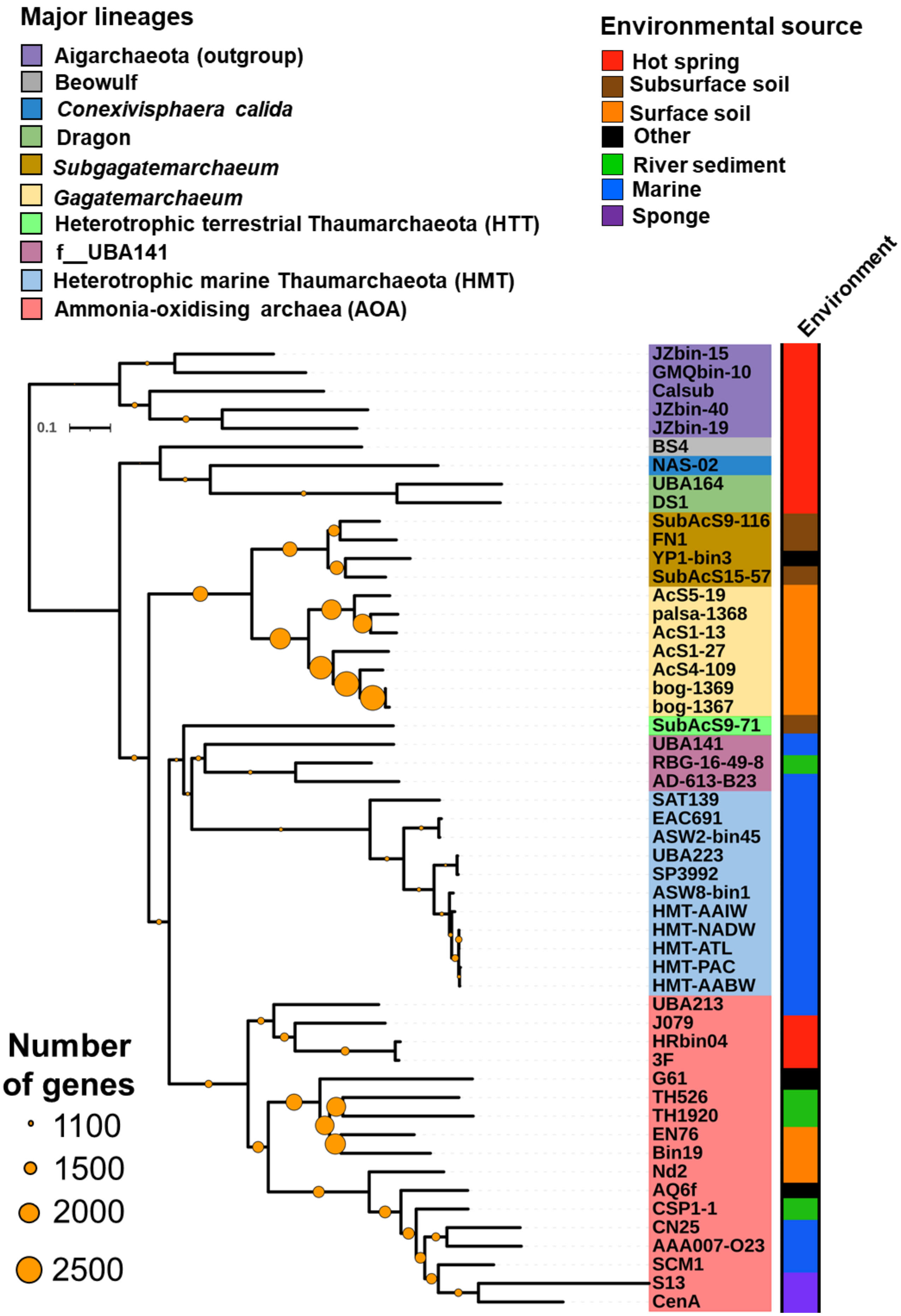
Supplementary Fig. 5. Phylogeny of the V/A-ATPase. Sequences from *Gagatemarchaeum* clustered with acid tolerant archaea, whereas those of *Subgagatemarchaeum* clustered with HMT, HTT and non-acidophilic AOA. The three largest subunits of V/A-ATPase (*atpA*, *atpB* and *atpI*) were individually aligned and then concatenated into a single partitioned supermatrix. A supermatrix tree was then estimated using the best fitting model for each partition and rooted using minimal ancestor deviation (MAD). Dots indicate branches with >70% of 1,000 UFBoot replicates. The alternating blue and red branches indicate different subfamilies of V/A-ATPase as determined by average pairwise distance between leaves.


Major lineages
🟪 Aigarchaeota
🟦 *Conexivisphaera calida*
🟩 Dragon
🟥 Ammonia-oxidising archaea (AOA)
Clades
🟧 Mostly Thermoproteota
⬜ Mostly other TACK phyla
▶ Diverse prokaryotic clades
0.2
A0A7C2XFA9 Thermoproteota archaeon
A0A7J4AAI8 Thermoproteota archaeon
UBA164
UBA160
A0A8D5ZFM8 Sulfolobales archaeon HS-7
A0A348B1A3 Sulfodiicoccus acidiphilus
A0A9E4ZRG0 Sulfolobaceae archaeon
A0A8J8PWZ1 Sulfolobus sp. A20-N-G8
A0A2U9IU75 Metallosphaera hakonensis JCM 8857
RGYR Sulfolobus acidocaldarius strain ATCC 33909
A0A2J6WUK1 Aciduliprofundum sp
A0A2T9WYQ1 Sulfolobales archaeon SCGC AB-777 J03
A0A497ET63 Candidatus Verstraetearchaeota archaeon
A0A662TAA2 Candidatus Geothermarchaeota archaeon
J079
A0A497GLA6 Thermoprotei archaeon
A0A497G4J2 Thermoprotei archaeon
NAS-02
A0A662HE97 Thermoprotei archaeon
A0A7C2ATT9 Thermofilum sp
A1RXQ0 Thermofilum pendens strain DSM 2475 / Hrk 5
A0A7C4D3D2 Thermofilum pendens
A0A7C4I0V9 Caldiarchaeum subterraneum
Calsub
JZbin-40
JZbin-19
A0A7J3MVS1 Nitrososphaeria archaeon
A0A523AV80 Hadesarchaea archaeon
GMQbin-10
A0A7C2QGM5 Nitrososphaeria archaeon
JZbin-15
A0A7C4NPI5 Nitrososphaeria archaeon
A0A3R7FUB8 Candidatus Bathyarchaeota archaeon
A0A7C5VP97 Candidatus Bathyarchaeota archaeon
A0A256YFG6 Candidatus Pacearchaeota archaeon ex4484 31
A0A662SWJ0 Candidatus Korarchaeota archaeon

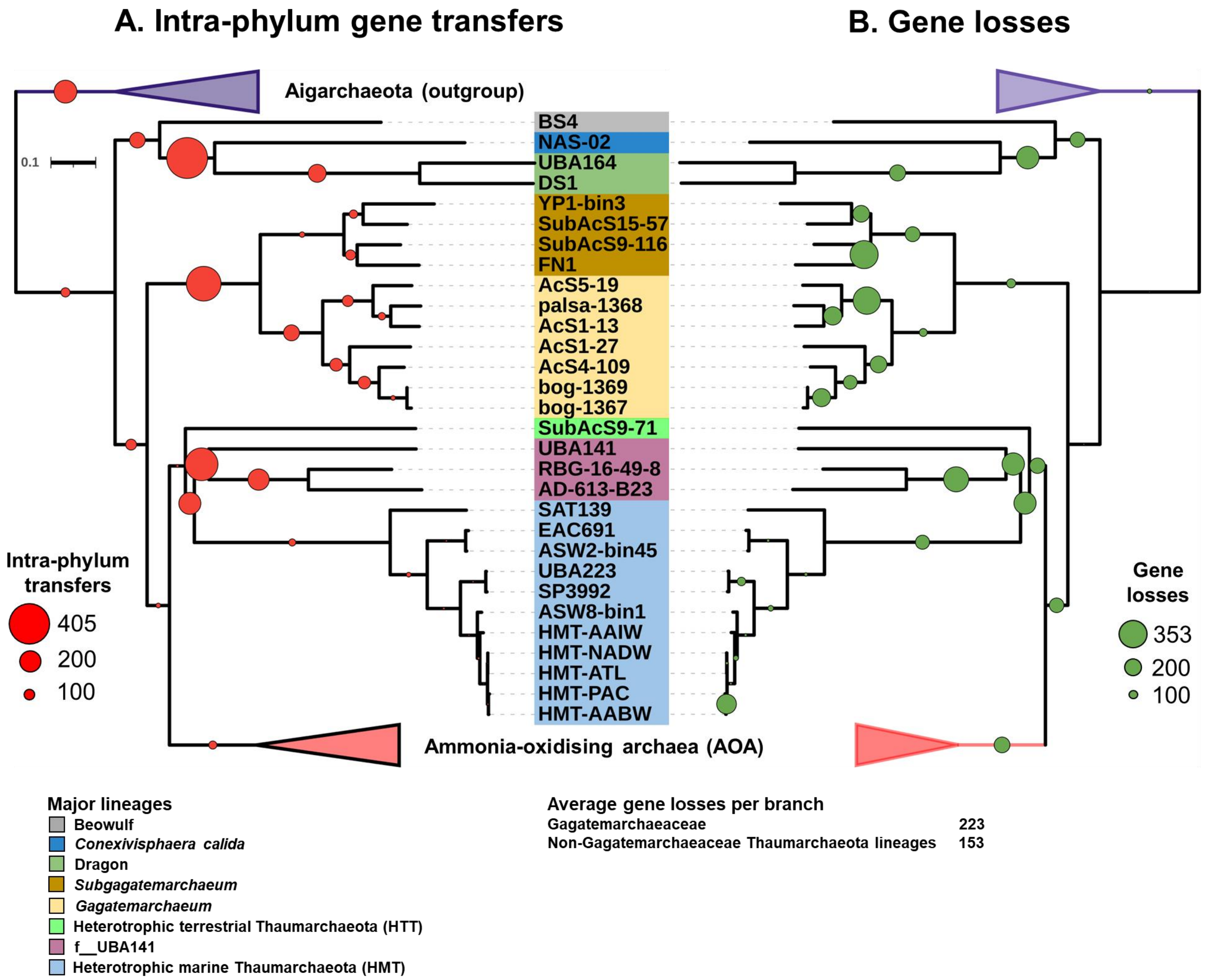
Supplementary Fig. 6. Phylogeny of Thaumarchaeota reverse gyrase, rgy. Dots indicate branches with >70% of 1,000 UFBoot replicates**.** The ML tree was predicted using the LG+F+R10 model.



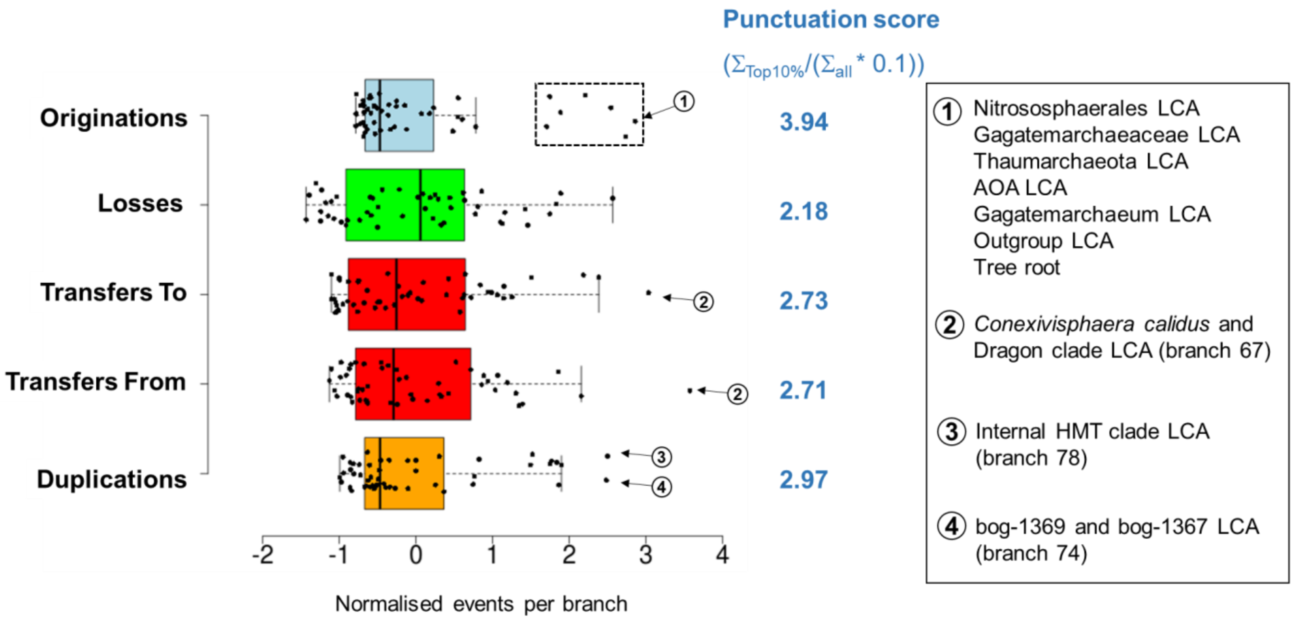
Supplementary Fig. 7. Phylogeny of higher-quality genomes (A) and corresponding branch numbers (B). The ML tree was predicted with the LG+C60+G+F model. Dots indicate branches with >95% of 2,000 UFBoot and 1,000 SH-aLRT replicates**.**



Supplementary Fig. 8. Genome expansion transitioning into terrestrial environments. Copy number of ancestor gene content reconstructions estimated using gene tree-species tree reconciliation are indicated.



Supplementary Fig. 9. Acquisition of genes by intra-phyla transfer and gene losses along non-AOA lineages. The quantitative and qualitative predictions of intra-phyla gene transfers (A) and gene loss (B) were estimated across the Thaumarchaeota history using a gene tree-species tree reconciliation approach. Scale numbers indicate the range of the predicted number of events for a given mechanism and circle sizes are proportional to the number of events.



Supplementary Fig. 10. Distribution of mechanism of gene content change across Thaumarchaeota evolution. Boxplots represent normalised events per branch ((events per branch - μ)/σ) for each mechanism. Numbered circles mark ancestors (species tree branches) with the highest numbers of events. Horizontal lines within boxes indicate the medians, box boundaries indicate the 1st and 3rd quartiles, whiskers indicate the minima and maxima, and points beyond these whiskers are outliers. A punctuation score is measured for each given mechanism. It represents the sum of events in the 10% of branches with the highest event numbers divided by 10% of the sum of events into all branches ($\Sigma_{\text{events in top10\%}}/(\Sigma_{\text{events in all branches}}$* 0.1). Source data are provided as a Source Data file.

Supplementary References

1. Vico Oton, E., Quince, C., Nicol, G. W., Prosser, J. I. & Gubry-Rangin, C. Phylogenetic congruence and ecological coherence in terrestrial Thaumarchaeota. *The ISME journal* **10**, 85-96 (2016).