Phylogenomics and ancestral reconstruction of Korarchaeota reveals genomic adaptation to habitat switching

Guillaume Tahon[1*], Stephan Köstlbacher[1], Erik A. Pelve[2], Brett J. Baker[3, 4], Jimmy H. Saw[5], Laura Eme[6], Daniel Tamarit[7], Max Emil Schön[8] and Thijs J. G. Ettema[1*]

[1] *Laboratory of Microbiology, Wageningen University, 6708WE Wageningen, The Netherlands*

[2] *Department of Anatomy, Physiology and Biochemistry, Swedish University of Agricultural Sciences, Uppsala, Sweden*

[3] Department of Integrative Biology, University of Texas at Austin, USA

[4] Department of Marine Science, University of Texas at Austin, USA

[5] *Department of Biological Sciences, The George Washington University, Washington, DC 20052, USA*

[6] *Unité d'Ecologie, Systématique et Evolution, CNRS, Université Paris-Saclay, Paris, France*

[7] *Theoretical Biology and Bioinformatics, Utrecht University, 3584CH Utrecht, The Netherlands*

[8] *Department of Biomolecular Mechanisms, Max Planck Institute for Medical Research, 69120 Heidelberg, Germany*

*** Correspondence:**

Guillaume Tahon

guillaume.tahon@wur.nl

Thijs J. G. Ettema

thijs.ettema@wur.nl

Keywords

Archaea, microbial diversity, phylogeny, hyperthermophiles, hot spring, genome evolution

Abstract

Our knowledge of archaeal diversity and evolution has expanded rapidly in the past decade. However, hardly any genomes of the phylum Korarchaeota have been obtained due to the difficulty in accessing their natural habitats and – possibly – their limited abundance. As a result, many aspects of Korarchaeota biology, physiology and evolution remain enigmatic. Here, we expand this phylum with five high-quality metagenome-assembled genomes. This improved taxon sampling combined with sophisticated phylogenomic analyses robustly places Korarchaeota at the base of TACK and Asgard clades, revisiting the phylum's long-assumed position. Furthermore, we observe a clear split between terrestrial and marine thermal clades. Gene tree-aware ancestral reconstructions suggest that the last Korarchaeota common ancestor was a thermophilic autotroph. In contrast, Korarchaeaceae, the lineage where environmental transitions occurred, shifted towards a heterotrophic lifestyle. Terrestrial Korarchaeota gained many *cas* and CARF genes indicating they may need to manage viral infections. Together, our study provides new insights into these early diverging Archaea and suggests that gradual gene gain and loss shaped their adaptation to different thermal environments.

Importance

Korarchaeota are an ancient group of archaea, but their biology, physiology and evolution have remained obscure. Analysis of five novel Korarchaeota MAGs, and publicly available reference data provides robust phylogenomic evidence that Korarchaeota are placed at the base of Asgard archaea and TACK, revisiting the phylum's long-assumed position. Gene content reconstruction suggests a versatile thermophilic and autotrophic last Korarchaeota common ancestor. Environmental distribution surveying of public databases places all Korcharchaeota in thermophilic environments and indicates that their habitat is limited to hydrothermal vents and hot springs. Our modeling indicates at least two transitions linked to habitat switching between these environments in the evolutionary history of Korarchaeota. Both are linked to a significant alteration of the inferred ancestral gene content, including a shift towards a heterotrophic and potential scavenging lifestyle. Furthermore, hot spring Korarchaeota acquired various genes participating in resistance to viruses, suggesting they may need to manage frequent viral threats.

Introduction

Archaea were only recognized as a third domain of life in 1977 (1), and were for a long time made up of only two phyla – Crenarchaeota and Euryarchaeota – represented by very few cultured strains (2). The third archaeal phylum to be discovered and the first one to be proposed without a cultured representative was *Candidatus* Korarchaeota (sometimes referred to as Korarchaeia (3) – hereafter referred to as Korarchaeota) (4). Amplicon sequences of the lineage were first sampled from Obsidian Pool, a near-neutral pH hot spring in the Hayden Valley of Yellowstone National Park (5). Subsequent 16S rRNA gene surveys and phylogenetic analyses supported the separation of this strain from the classical archaeal groups Euryarchaeota and Crenarchaeota, inspiring the name Korarchaeota, for the early divergence of the taxon in archaeal evolution ("koros/kore" referring to a young man/woman in Greek) (4). In 2008, ultrathin filamentous cells isolated from an anaerobic enrichment culture inoculated with Obsidian Pool sediment allowed initial characterization of a Korarchaeon - "*Candidatus* Korarchaeum cryptofilum" OPF8 (6). Its 1.59 Mb complete genome suggested an anaerobic heterotrophic lifestyle relying on a simple mode of peptide fermentation for generating biomass and energy. Also, the inability to synthesize various essential biomolecules indicated a potential dependency on other community members.

With the development of high-throughput DNA sequencing technologies and phylogenetic approaches in the four decades since the recognition of Archaea, the number and diversity of lineages described has increased exponentially. Currently, Archaea comprise four main clades (i.e., TACK, DPANN, Asgard and Euryarchaeota) which encompass over 30 phylum-level lineages and at least 50,000 species (See G. Tahon et al. (7) for a detailed overview). However, Korarchaeota have not undergone nearly the same expansion. Nearly 30 years after its initial discovery, very few Korarchaeota genomes have been sequenced, all of which have been recovered from either marine hydrothermal vents or terrestrial hot springs (Table S2) (8). Furthermore, all but the *Ca.* Korarchaeum cryptofilum OPF8 genome were published after 2018. As a result, many aspects of Korarchaeota biology, physiology and evolution have remained obscure. In addition, their phylogenetic position has been difficult to establish precisely. Initially, the two original environmental korarchaeotal 16S rRNA gene sequences branched deeply within Crenarchaeota, as a sister group

101 to Eukarya or branching off before the common ancestor of Cren- and Euryarchaeota
102 (4). Subsequent phylogenomic analyses based on concatenated sets of marker
103 genes have not managed to unequivocally resolve the position of Korarchaeota.
104 Depending on the phylogenetic method and dataset used, the phylum was placed
105 either at the base of the TACK superphylum, as a sister group of Crenarchaeota
106 inside TACK or as a sister lineage to Asgardarchaeota (9-13). However, the varying
107 and frequently low-supported position of Korarchaeota within Archaea may be
108 explained in part because of the use of less sophisticated models of evolution at the
109 time, as well as by the lack of sequence representative of the korarchaeotal diversity.
110 The latter led to the existence of a long branch at the base of Korarchaeota, which is
111 well-known to yield phylogenetic reconstruction artefacts (14).
112 In this project, we expand the Korarchaeota genomic and taxonomic diversity with
113 five distinct draft genomes obtained from sediment collected from the Little Hot Creek
114 terrestrial hot spring (California, USA) and the Taketomi Island shallow submarine
115 hydrothermal field of the Southern Ryukyu Archipelago (Japan). Using sophisticated
116 phylogenomic analyses relying on multiple marker protein sets, we do not only
117 accurately resolve the placement of Korarchaeota within the archaeal domain, but
118 also the internal topology of the phylum. Furthermore, using gene tree-species tree
119 reconciliation methods, we infer the evolution gene content along the korarchaeotal
120 tree and describe how living in different marine and aquatic terrestrial thermal
121 habitats shaped the genetic composition of the Korarchaeota.
122

Materials and Methods

Sampling

Samples were collected from the Taketomi Island shallow submarine hydrothermal field (TIV) of the Southern Ryukyu Archipelago (N24°20.9' E124°06.10') (15, 16) and from the Little Hot Creek terrestrial hot spring (N37°41.436' W118°50.664') (17, 18).

Samples from Taketomi Island were collected as previously described by K. Zaremba-Niedzwiedzka et al. (19). Briefly, a sediment core was collected at 18 m depth, 5 m from the hydrothermal vent outflow by scuba diving, and material from the 10-15 cm section below the seafloor was extracted and stored at –25 °C and –80 °C. The core temperatures at 0, 20 and 30 cm below the sea floor were determined to be 31.0, 57.6 and 59.0 ˚C, respectively (RMT-2000; RIGO, Tokyo, Japan).

The Little Hot Creek (LHC) sample was collected as previously described (17, 18, 20). Briefly, cells from 80 °C sediment collected from the LHC4 hot spring in the Long Valley Caldera, California, USA, were separated by Nycodenz density centrifugation and stored either on ice (4 °C, in 10% v/v ethanol) or dry ice (80 °C, in 6% w/v betaine).

DNA extraction and sequencing

DNA from the TIV sample was extracted using the Power Max Soil DNA isolation kit (MoBio Labs) and purified with the Aurora system (Boreal Genomics).

Metagenome libraries from the sample were prepared from 1 ng purified DNA using the Nextera XT DNA library kit and sequenced with the Illumina HiSeq 2000 instrument at the Uppsala SNP&SEQ Technology Platform, generating 49 Gb of 2x250 bp paired-end reads, as previously described by Zaremba-Niedzwiedzka et al. (2017).

Cells from the LHC samples were sorted within 1 week of collection and subsequently lysed. The DNA was amplified using multiple displacement amplification (MDA) in a microfluidic device and individual cells were screened with PCR, as previously described by J. A. Dodsworth et al. (20). Subsequently, 1 µl of the original single-amplified genome (SAG) DNA of the Korarchaeon was used in another round of MDA to generate additional DNA of this SAG for library preparation (18). The MDA reaction was performed using the Qiagen REPLI-g Mini kit and purified with the QIAamp DNA Mini kit (Qiagen, Hilden, Germany). Next, a

157 sequencing library was generated from 1 ng of the resulting SAG DNA with the
158 NexteraXT library preparation kit (Illumina, San Diego, CA, USA) and the library was
159 then sequenced on an Illumina MiSeq instrument (2x300 bp), resulting in 4.9 Gbp of
160 raw data.
161 Additionally, DNA was extracted from LHC sediment using the FastDNA Spin Kit for
162 Soil (MP Biomedicals, Solon, Ohio, USA). Two metagenome libraries were generated
163 using the Nextera DNA Sample Preparation kit (Illumina): a MiSeq and a HiSeq
164 library which generated 7.3 and 10.8 Gbp, respectively.
165
166 **Assembly and binning**
167 Illumina adapters in the raw reads of each metagenome dataset were removed using
168 Scythe (https://github.com/vsbuffalo/scythe) with default parameters. Low quality
169 regions in the sequences were trimmed with Sickle (https://github.com/najoshi/sickle)
170 with the *-q 30* parameter to only keep reads with a PHRED score above 30.
171 Afterwards, reads were assembled using SPAdes (version 3.5.0;
172 https://github.com/ablab/spades) with the following settings *--meta -k 21, 33, 55*; and
173 IBDA-UA (version 1.1.1; (21)) with the following settings: *--maxk 124*. Assembled
174 contigs larger than 1 kbp were binned with the Emergent Self-Organizing Maps
175 approach (ESOM) (22) with the following parameters: *-min 5000 -max 10000 -t*
176 *robust*. korarchaeotal bins were visually inspected and manually selected using
177 Databionic ESOM Tools (23). Bins containing SSU genes displaying high similarity to
178 the *Ca.* K. cryptofilum SSU gene sequence were selected for further processing. A
179 description of the *Ca*. Korarchaeota archaeon LHC assembly can be found in J. H.
180 Saw et al. (18).
181
182 **Cleaning of new and reference Korarchaeota bins**
183 The five korarchaeotal bins (one from LHC and four from TIV) were screened for
184 non-korarchaeotal contigs using the R package multi-metagenome
185 (https://github.com/MadsAlbertsen/multi-metagenome) (24). Contigs with deviant
186 coverage, GC- or kmer content or with marker genes not phylogenetically clustered
187 to other korarchaeotal genes were manually inspected and removed if deemed not
188 likely to belong to the korarchaeotal clade.
189 Together with the five newly assembled Korarchaeota metagenome-assembled
190 genomes (MAGs), a total of 34 publicly available MAGs classified as Korarchaeota in

191 the NCBI (8), SILVA (25) and GTDB (3) databases were downloaded from NCBI in
192 October 2020. The completeness and contamination levels of all MAGs were
193 determined using CheckM v1.1.3 (26) and CheckM2 v1.0.1 (27). Additionally, MAGs
194 were screened using an in-house RP15 pipeline, as previously described (28).
195 Briefly, this pipeline was used to identify contigs that contain between 3 and 15
196 ribosomal proteins (RPs). The RPs found on these '*ribocontigs'* are then aligned to
197 the orthologous RPs from a set of reference taxa. Subsequently, the separate
198 alignments are trimmed, concatenated, and used to infer a phylogenetic tree using
199 FastTree 2 (*-wag model*) (29) and IQ-TREE 2.1.2 (*LG+C60+F+G4 model*, *-bb 1000 -*
200 *bcor 0.98*) (30).
201 Based on the position of ribocontigs in this tree, we observed that five of the publicly
202 available MAGs clustered outside of Korarchaeota and we removed them from
203 subsequent analyses.
204
205 **Phylogeny of 16S rRNA genes**
206 To construct a phylogenetic maximum likelihood (ML) tree of the TACK superphylum,
207 partial 16S rRNA gene sequences were extracted from Korarchaeota MAGs using
208 barrnap v0.9 (https://github.com/tseemann/barrnap) with the flags *--kingdom*
209 *arc/bac/euk*. Additionally, a total of 305 korarchaeotal sequences (Table S3) were
210 retrieved from the SILVA Ref NR database SSU r138.1 (25), together with 2367 near
211 full-length reference sequences (>1300 bp) representing the other taxa
212 encompassed in the TACK superphylum. Asgardarchaeaota 16S rRNA gene
213 sequences (>1300 bp; 305 sequences) were retrieved to be used as an outgroup.
214 Several Korarchaeota 16S rRNA gene sequences retrieved from the MAGs were
215 already present in the SILVA database and therefore removed from the dataset to
216 avoid duplicate sequence entries.
217 The dataset was aligned using MAFFT v7.310 (*--auto*) (31). Subsequently, the
218 alignment was processed using trimAl v1.4.rev22 to remove all columns with gaps in
219 more than 80% of the sequences (*-gt 0.2*) (32). The resulting alignment was then
220 used to construct a ML tree using IQ-TREE 2.1.2 (30), with settings: "*-m GTR+F -bb*
221 *1000 -altr 1000 --bcor 0.98*" (33).
222 The initial phylogenetic tree revealed that several publicly available Korarchaeota
223 MAGs contained contigs harboring non-korarchaeotal 16S rRNA gene copies. These
224 contigs were removed from the MAGs before downstream analyses. Additionally, two

225 Korarchaeota 16S rRNA gene sequences from the SILVA database were shown not
226 to cluster with this phylum. These non-Korarchaeota 16S rRNA gene sequences
227 were removed from the dataset after which the above steps were repeated.
228 If publicly available and documented, the sample origin (i.e., country and
229 environment) from which the MAGs and 16S rRNA gene sequences originate was
230 mapped onto the phylogenetic tree to visualize in which environments korarchaeotal
231 clades have been detected. All trees were visualized using iTOL v6 (34).
232
233 **Environmental distribution assessment**
234 To detect the presence of Korarchaeota in the environment, the 317 Korarchaeota
235 16S rRNA gene sequences were subjected to an integrated microbial NGS platform
236 (IMNGS) search using a similarity threshold of 90% and a minimum size of 200 bp
237 (35). From the results, sequences from datasets with less than 10.000 reads were
238 deleted. The remaining Sequence Read Archive (SRA) sequence hits returned by the
239 analysis were added to the aforementioned 16S rRNA gene alignment using the --
240 *add* command in MAFFT. This alignment was then used to construct a ML tree using
241 the GTR+F model and ultrafast bootstrap support values as described above. Only
242 SRA sequences clustering with Korarchaeota were retained from the IMNGS results.
243 The origin of these SRA datasets was retrieved from NCBI and used to create an
244 overview of the relative abundance of Korarchaeota in the environment (Table S4).
245
246 **16S rRNA gene primer coverage of Korarchaeota**
247 To analyze how well primers previously used to obtain Korarchaeota 16S rRNA gene
248 sequences target different Korarchaeota clades, amplification primer information
249 was, if documented, retrieved for the 294 PCR-amplicon-derived Korarchaeota 16S
250 rRNA gene sequences in SILVA (Table S3). These primer sequences were then
251 used in an *in silico* analysis which included all 317 Korarchaeota 16S rRNA gene
252 sequences, together with 12 recently designed universal primers to determine primer
253 coverage of different korarchaeotal lineages as described in G. Tahon et al. (7)
254 (Table S5).
255
256 **Selection of Korarchaeota MAGs for species tree inference**
257 To investigate the taxonomic diversity of Korarchaeota, average amino acid identity
258 (AAI) of verified Korarchaeota MAGs was calculated using EzAAI

259 (https://github.com/endixk/ezaai) (Table S6). Additionally, the 32 Korarchaeota MAGs
260 were used as input for a phylogenetic analysis pipeline using GToTree (36) with
261 settings "*-H Archaea -G 0.25",* together with a previous in-house selection of 119
262 archaeal proteomes spanning archaeal diversity with an emphasis on Asgard
263 archaeal representatives (Table S1). DPANN archaea were not included because
264 they are only remotely related to TACK archaea and tend to be long-branching taxa
265 in phylogenies, which can artefactually distort the inferred topology. The
266 concatenated alignment obtained by GToTree was subsequently used to construct a
267 ML phylogenetic tree with IQ-TREE v2.1.2 using 1000 ultrafast bootstraps and aLRT
268 under the LG+C60+F+G4 model. Using this tree, the 16S rRNA gene tree, CheckM
269 results, AAI values and the RP15 pipeline results, a subset of 18 MAGs - covering all
270 major korarchaeotal lineages (Table S2) - was selected for amalgamated likelihood
271 estimation (ALE) (37).
272
273 **Phylogenetic analyses of concatenated protein alignments for species tree**
274 **inference**
275 Two sets of phylogenetic markers were used to infer the species tree. The first one
276 (RP56) is based on a previously published dataset of 56 RPs (19). The second one
277 (NM54, for 'new markers') corresponds to 54 proteins extracted from a set of 200
278 markers previously identified as core archaeal proteins that can be used to robustly
279 infer the tree of archaea (38, 39).
280 For both RP56 and NM54 marker sets, we gathered orthologs from all korarchaeotal
281 predicted proteomes - predicted using prodigal (40) - as well as the other 119
282 archaeal proteomes. To retrieve orthologs from the set of selected taxa, we used
283 sequences from the previously published alignments (19, 38) as queries for BLASTp
284 against all proteomes. For each marker, the best BLAST hit from each proteome was
285 added to the dataset. For the first iteration, each dataset was aligned using MAFFT-
286 linsi (41) and ambiguously aligned positions were trimmed using BMGE 1.12 (*-m*
287 *BLOSUM30*) (42). ML individual protein phylogenies were inferred using IQ-TREE
288 v1.5.5 (43) under the LG+C20+F+G4 substitution model with 1000 ultrafast
289 bootstraps and were manually inspected.
290 Individual marker phylogenies were inspected for erroneous inclusion of paralogous
291 or contaminated sequences and to identify putative horizontal gene transfers. After

292 deleting dubious sequences, datasets were realigned and trimmed, and phylogenies
293 re-estimated as described above.
294 Trimmed alignments were then concatenated into two supermatrices (RP56 and
295 NM54 supermatrices). For each dataset, a phylogeny was inferred using IQ-TREE
296 under the LG+C60+F+G4 model. Statistical support for branches was calculated
297 using 1000 replicates for ultrafast bootstrap and aLRT analyses.
298 The RP56 and NM54 supermatrix alignments were used to create four more
299 alignments: an SR4-recoded alignment (44), an alignment where 50% of the fastest
300 evolving sites were removed, a χ^2-trimmed (28, 45) alignment where the 50% most
301 heterogeneous sites were removed in steps of 1% and a custom recoded alignment.
302 For the latter we calculated custom reduced amino acid schemes using the minmax
303 chisq method (https://www.mathstat.dal.ca/~tsusko/doc/minmax-chisq.pdf) (44) and
304 used these to recode sequence alignments.
305 For the 8 recoded or trimmed alignments, a phylogenetic ML tree was inferred using
306 1000 ultrafast bootstraps and aLRT under the GTR+G4+C60SR4 model (SR4-
307 recoded alignments) or LG+C60+F+G4 model (other alignments). The ten trees (i.e.,
308 two supermatrices, five alignments each) were subsequently used to infer 100 non-
309 parametric bootstrap trees using the posterior mean site frequency (PMSF) method
310 under the transfer bootstrap expectation (TBE) interpretation (*-b 100 --tbe*) (46).
311 Additionally, we ran four independent PhyloBayes-MPI v1.7a (47) Monte Carlo
312 Markov Chains (*-cat -gtr*) for each of the alignments. Results were checked to
313 determine whether convergence (maxdiff <0.3) and/or sufficient effective sample size
314 (effsize > 300) was reached. Subsequently, posterior predictive checks were
315 performed with PhyloBayes-MPI to determine whether the inferred phylogenetic
316 models adequately captured the across-taxa compositional heterogeneity and site-
317 specific pattern diversity present in the alignments.
318
319 **Homology detection and protein family phylogenetic analysis**
320 We used the annotation of arCOGs from EggNOG v5.0 (using eggnog-mapper
321 v2.0.5) (48) to assign all 290,917 proteins from 137 taxa (both MAGs and reference
322 genomes, Tables S1-2) into orthologous groups. All proteins without arCOG
323 annotation were subjected to *de novo* clustering using OrthoFinder v2.4.0 *(-os -M*
324 *msa*) (49), resulting in a total of 52,724 clusters (9,241 arCOG and 43,483 *de novo*).

For all protein clusters with at least four members (i.e. 9,108) multiple sequence alignments were computed using MAFFT v7.475 E-INS-I (41) and subsequently processed with trimAl v1.4.1 to filter gap-rich alignment columns (*-gt 0.01*) (32). Single-gene trees were inferred from these 9,108 alignments with IQ-TREE v2.0.3 with 1000 ultrafast bootstraps (*-bb 1000 -bnni -wbtl*) (30, 33). A model test was performed (*-m TESTNEW -mset LG -madd LG+C10, .., LG+C60*) for each tree inference (50). For clusters with only two or three members, only a single topology is possible for an unrooted tree and the unrooted gene trees were created for them.

Gene tree-aware ancestral reconstruction

ALE v0.4 (37) was applied to reconcile single gene trees with the species trees based on ribosomal markers (RP56) and the new marker set (NM54). Firstly, conditional clade probabilities were computed from the bootstrap samples (*ALEobserve*) and 100 reconciliations with the species tree were sampled (*ALEml_undated*). We thereby adjusted the reconstructed genome copy number with the extinction probability per cluster within ALE (see github.com/maxemil/ALE/commit/136b78e).

We applied a threshold of 0.3 to the raw reconciliation frequencies that are the output of ALE, meaning that events (loss, transfer, origination or duplication) and presence/absence (copies) were counted as such if they had a frequency of at least 0.3 (51). Singleton clusters were counted as originations at the corresponding species node.

Annotation of carbohydrate-active enzymes and transporters

Presence and distribution of glycosyl hydrolases, glycosyl transferases, polysaccharide lyases, carboxyl esterases, carbohydrate-binding modules and auxiliary activities were identified through the dbCAN2 meta server (52), using the CAZy database (July 2021). Carbohydrate-active enzymes predicted with at least two of the integrated tools were used. Transporters were identified using TransAAP (July 2021) (53).

Korarchaeota optimal growth temperature

The optimal growth temperature of Korarchaeota MAGs was predicted using OGT_prediction (54).

359 Inference of growth temperature of the last Korarchaeota common ancestor was
360 performed using the alignment and ML tree of the nucleoside diphosphate kinase
361 sequences from the MAGs, obtained during the protein family phylogenetic analysis.
362 Ancestral sequences and their unfolding midpoint temperatures (T_m) were predicted
363 using FireProt^{ASR} (55) and ProTstab2 (56), respectively.
364

365 **Reverse gyrase phylogeny**

366 To create a phylogenetic tree of reverse gyrase, all corresponding protein sequences
367 were retrieved from UniProt Release 2023_02 (57). Reverse gyrase sequences from
368 the 137 MAGs, if not yet present in UniProt, were added to this dataset. Sequences
369 were aligned with MAFFT v7.310 (*--auto*) after which the alignment was trimmed with
370 trimAl v1.4.rev22 to remove all columns with gaps in more than 80% of the
371 sequences (*-gt 0.2*). The ML phylogenetic tree was then generated using IQ-TREE
372 v2.0.3 (*-bb 1000 -alrt 1000*) with the model LG+C60+F+R8, as chosen by the
373 Bayesian information criterion (BIC) using ModelFinder Plus *(-m MFP -mset JTT,*
374 *JTT+C10..C60, WAG, WAG+C10..C60, LG, LG+C10..C60, Q.pfam,*
375 *Q.pfam+C10…+C60 -mrate G4, G6, G8, R4, R6, R8 -mfreq F*) (50).
376

377 **Relative evolutionary divergence**

378 To determine a robust phylogenomic placement of a potential novel family-level
379 lineage placed at the base of Korarchaeota by F. Vulcano et al. (58), seven MAGs
380 from lineages Kg_2, Kg_3 and Kg_5, with a completeness estimation above 75%
381 were added to our dataset containing the original 137 MAGs, and MAGs classified as
382 *Ca*. Njordarchaeales. The single MAG from lineage Kg_4 could not be included due
383 to its low completeness. The expanded dataset was subjected to the same analyses
384 as before to create a NM54 supermatrix alignment. This alignment was χ^2-trimmed
385 as mentioned before, after which it was used to create a non-parametric bootstrap
386 tree using the PMSF method under the TBE interpretation. The resulting ML tree was
387 used as input for a phylogenomic analysis using the Castor package (59). Relative
388 evolutionary divergence values were calculated for all nodes of the tree to determine
389 whether this new lineage was part of the phylum Korarchaeota.
390

391 **Code availability**

392 All scripts needed to reproduce the protein clustering and running ALE are available
393 at https://github.com/maxemil/korarchaeota-scripts and
394 https://github.com/maxemil/ALE-pipeline.
395
396 **Data availability**
397 Sequencing raw reads and draft genome assemblies for the five new Korarchaeota
398 MAGs have been deposited in the NCBI GenBank database under the accession
399 numbers PRJNA983117 (TIV) and PRJNA983228 (LHC), respectively.
400

401 ## Results and Discussion

402

403 **Contamination and misclassification among Korarchaeota**

404 A first round of quality control using all contigs that contained (part of) the ribosomal
405 protein (RP) operon, contigs that harbor partial 16S rRNA genes, and 76 single-copy
406 archaeal marker proteins (see methods for detail) revealed that several of the
407 publicly available metagenome-assembled genomes (MAGs) assigned as
408 Korarchaeota were prone to contamination (Table S2). One MAG contained a 5 kbps
409 contig harboring a bacterial 16S rRNA gene. Another MAG contained a contig
410 harboring bacterial RPs. Finally, one MAG turned out to be a hybrid of bacterial and
411 fungal contigs. In addition, some MAGs did not truly belong to Korarchaeota. Among
412 these, three MAGs were found to represent *Ca*. Njordarchaeales, a recently
413 described clade belonging to the Asgard archaea (39, 60), rather than Korarchaeota.
414 A third MAG (DRAE01.1) was found to represent a yet undescribed lineage in TACK.
415 Although sometimes classified as a lineage at the base of Korarchaeota, this
416 placement was not supported and resulted from inferior phylogenomic analyses (3,
417 58). These findings underline that automated results generated by commonly used
418 binning software, quality control tools and phylogenomic placement pipelines such as
419 CheckM (26) and GTDB-tk (61) cannot be blindly trusted, because these tools do not
420 perform phylogenetic validation of marker genes. For example, hybrid MAGs may
421 have high completeness and low contamination values and even have robust
422 phylogenetic placement. Therefore, thorough quality control and manual curation of
423 MAGs is needed prior to their public deposit and classification. Phylogenetic
424 analyses of the MAGs verified to be Korarchaeota, based on the 16S rRNA gene,
425 contigs encoding at least 3 ribosomal proteins encoded on of a conserved ribosomal
426 protein operon, and 76 single-copy archaeal marker proteins all revealed that
427 Korarchaeota formed a monophyletic group in Archaea, however, with an unstable
428 phylogenetic position and internal topology (Figs. S1-3). This urged us to perform in-
429 depth phylogenomic analyses to resolve the phylogenetic position of Korarchaeota
430 and its internal topology.

431

432 **Korarchaeota are placed at the base of the 'ATACK superphylum'**

433 The availability of 32 *bona fide* Korarchaeota MAGs in October 2020 allowed us to
434 robustly place the phylum in the archaeal tree of life. To this end, we subsampled the

435 dataset to reduce genomic redundancy of highly similar taxa in order to reduce
436 computational burden. More Korarchaeota MAGs became publicly available after our
437 initial selection (8, 58, 62). The vast majority of these comprised taxa already
438 covered by our dataset. Some MAGs, however, represented a novel order at the
439 base of Korarchaeota (58), as supported by estimation of relative evolutionary
440 divergence values (Fig. S4). These MAGs became public when our project was in its
441 final stage and could thus not be included in the analyses anymore. Further analysis
442 of these MAGs indicated they did not change our findings.
443 Based on our initial phylogenomic tree including all MAGs (Fig. S3) and average
444 amino acid identity (AAI) (Table S6) (63), we observed that the 32 MAGs could be
445 grouped into 20 species, eight genera, four families and one order (Table S2, Fig.
446 S3). The five MAGs obtained from in-house sequencing efforts all represented
447 different species – four of which were novel – belonging to different genera. A total of
448 18 MAGs from all genera were then selected (with completeness >65%) for
449 downstream analyses. For each taxon, we here propose *Candidatus* names that will
450 be used throughout the text. Details on the taxonomic descriptions are given in the
451 Supplementary Note. All names and the relevant nomenclatural types have been
452 proposed under the SeqCode (64).
453 A first set of phylogenomic trees was inferred using 56 concatenated conserved
454 ribosomal proteins (RP56) (39). In addition to the original concatenation, we
455 performed four treatments aiming to reduce potential phylogenetic reconstruction
456 artefacts yielded by saturation and compositionally biased sites (39, 65). For this, we
457 first recoded our alignments into four character-states (SR4-recoding or custom-
458 recoding) (44). We also removed the 50% fastest-evolving sites. Finally, we used a
459 χ^2 trimmer to remove the 50% most compositionally heterogeneous sites (28). The
460 initial alignment and its four derivations (i.e., SR4-recoded, custom recoded, 50%
461 fastest evolving sites removed, and χ^2-trimmed) were then used to infer maximum
462 likelihood (ML) trees using the posterior mean site frequency (PMSF) method (Figs.
463 S5-9), and for phylogenetic reconstruction using PhyloBayes-MPI (Figs. S10-14).
464 Phylogenies based on all variations of the supermatrix, whether reconstructed in ML
465 or Bayesian framework, showed Korarchaeota branching at the base of the TACK
466 superphylum, although statistical support ranged from weak to high (80-99%
467 bootstrap support; 0.69-1 posterior probability). The only exception was the Bayesian
468 inference based on the RP56-SR4-recoded dataset, which placed Korarchaeota

469 nested within TACK. Regarding the internal topology, we consistently recovered two
470 distinct groups of MAGs originating from hot springs, representing the genera *Ca*.
471 Korarchaeum and *Ca*. Methanodesulfokora, respectively. However, these were never
472 found to be monophyletic, since the hydrothermal vent MAG representing the genus
473 *Ca*. Fumariihydrokora always placed as sister-lineage to *Ca*. Korarchaeum. The rest
474 of relationships between deep branches within the Korarchaeota varied depending
475 on the analysis. In particular, we observed various placements of the hydrothermal
476 vent genera *Ca*. Calidiprofundikora. and *Ca*. Sedimenticalidikora which are
477 characterized by high genomic GC content (Fig. 2). This is in part reflected by the
478 fact that, out of the five RP56 alignments, only the χ^2-trimmed one was appropriately
479 modelled by the sophisticated CAT-GTR model of evolution, based on bayesian
480 posterior prediction tests (Table S7). And even in this case, the monophyly between
481 the two clades from hot spring environments was poorly supported (Figs. S8 and
482 S13).
483
484 As previously shown, RPs can yield phylogenomic reconstruction artefacts
485 depending on the lineage considered (39). Despite them being highly conserved
486 across the different domains of life, they are short in length (66) resulting in short
487 concatenated alignments (i.e. 3575-7158 alignment positions in our case) and thus
488 limited phylogenetic signal (Table S7). Furthermore, RPs are all part of a unique
489 multisubunit complex, meaning they have undergone strong co-evolution. As a result,
490 the assumptions of most evolutionary models of independence of evolution at sites
491 are violated. Additionally, this implies that any departure from an average amino-acid
492 composition will be propagated across subunits, and the noise/signal ratio amplified
493 during phylogenetic reconstruction. This is particularly relevant since the amino-acid
494 composition of RPs is strongly influenced by adaptation to high temperature
495 environments, even more than the rest of the proteome (39). All Korarchaeota
496 genomes were obtained from thermal environments, and using genomic information
497 we estimated phylum members to have an optimal growth temperature of 70-85 °C
498 (Table S2). One recently published Korarchaeota MAG was suggested to be
499 mesophilic (62). However, the authors did not determine any optimal growth
500 parameters of the organism. Here, we determined the optimal growth temperature of
501 this MAG to be 79.3 °C, a value similar that of all other known Korarchaeota.
502 Therefore, it can be assumed that this organism is a thermophile and not a

503 mesophile. As a result of this adaptation to high temperature, the amino acid
504 composition of their RPs has evolved strong thermostability-related signatures, such
505 as a high ratio of charged versus polar amino acids, which can yield phylogenomic
506 artefacts (39). An illustrative example of how problematic this can be is the recent
507 case of the so-called *Ca*. Panguiarchaeales (67). Although described as a novel
508 korarchaeotal lineage, phylogenomic and AAI analyses (Fig. S15, Table S8) show
509 that all genomes described under this lineage name correspond to a low diverse
510 taxon that instead belongs to the *Ca.* Njordarchaeales, a hyperthermophilic Asgard
511 archaeal order (60). A similar problem arises in the work of J. Pan et al. (62), where
512 several *Ca*. Njordarchaeales MAGs were considered as Korarchaeota (i.e.,
513 3300022188_28 and GCA_011042815.1). *Ca*. Njordarchaeales have been shown to
514 artefactually branch outside Asgard archaea as a sister lineage to Korarchaeota
515 when inferring phylogenies based on RPs (39). The above examples show the
516 importance of constructing robust phylogenomic trees including a broad diversity of
517 reference data. Failure to do so (e.g., the work of Y. Liu and M. Li (68)) ultimately
518 leads to incorrect results and conclusions.
519 To overcome limitations and inconsistencies specific to RP datasets, a second
520 supermatrix alignment was constructed based on 54 core archaeal proteins that can
521 be robustly used to infer the tree of Archaea (39). This new marker (NM54)
522 supermatrix alignment and its four derivations were subjected to the same analyses
523 as the RP56 alignments. Contrary to the RP-based phylogenies, all resulting NM54-
524 based trees showed Korarchaeota not at the base of or within TACK, but instead at
525 the base of a large superphylum composed of 'TAC' (i.e., TACK clade without
526 Korarchaeota) and Asgard archaea, with high support (Figs. S16-24); altogether,
527 they compose the informally named ATACK superphylum. Given this revisited
528 position of Korarchaeota in the tree of Archaea using robust and sophisticated
529 phylogenomic analyses, we show Korarchaeota to be a phylum-level lineage and not
530 a class (i.e., *Ca*. Korarchaeia) of the Thermoproteota, as suggested by D. H. Parks et
531 al. (3). Additionally, the internal Korarchaeota topology of the NM54 supermatrix
532 alignments always showed a clear highly supported grouping of MAGs from hot
533 springs, with the hydrothermal vent lineage *Ca*. Fumariihydrokora interrupting their
534 monophyly. The χ^2-trimmed alignment was again the only one appropriately modelled
535 based on posterior prediction tests (Table 1). Contrary to its RP56 counterpart, the
536 consensus tree of the four PhyloBayes-MPI Monte Carlo Markov Chains now showed

537 maximum support for branches within Korarchaeota (Fig. 1). Additionally, the
538 PhyloBayes sampling statistics showed sufficient effective sample size, and posterior
539 prediction tests indicated that the model adequately captured the site-specific pattern
540 diversity present in the NM54-χ^2-trimmed alignment (Table 1). Therefore, we opted to
541 use this alignment and the resulting phylogeny as species tree for the ancestral
542 reconstruction analyses (below). We also performed phylogenomic analyses of our
543 dataset (of the untreated and four treated alignments) expanded with *Ca*.
544 Njordarchaeales (including the so-called *Ca*. Panguiarchaeales) MAGs. In the NM54
545 phylogenomic trees, contrary to the RP56 marker set, these lineages are not
546 misplaced outside of the Asgard archaea and next to Korarchaeota but remain in
547 place within the Asgard archaea (Fig. S25), thus showing that the NM54 marker set
548 is better suited to resolve deep evolutionary relationships compared to marker sets
549 based on ribosomal proteins. This highlights the importance of performing such
550 sophisticated phylogenomic analyses prior to ALE. The different yet more robust and
551 accurate placement of Korarchaeota in our study, compared to F. Vulcano et al. (58),
552 J. Pan et al. (62) and Y. Liu and M. Li (68) undoubtedly impacts downstream
553 analyses. For example, the previously estimated first divergence time of
554 Korarchaeota (62) – when placed at the base of TACK – will greatly differ. With
555 Korarchaeota placed at the base of ATACK, their estimated first divergence time
556 would be in the Archean, long before the Great Oxidation Event. Therefore,
557 Korarchaeota evolution was very unlikely driven by the availability of oxygen.
558
559 **Korarchaeota have a limited and environmentally confined distribution**
560 Interestingly, all *bona fide* Korarchaeota MAGs originated from either hot springs or
561 hydrothermal vents (Table S2), as was the case for all korarchaeotal 16S rRNA gene
562 sequences from SILVA (Table S3). Because of the limited number of MAGs and 16S
563 rRNA gene sequences, we also explored the phylogenetic diversity of the phylum
564 using ~280000 16S rRNA gene amplicon datasets included in the IMNGS platform
565 (35). Only 161 datasets were found to contain Korarchaeota sequences (Table S4),
566 as verified with a phylogenetic analysis (Fig. S2). Of these, only 5 and 24 had a
567 Korarchaeota relative abundance of more than 1% and 0.1%, respectively. All but 4
568 datasets originated from known hydrothermal environments (Fig. S26). Together,
569 these results suggest that Korarchaeota exhibit a high level of endemism and low
570 abundance, even in their niche.

571 To investigate potential explanations for the restricted distribution and limited
572 abundance of Korarchaeota, metadata was gathered for all MAGs, 16S rRNA gene
573 sequences and IMNGS datasets. The majority of amplicon-based 16S rRNA gene
574 sequences were obtained using Korarchaeota-specific and universal primers that
575 were designed at least one decade ago (Table S5). An *in silico* primer coverage
576 analysis showed that these primers almost exclusively target *Ca.* Korarchaeaceae
577 with no or relatively few mismatches (Table S5). At the time of designing, only 16S
578 rRNA gene sequences from these clades were available, which explains why these
579 primers fail to target more distant Korarchaeota lineages. Various more recently
580 developed universal primers, however, target the Archaea much better (Table S5)
581 (7). Accordingly, the IMNGS datasets containing the highest Korarchaeota relative
582 abundances were generated using these universal primers. Altogether, this suggests
583 that the limited availability of Korarchaeota data is the result not only of their true low
584 abundance and restricted distribution, but also of limited sequencing depth and the
585 use of inaccurate primers.
586 All phylogenomic trees (Figs. S3-25) showed the presence of two distinct, yet low
587 phylogenetically diverse clusters of MAGs originating from hot springs, representing
588 the genera *Ca*. Korarchaeum and *Ca*. Methanodesulfokora. Each hot spring cluster
589 was composed of a clade of multiple closely related MAGs from USA hot springs,
590 with a non-USA (i.e., Russia and New Zealand, respectively) MAG at their base.
591 Hydrothermal vent MAGs, on the other hand, form six distinct genera, their species
592 originating either from the Taketomi Island shallow submarine hydrothermal field,
593 Guaymas Basin or the Mid-Cayman Rise.
594 We wanted to check if the environmental and geographical split observed on
595 Korarchaeota MAG phylogenies could be expanded to the larger amount of 16S
596 rRNA gene sequence data available. The Korarchaeota topology of the 16S rRNA
597 gene trees (Figs. S1-2) revealed that, compared to hot spring lineages, only few
598 sequences grouped with the hydrothermal vent clades – especially the basal ones –
599 which could be explained by primer bias and undersampling of these latter
600 environments (see above). Overall, the two 16S rRNA gene trees were congruent
601 with the phylogenomic ones. However, the relationships between hot spring and
602 hydrothermal vent sequences were less well resolved. The hot spring lineages of *Ca.*
603 Methanodesulfokora and *Ca.* Korarchaeum were comprised of closely related
604 subclusters originating from Icelandic, Russian, USA and New-Zealand hot springs,

605 and Malaysian, New Zealand, Japanese, Russian, Chinese, Bulgarian and USA
606 locations, respectively (Table S3). Although similar in temperature and pH, these
607 locations have vastly different chemistries. Therefore, despite their lower
608 phylogenetic diversity, hot spring Korarchaeota could still be genetically distinct. To
609 date, however, no MAGs are available from most of these sites. Interestingly, *Ca.*
610 Methanodesulfokora contained one single hydrothermal vent sequence (accession
611 number AF411237.1), whereas *Ca*. Fumariihydrokora, close relative of *Ca*.
612 Korarchaeum, contained a single hot spring sequence (accession number
613 DQ300332.1). These sequences originated respectively from a coastal shallow water
614 hydrothermal vent and a geothermal well less than 100m from the beach, thus
615 possibly representing environments with mixed characteristics. Nevertheless,
616 biogeographic clustering suggests the existence of geographic variants.
617 All analyses support a Korarchaeota niche-invasion from marine to terrestrial
618 geothermal habitats and a high degree of endemism. How this transition exactly took
619 place and how genetically similar Korarchaeota from distant locations are remains
620 unclear. Genetically similar korarchaeotal hot spring populations in the USA have
621 been suggested to follow a spatial distribution resulting from a historical connection
622 or an invasion-type event (69). Although these hypotheses could explain similar
623 phylotypes in locations that are geographically close to each other, they fail to explain
624 the clustering of phylotypes from distant locations. Instead, Korarchaeota probably
625 adapted to specific environmental conditions of their respective habitats, and spread
626 globally. For example, many hot springs where Korarchaeota have been detected
627 share multiple environmental conditions (69, 70) that may have influenced the
628 evolutionary origin of the two major hot spring phylotypes.
629

630 **The last Korarchaeota common ancestor: a versatile anaerobic autotrophic**
631 **thermophile**

632 To explore the genetic basis of the transitions between hydrothermal vents and hot
633 springs or *vice versa*, and more generally how ancestral gene content evolved along
634 the Korarchaeota species tree, we reconciled the histories of gene families and
635 species tree using ALE (37). First, all proteins from the 137 MAGs were clustered
636 into gene families, after which single-gene trees were inferred for all gene families.
637 Gene family histories were then inferred by reconciling gene trees with the species
638 tree obtained from the NM54-χ^2-trimmed alignment (Fig. 1) to infer four types of

639 ancestral events (i.e., gene transfer, loss, duplication, and origination) as well as their
640 resulting gene copy numbers on all ancestral nodes of the species tree (Fig. 2, Table
641 S9). First, we screened the inferred gene content of the last Korarchaeota common
642 ancestor (LKorCA).
643 The ALE analysis suggested that LKorCA produced biomass autotrophically by
644 assimilating inorganic carbon via the Wood-Ljungdahl pathway, thus turning CO_2 into
645 acetyl-CoA (Fig. 3). Apart from the carbon monoxide dehydrogenase/acetyl-CoA
646 synthase, LKorCA also encoded the genetic potential of pyruvate:ferredoxin
647 oxidoreductase (*porABDG*). This key enzyme in anaerobic metabolism indicates that
648 LKorCA could further reduce and carboxylate acetyl-CoA to pyruvate and *vice versa*
649 (71).
650 LKorCA also contained a single *rbcL* copy, which encodes the archaeal type III
651 ribulose-1, 5-bisphosphate carboxylase-oxygenase (RuBisCO) of the Calvin-Benson-
652 Bassham (CBB) cycle. However, a complete CBB pathway was not present, which is
653 common in archaeal genome sequences (72). Instead, the presence of thymidine
654 phosphorylase (*deoA*) and eukaryotic translation initiation factor 2B (*e2b2*) hint that
655 *rbcL* is involved in the utilization of ribonucleotides and ribonucleosides via the
656 pentose-bisphosphate pathway (Fig. 3) (72, 73).
657 Inference of 3-hexulose-6-phosphate synthase (*hps*) and 6-phospho-3-
658 hexuloisomerase (*phi*) in LKorCA suggests LKorCA could potentially rely on the
659 reverse ribulose monophosphate pathway for biosynthesizing ribulose 5-phosphate
660 in absence of the pentose phosphate pathway (74).
661 The inferred genetic potential of LKorCA also showed evidence for assimilatory
662 sulfate reduction from sulfate to sulfite, a trait often found in conjunction with the
663 Wood-Ljungdahl pathway (Fig. 3). LKorCA likely dwelled in deep-sea hydrothermal
664 ecosystems. In these anoxic sediments niches carbon and sulfur cycling are
665 intimately linked, and the diffusion of sulfate from seawater makes sulfate reduction a
666 highly favorable process (75). Therefore, LKorCA might have played an important
667 role in sulfur cycling. For nitrogen intake, LKorCA relied on a glutamine synthetase
668 for the conversion of ammonium to glutamine. Phosphate, saccharides, metallic
669 cations and peptides could be imported using ABC-type transporters. The presence
670 of a complete branched-chain amino acid transport system (*liv* complex) suggests
671 LKorCA could take up and utilize protein-degradation products from its environment.
672 LKorCA also contained genes for the degradation of various amino acids, including

673 alanine, serine and aspartate. The presence of genes encoding proteins for β-
674 oxidation indicate at the genetic capacity of LKorCA to degrade fatty acids (Fig. 3).
675 LKorCA also encoded several cytoplasmic and membrane-bound complexes which
676 are relevant for energy conservation. Presence of the respiratory complex NADH-
677 ubiquinone oxidoreductase (*nuo*) would allow reoxidizing reduced ferrodoxin, which
678 can in turn be used as electron donor, as supported by the absence of *nuoEFG*, the
679 subunits required for the binding and oxidation of NADH (76). The Mvh/Hdr complex
680 potentially coupled the endergonic reduction of ferredoxin by H_2 to the exergonic
681 reduction of a disulfide bond (77). Presence of a FAD-containing hydrogenase (*glcD*)
682 could also indicate the use of HdrD/GlcD for heterodisulfide reduction (76). Hya and
683 Mvh [NiFe]-type hydrogenases act in mediating the oxidation of H_2, while an archaeal
684 ATPase (Ntp) ensures energy conservation (Fig. 3). In addition, LKorCA showed
685 genetic potential of HypA-F which are involved in the maturation of hydrogenases
686 (78).
687 Although a complete *de novo* pyrimidine synthesis pathway was inferred in LKorCA,
688 it lacked the purine counterpart. Therefore, LKorCA may have relied on transporting
689 nucleosides across the membrane using the ABC uptake system encoded by
690 *nupABC* and *bmpA* (Fig. 3).
691 Reverse gyrase, an enzyme which induces positive supercoiling in closed circular
692 DNA in an ATP-dependent manner (79) and unique to (hyper)thermophilic
693 organisms, was inferred as absent in LKorCA. However, all internal Korarchaeota
694 ancestors and MAGs have a reverse gyrase. Phylogenetic analysis revealed that
695 Korarchaeota reverse gyrase sequences are not monophyletic (Fig. S27).
696 Sequences from *Ca*. Caldabyssikora were distinctly placed from those of the other
697 Korarchaeota. Since Korarchaeota reverse gyrase sequences are not monophyletic,
698 we cannot exclude that LKorCA already had it based on our inferences and that it
699 might have been replaced in the last common ancestor of *Ca*. Caldabyssikora.
700 Although potential absence of reverse gyrase could indicate that LKorCA was not a
701 hyperthermophile, this is very unlikely. Firstly, reverse gyrase is not necessarily a
702 prerequisite for life at high temperatures, because the growth rate of a reverse
703 gyrase knockout of *Pyrococcus furiosus* was only impacted above 90 °C (80).
704 Therefore, the enzyme might not have been essential even if LKorCA had a
705 thermophilic lifestyle. Secondly, the last common ancestors of each of the major
706 archaeal clades were inferred to be thermophiles (81). Thirdly, nucleoside

707 diphosphate kinase is a well conserved enzyme of which the stability is directly
708 related to the host organism's environmental temperature (82). The unfolding
709 midpoint temperature of the reconstructed korarchaeotal ancestral nucleoside
710 diphosphate kinase sequence (T_m = 78 °C) was more stable than that of the
711 hyperthermophilic archaeon *Archaeoglobus fulgidus* DSM4304 (T_m = 62 °C) and thus
712 indicated LKorCA was a thermophile.
713 LKorCA contains several genes encoding antioxidant enzymes (e.g., superoxide
714 reductase, rubrerythrin). Previously, J. Pan et al. (62) suggested these genes to be
715 absent from the *Ca*. Caldabyssikora (i.e., Kor 7) and *Ca*. Thermotainarokora (i.e., Kor
716 8). However, our extended Korarchaeota dataset including five newly constructed in-
717 house MAGs, together with robust sophisticated phylogenomic analyses indicate that
718 this is not the case.
719
720 **A first environmental transition is coupled to the switch to a likely**
721 **heterotrophic lifestyle and higher virus predation**
722 The NM54-χ^2-trimmed alignment species tree indicated two environmental
723 transitions, both appearing within the *Ca*. Korarchaeaceae family (Fig. 1). The first
724 transition – from hydrothermal vents to hot springs – appeared at the branch before
725 the diversification of the hot spring genera *Ca*. Korarchaeum, *Ca*.
726 Methanodesulfokora and the hydrothermal vent genus Ca. Fumariihydrokora ('last
727 *Ca*. Korarchaeum, Methanodesulfokora and Fumariihydrokora common ancestor' -
728 LKMFCA). Upon this transition, the genetic repertoire of Korarchaeota was
729 significantly altered (Fig. 4). Contrary to LKorCA, LKMFCA was not an autotroph. The
730 Wood-Ljungdahl pathway, apart from *mch*, was lost in the last Korarchaeaceae
731 ancestor and replaced by the glycolysis pathway. The latter, however, lacks the
732 genes required for the reversible transformation between glucose and glucose-6-
733 phosphate. LKMFCA's lifestyle shifted towards that of fermentation, as could be
734 observed by the gain of the *acs* gene allowing the reversible conversion of acetyl-
735 CoA into acetate, and formate acetyltransferase.
736 Likewise, a shift took place in the sulfur metabolism, from assimilatory to dissimilatory
737 (DsrAB) reduction of sulfur compounds (Fig. 4). This complex, together with the Mvh
738 hydrogenase, indicates that hydrogen may have been used as an electron donor for
739 the reduction of sulfite (83). The nitrogen metabolism repertoire was significantly
740 expanded, including the glutamate dehydrogenase and glutamate synthase allowing

741 the conversion of ammonia and glutamine, respectively, to glutamate. Additionally,
742 presence of nitrate reductase suggests the ability to reduce nitrate to nitrite. A
743 formate dehydrogenase allowed the conversion of formate to carbon dioxide.
744 Compared to LKorCA, the genetic repertoire of LKMFCA also enlarged with various
745 glycosyltransferases, mainly of families GT2 and GT4 which are commonly involved
746 in processes such as N-glycosylation, cell wall biogenesis, cellulose, and sucrose
747 biosynthesis (84). Where the electron transport chain of LKorCA consisted of only
748 Complex I, that of LKMFCA was expanded with Complex II, the succinate
749 dehydrogenase. The environmental switch also coincided with major changes in the
750 amino acid metabolism of LKMFCA. Many of the biosynthesis pathways gained
751 genes (e.g., LeuA-D), whereas genes required for degradation were lost. Pathways
752 related to the metabolism of cofactors and vitamins were also subject to gene gain
753 and now included a complete NAD^+ biosynthesis pathway and greatly expanded
754 thiamine metabolism. Furthermore, LKMFCA gained multiple transporter systems
755 including those for arsenite efflux and transport of metal ions, amino acids, cations,
756 tungstate and nitrate/bicarbonate. As a result, LKMFCA had an increased potential to
757 scavenge its environment for nutrients and to efflux toxins out of its cell to survive
758 (85). Interestingly, at the time of environmental transition, LKMFCA gained many
759 genes participating in resistance to viruses, particularly those of type I and III *cas*,
760 and CARF systems (86, 87). Given that both environments are hyperthermophilic, it
761 is unclear exactly why these genes became enriched in hot spring Korarchaeota.
762 One potential explanation could be a different virome, although further research is
763 needed to answer this question. Several of the inferred type I and III genes (e.g.
764 Cmr4g7, Cmr1g7) are known to be associated with hyperthermophiles, indicating
765 they may be essential for managing viral infections in the hot spring habitats where
766 many archaeal viruses reside (88, 89). Others, such as the BREX system, might
767 allow phage adsorption while blocking its DNA replication (90). As such, the phage
768 could be used as a source of nucleotides in support of growth of LKMFCA (Fig. 4).
769

770 **A reverse transition towards a scavenging or symbiont lifestyle**

771 A second environmental transition, this time from hot spring to hydrothermal vent
772 adaptation, happened before the diversification of *Ca*. Korarchaeum and Ca.
773 Fumariihydrokora (LKFCA). While the central carbon metabolism of LKFCA did not
774 change compared to LKMFCA, many changes, particularly gene losses,

775 characterized this second transition (Fig. 5). LKFCA lost the ability to perform
776 dissimilatory reduction of sulfur compounds and to reversibly convert acetyl-CoA into
777 acetate. Likewise, the inferred genome of LKFCA is devoid of thymidine
778 phosphorylase (*deoA*), the eukaryotic translation initiation factor 2B (*e2b2*) and the
779 archaeal type III RuBisCO (*rbcL*) and thus lost the potential to use ribonucleotides
780 and ribonucleosides via the pentose-bisphosphate pathway. The thiamine
781 metabolism lost various key genes and was likely no longer functional. Given the
782 absence of the *acs* and *mobAB* genes, the interconversion of acetate and acetyl-
783 CoA, and synthesis of enzymes belonging to the DMSO reductase family were not
784 possible anymore, respectively. Additionally, the repertoire of LKFCA no longer
785 contained a tungstate ABC transporter. Despite many losses, LKFCA also gained
786 various novel genes and thus pathways. Like in the transition to LKMFCA, many
787 genes involved in amino acid biosynthesis were gained. Pathways for the
788 biosynthesis of valine, leucine, isoleucine, lysine and phenylalanine expanded as
789 well. A newly obtained spermidine/putrescine ABC transporter allowed regulating
790 these polyamines which play vital roles in ion homeostasis and cell proliferation.
791 Taken together, the second transition indicated that LKFCA underwent even more
792 metabolic changes towards a simplified genomic machinery. Contrary to the once
793 rather metabolically complex LKorCA, LKFCA transitioned to a simple anaerobic
794 heterotrophic lifestyle of peptides and amino acids (6). Whether LKFCA is just a very
795 efficient scavenger or a symbiont, or perhaps both, awaits to be proven with the
796 cultivation of Korarchaeum strains.
797
798 **The last Methanodesulfokora common ancestor: a virus-combatting**
799 **methanogen**
800 The last *Ca*. Methanodesulfokora common ancestor (LMCA) followed a different
801 evolutionary trajectory. Where LKFCA lost the *rbcL*, *e2be*, *deoA* and *dsrAB* genes,
802 and several key genes of the thiamine metabolism, LMCA kept this genomic
803 potential. LMCA also had an expanded amino acid metabolism, yet less extensive
804 than LKFCA. Several more type I and type III *cas* system genes participating in
805 resistance to viruses were acquired, indicating that *Ca*. Methanodesulfokora spp.
806 may need to manage additional viral threats in the hot springs where they reside (Fig.
807 6). Interestingly, we inferred in LMCA a whole suite of genes encoding methanol
808 methyltransferase (MtaABC) and methyl-coenzyme M reductase (McrABG) for

809 methanol reduction and methane production. Although *dsr* and *mcr* genes could hint
810 at a sulfite-dependent, anaerobic oxidation of methane to methanol (83), recent
811 cultivation of a strain of a *Ca*. Methanodesulfokora washburnensis has validated its
812 methylotrophic methanogenic potential (Personal communication with Roland
813 Hatzenpichler). Using genomic data, only *Ca*. M. washburnensis MAGs – all obtained
814 from USA hot springs – appear to have the Dsr, Mcr and Mta complexes. Whether
815 they are truly absent in the other *Ca*. Methanodesulfokora species, currently
816 comprises a single MAG obtained from a Russian hot spring, needs to be further
817 investigated by recovery of more complete MAGs and cultivated strains.
818
819 **Conclusion**
820 Despite rapid advancements in metagenomic sequencing analysis which particularly
821 expanded our view of the archaeal branch of the Tree of Life, our understanding of
822 Korarchaeota genomics is limited. Here, the addition of five new MAGs improved the
823 phylogenomic confidence, placing the phylum at the base of TACK and Asgard,
824 creating an ATACK superphylum, and showed a clear korarchaeotal endemism and
825 two habitat transition between marine and terrestrial geothermal habitats. The
826 inclusion of MAGs from all currently known korarchaeotal genera in gene tree-aware
827 ancestral genome reconstructions revealed that these environmental transitions were
828 subject to significant metabolic changes, thereby completely reshaping their genetic
829 composition. Our work provides new insights into the distribution, evolution and
830 physiology of the Korarchaeota and guides future cultivation-dependent and -
831 independent studies to further unravel its enigmatic biology and lifestyle.
832

Acknowledgements

The authors would like to express their appreciation to Nina Dombrowski and Kassiani Panagiotou for their support. We thank Peter Vandamme for advice on the etymology of the new names proposed.

Funding

This work was supported by grants from the European Research Council (ERC) (Consolidator grant 817834 to T.E. and Starting grant 803151 to L.E.), the Swedish Research Council (International Postdoc 2018-06609 to D.T.) and the Volkswagen Foundation (grant no. 96725 to T.E.). This work made use of the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-2953.

Conflict of interest

The authors declare there is no conflict of interest.

848 ## References

849

850 1. Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: The
851 primary kingdoms. Proceedings of the National Academy of Sciences 74:5088-5090.
852 2. Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms:
853 proposal for the domains Archaea, Bacteria, and Eucarya. Proceedings of the
854 National Academy of Sciences 87:4576-4579.
855 3. Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil P-A, Hugenholtz P. 2021.
856 GTDB: an ongoing census of bacterial and archaeal diversity through a
857 phylogenetically consistent, rank normalized and complete genome-based taxonomy.
858 Nucleic Acids Res 50:D785-D794.
859 4. Barns SM, Delwiche CF, Palmer JD, Pace NR. 1996. Perspectives on archaeal diversity,
860 thermophily and monophyly from environmental rRNA sequences. Proceedings of
861 the National Academy of Sciences 93:9188-9193.
862 5. Barns SM, Fundyga RE, Jeffries MW, Pace NR. 1994. Remarkable archaeal diversity
863 detected in a Yellowstone National Park hot spring environment. Proc Natl Acad Sci U
864 S A 91:1609-1613.
865 6. Elkins JG, Podar M, Graham DE, Makarova KS, Wolf Y, Randau L, Hedlund BP,
866 Brochier-Armanet C, Kunin V, Anderson I, Lapidus A, Goltsman E, Barry K, Koonin EV,
867 Hugenholtz P, Kyrpides N, Wanner G, Richardson P, Keller M, Stetter KO. 2008. A
868 korarchaeal genome reveals insights into the evolution of the Archaea. Proceedings
869 of the National Academy of Sciences 105:8102-8107.
870 7. Tahon G, Geesink P, Ettema TJG. 2021. Expanding Archaeal Diversity and Phylogeny:
871 Past, Present, and Future. Annu Rev Microbiol 75:359-381.
872 8. NCBI Resource Coordinators. 2017. Database resources of the National Center for
873 Biotechnology Information. Nucleic Acids Res 46:D8-D13.
874 9. Spang A, Caceres EF, Ettema TJG. 2017. Genomic exploration of the diversity,
875 ecology, and evolution of the archaeal domain of life. Science 357:eaaf3883.
876 10. Dombrowski N, Williams TA, Sun J, Woodcroft BJ, Lee J-H, Minh BQ, Rinke C, Spang A.
877 2020. Undinarchaeota illuminate DPANN phylogeny and the impact of gene transfer
878 on archaeal evolution. Nature Communications 11:3939.
879 11. Guy L, Ettema TJG. 2011. The archaeal 'TACK' superphylum and the origin of
880 eukaryotes. Trends Microbiol 19:580-587.
881 12. Shu W-S, Huang L-N. 2022. Microbial diversity in extreme environments. Nat Rev
882 Microbiol 20:219-235.
883 13. Williams TA, Cox CJ, Foster PG, Szöllősi GJ, Embley TM. 2020. Phylogenomics provides
884 robust support for a two-domains tree of life. Nature Ecology & Evolution 4:138-147.
885 14. Susko E, Roger AJ. 2021. Long Branch Attraction Biases in Phylogenetics. Syst Biol
886 70:838-843.
887 15. Nakamura T, Yamazaki SS, Sakai K, Yamasaki H, Furushima Y, Yamamoto H. 2006.
888 Acroporid corals growing over a methane-bubbling hydrothermal vent, Southern
889 Ryukyu Archipelago. Coral Reefs 25:382-382.
890 16. Furushima Y, Nagao M, Suzuki A, Yamamoto H, Maruyama T. 2009. Periodic Behavior
891 of the Bubble Jet (Geyser) in the Taketomi Submarine Hot Springs of the Southern
892 Part of Yaeyama Archipelago, Japan. Mar Technol Soc J 43:13-22.

893 17. Vick TJ, Dodsworth JA, Costa KC, Shock EL, Hedlund BP. 2010. Microbiology and
894 geochemistry of Little Hot Creek, a hot spring environment in the Long Valley
895 Caldera. Geobiol 8:140-154.
896 18. Saw JH, Spang A, Zaremba-Niedzwiedzka K, Juzokaite L, Dodsworth JA, Murugapiran
897 SK, Colman DR, Takacs-Vesbach C, Hedlund BP, Guy L, Ettema TJG. 2015. Exploring
898 microbial dark matter to resolve the deep archaeal ancestry of eukaryotes.
899 Philosophical Transactions of the Royal Society B: Biological Sciences 370:20140328.
900 19. Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester
901 E, Seitz KW, Anantharaman K, Starnawski P, Kjeldsen KU, Stott MB, Nunoura T,
902 Banfield JF, Schramm A, Baker BJ, Spang A, Ettema TJG. 2017. Asgard archaea
903 illuminate the origin of eukaryotic cellular complexity. Nature 541:353-358.
904 20. Dodsworth JA, Blainey PC, Murugapiran SK, Swingley WD, Ross CA, Tringe SG, Chain
905 PSG, Scholz MB, Lo C-C, Raymond J, Quake SR, Hedlund BP. 2013. Single-cell and
906 metagenomic analyses indicate a fermentative and saccharolytic lifestyle for
907 members of the OP9 lineage. Nature Communications 4:1854.
908 21. Peng Y, Leung HCM, Yiu SM, Chin FYL. 2012. IDBA-UD: a de novo assembler for single-
909 cell and metagenomic sequencing data with highly uneven depth. Bioinformatics
910 28:1420-1428.
911 22. Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP, Banfield JF.
912 2009. Community-wide analysis of microbial genome sequence signatures. Genome
913 Biol 10:R85.
914 23. Ultsch A, Mörchen F. ESOM-Maps : tools for clustering , visualization , and
915 classification with Emergent SOM, p. *In* (ed),
916 24. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. 2013.
917 Genome sequences of rare, uncultured bacteria obtained by differential coverage
918 binning of multiple metagenomes. Nat Biotechnol 31:533-538.
919 25. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO.
920 2012. The SILVA ribosomal RNA gene database project: improved data processing
921 and web-based tools. Nucleic Acids Res 41:D590-D596.
922 26. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM:
923 assessing the quality of microbial genomes recovered from isolates, single cells, and
924 metagenomes. Genome Res 25:1043-1055.
925 27. Chklovski A, Parks DH, Woodcroft BJ, Tyson GW. 2022. CheckM2: a rapid, scalable
926 and accurate tool for assessing microbial genome quality using machine learning.
927 bioRxiv doi:10.1101/2022.07.11.499243:2022.07.11.499243.
928 28. Martijn J, Vosseberg J, Guy L, Offre P, Ettema TJG. 2018. Deep mitochondrial origin
929 outside the sampled alphaproteobacteria. Nature 557:101-105.
930 29. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – Approximately Maximum-Likelihood
931 Trees for Large Alignments. PLoS One 5:e9490.
932 30. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A,
933 Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic
934 Inference in the Genomic Era. Molecular Biology and Evolution 37:1530-1534.
935 31. Katoh K, Misawa K, Kuma Ki, Miyata T. 2002. MAFFT: a novel method for rapid
936 multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res
937 30:3059-3066.

938 32. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for
939 automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics
940 25:1972-1973.
941 33. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2017. UFBoot2:
942 Improving the Ultrafast Bootstrap Approximation. Molecular Biology and Evolution
943 35:518-522.
944 34. Letunic I, Bork P. 2021. Interactive Tree Of Life (iTOL) v5: an online tool for
945 phylogenetic tree display and annotation. Nucleic Acids Res 49:W293-W296.
946 35. Lagkouvardos I, Joseph D, Kapfhammer M, Giritli S, Horn M, Haller D, Clavel T. 2016.
947 IMNGS: A comprehensive open resource of processed 16S rRNA microbial profiles for
948 ecology and diversity studies. Sci Rep 6:33721-33721.
949 36. Lee MD. 2019. GToTree: A user-friendly workflow for phylogenomics. Bioinformatics
950 35:4162-4164.
951 37. Szöllősi GJ, Tannier E, Daubin V, Boussau B. 2014. The Inference of Gene Trees with
952 Species Trees. Syst Biol 64:e42-e62.
953 38. Petitjean C, Deschamps P, López-García P, Moreira D, Brochier-Armanet C. 2015.
954 Extending the Conserved Phylogenetic Core of Archaea Disentangles the Evolution of
955 the Third Domain of Life. Molecular Biology and Evolution 32:1242-1254.
956 39. Eme L, Tamarit D, Caceres EF, Stairs CW, Anda VD, Schön ME, Seitz KW, Dombrowski
957 N, Lewis WH, Homa F, Saw JH, Lombard J, Nunoura T, Li W-J, Hua Z-S, Chen L-X,
958 Banfield JF, John ES, Reysenbach A-L, Stott MB, Schramm A, Kjeldsen KU, Teske AP,
959 Baker BJ, Ettema TJG. 2023. Inference and reconstruction of the heimdallarchaeial
960 ancestry of eukaryotes. bioRxiv doi:10.1101/2023.03.07.531504:2023.03.07.531504.
961 40. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal:
962 prokaryotic gene recognition and translation initiation site identification. BMC
963 Bioinformatics 11:119.
964 41. Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version
965 7: Improvements in Performance and Usability. Molecular Biology and Evolution
966 30:772-780.
967 42. Criscuolo A, Gribaldo S. 2010. BMGE (Block Mapping and Gathering with Entropy): a
968 new software for selection of phylogenetic informative regions from multiple
969 sequence alignments. BMC Evol Biol 10:210.
970 43. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2014. IQ-TREE: A Fast and
971 Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.
972 Molecular Biology and Evolution 32:268-274.
973 44. Susko E, Roger AJ. 2007. On Reduced Amino Acid Alphabets for Phylogenetic
974 Inference. Molecular Biology and Evolution 24:2139-2150.
975 45. Viklund J, Ettema TJG, Andersson SGE. 2011. Independent Genome Reduction and
976 Phylogenetic Reclassification of the Oceanic SAR11 Clade. Molecular Biology and
977 Evolution 29:599-615.
978 46. Wang H-C, Minh BQ, Susko E, Roger AJ. 2017. Modeling Site Heterogeneity with
979 Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic
980 Estimation. Syst Biol 67:216-235.
981 47. Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes MPI: Phylogenetic
982 Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment. Syst Biol
983 62:611-615.

48. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, Bork P. 2017. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. Molecular Biology and Evolution 34:2115-2122.
49. Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20:238.
50. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods 14:587-589.
51. Schön ME, Martijn J, Vosseberg J, Köstlbacher S, Ettema TJG. 2022. The evolutionary origin of host association in the Rickettsiales. Nature Microbiology 7:1189-1199.
52. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, Busk PK, Xu Y, Yin Y. 2018. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res 46:W95-W101.
53. Elbourne Liam DH, Tetu SG, Hassan KA, Paulsen IT. 2016. TransportDB 2.0: a database for exploring membrane transporters in sequenced genomes from all domains of life. Nucleic Acids Res 45:D320-D324.
54. Sauer DB, Wang D-N. 2019. Predicting the optimal growth temperatures of prokaryotes using only genome derived features. Bioinformatics 35:3224-3231.
55. Musil M, Khan RT, Beier A, Stourac J, Konegger H, Damborsky J, Bednar D. 2020. FireProtASR: A Web Server for Fully Automated Ancestral Sequence Reconstruction. Brief Bioinform 22.
56. Yang Y, Zhao J, Zeng L, Vihinen M. 2022. ProTstab2 for Prediction of Protein Thermal Stabilities. Int J Mol Sci 23:10798.
57. The UniProt Consortium. 2022. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res 51:D523-D531.
58. Vulcano F, Hribovšek P, Denny EO, Steen IH, Stokke R. 2023. Potential for homoacetogenesis via the Wood–Ljungdahl pathway in Korarchaeia lineages from marine hydrothermal vents. Environ Microbiol Rep n/a.
59. Louca S, Doebeli M. 2017. Efficient comparative phylogenetics on large trees. Bioinformatics 34:1053-1055.
60. Xie R, Wang Y, Huang D, Hou J, Li L, Hu H, Zhao X, Wang F. 2022. Expanding Asgard members in the domain of Archaea sheds new light on the origin of eukaryotes. Science China Life Sciences 65:818-829.
61. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. 2019. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics 36:1925-1927.
62. Pan J, Zhang X, Xu W, Liu Y, Liu L, Luo Z, Li M. 2023. Wood–Ljungdahl pathway found in novel marine Korarchaeota groups illuminates their evolutionary history. mSystems 0:e00305-23.
63. Konstantinidis KT, Rosselló-Móra R, Amann R. 2017. Uncultivated microbes in need of their own taxonomy. The ISME Journal 11:2399-2406.
64. Hedlund BP, Chuvochina M, Hugenholtz P, Konstantinidis KT, Murray AE, Palmer M, Parks DH, Probst AJ, Reysenbach A-L, Rodriguez-R LM, Rossello-Mora R, Sutcliffe IC, Venter SN, Whitman WB. 2022. SeqCode: a nomenclatural code for prokaryotes described from sequence data. Nature Microbiology 7:1702-1708.
65. Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet 22:225-231.

1030 66. Yutin N, Puigbò P, Koonin EV, Wolf YI. 2012. Phylogenomics of Prokaryotic Ribosomal
1031 Proteins. PLoS One 7:e36972.
1032 67. Qu Y-N, Rao Y-Z, Qi Y-L, Li Y-X, Li A, Palmer M, Hedlund BP, Shu W-S, Evans PN, Nie G-
1033 X, Hua Z-S, Li W-J. 2023. Panguiarchaeum symbiosum, a potential hyperthermophilic
1034 symbiont in the TACK superphylum. Cell Rep 42:112158.
1035 68. Liu Y, Li M. 2022. The unstable evolutionary position of Korarchaeota and its
1036 relationship with other TACK and Asgard archaea. mLife 1:218-222.
1037 69. Miller-Coleman RL, Dodsworth JA, Ross CA, Shock EL, Williams AJ, Hartnett HE,
1038 McDonald AI, Havig JR, Hedlund BP. 2012. Korarchaeota Diversity, Biogeography, and
1039 Abundance in Yellowstone and Great Basin Hot Springs and Ecological Niche
1040 Modeling Based on Machine Learning. PLoS One 7:e35964.
1041 70. Auchtung TA, Shyndriayeva G, Cavanaugh CM. 2011. 16S rRNA phylogenetic analysis
1042 and quantification of Korarchaeota indigenous to the hot springs of Kamchatka,
1043 Russia. Extremophiles 15:105-116.
1044 71. Blamey JM, Adams MWW. 1993. Purification and characterization of pyruvate
1045 ferredoxin oxidoreductase from the hyperthermophilic archaeon Pyrococcus
1046 furiosus. Biochim Biophys Acta 1161:19-27.
1047 72. Frolov EN, Kublanov IV, Toshchakov SV, Lunev EA, Pimenov NV, Bonch-Osmolovskaya
1048 EA, Lebedinsky AV, Chernyh NA. 2019. Form III RubisCO-mediated transaldolase
1049 variant of the Calvin cycle in a chemolithoautotrophic bacterium. Proceedings of the
1050 National Academy of Sciences 116:18638-18646.
1051 73. Sato T, Atomi H, Imanaka T. 2007. Archaeal Type III RuBisCOs Function in a Pathway
1052 for AMP Metabolism. Science 315:1003-1006.
1053 74. Orita I, Sato T, Yurimoto H, Kato N, Atomi H, Imanaka T, Sakai Y. 2006. The Ribulose
1054 Monophosphate Pathway Substitutes for the Missing Pentose Phosphate Pathway in
1055 the Archaeon Thermococcus kodakaraensis. J Bacteriol 188:4698-4704.
1056 75. Frank KL, Rogers DR, Olins HC, Vidoudez C, Girguis PR. 2013. Characterizing the
1057 distribution and rates of microbial sulfate reduction at Middle Valley hydrothermal
1058 vents. The ISME Journal 7:1391-1401.
1059 76. Berghuis BA, Yu FB, Schulz F, Blainey PC, Woyke T, Quake SR. 2019. Hydrogenotrophic
1060 methanogenesis in archaeal phylum Verstraetearchaeota reveals the shared ancestry
1061 of all methanogens. Proceedings of the National Academy of Sciences 116:5037-
1062 5044.
1063 77. Appel L, Willistein M, Dahl C, Ermler U, Boll M. 2021. Functional diversity of
1064 prokaryotic HdrA(BC) modules: Role in flavin-based electron bifurcation processes
1065 and beyond. Biochim Biophys Acta 1862:148379.
1066 78. Thauer RK, Kaster A-K, Goenrich M, Schick M, Hiromoto T, Shima S. 2010.
1067 Hydrogenases from Methanogenic Archaea, Nickel, a Novel Cofactor, and H2 Storage.
1068 Annu Rev Biochem 79:507-536.
1069 79. Kikuchi A, Asai K. 1984. Reverse gyrase—a topoisomerase which introduces positive
1070 superhelical turns into DNA. Nature 309:677-681.
1071 80. Lipscomb GL, Hahn EM, Crowley AT, Adams MWW. 2017. Reverse gyrase is essential
1072 for microbial growth at 95 °C. Extremophiles 21:603-608.
1073 81. Williams TA, Szöllősi GJ, Spang A, Foster PG, Heaps SE, Boussau B, Ettema TJG,
1074 Embley TM. 2017. Integrative modeling of gene and genome evolution roots the
1075 archaeal tree of life. Proceedings of the National Academy of Sciences 114:E4602-
1076 E4611.

1077 82. Akanuma S, Nakajima Y, Yokobori S-i, Kimura M, Nemoto N, Mase T, Miyazono K-i,
1078 Tanokura M, Yamagishi A. 2013. Experimental evidence for the thermophilicity of
1079 ancestral life. Proceedings of the National Academy of Sciences 110:11067-11072.
1080 83. McKay LJ, Dlakić M, Fields MW, Delmont TO, Eren AM, Jay ZJ, Klingelsmith KB, Rusch
1081 DB, Inskeep WP. 2019. Co-occurring genomic capacity for anaerobic methane and
1082 dissimilatory sulfur metabolisms discovered in the Korarchaeota. Nature
1083 Microbiology 4:614-622.
1084 84. Amin K, Tranchimand S, Benvegnu T, Abdel-Razzak Z, Chamieh H. 2021. Glycoside
1085 Hydrolases and Glycosyltransferases from Hyperthermophilic Archaea: Insights on
1086 Their Characteristics and Applications in Biotechnology. Biomolecules 11:1557.
1087 85. Saier MH. 2000. A Functional-Phylogenetic Classification System for Transmembrane
1088 Solute Transporters. Microbiology and Molecular Biology Reviews 64:354-411.
1089 86. Steens JA, Salazar CRP, Staals RHJ. 2022. The diverse arsenal of type III CRISPR–Cas-
1090 associated CARF and SAVED effectors. Biochem Soc Trans 50:1353-1364.
1091 87. Makarova KS, Wolf YI, Iranzo J, Shmakov SA, Alkhnbashi OS, Brouns SJJ, Charpentier
1092 E, Cheng D, Haft DH, Horvath P, Moineau S, Mojica FJM, Scott D, Shah SA, Siksnys V,
1093 Terns MP, Venclovas Č, White MF, Yakunin AF, Yan W, Zhang F, Garrett RA, Backofen
1094 R, van der Oost J, Barrangou R, Koonin EV. 2020. Evolutionary classification of
1095 CRISPR–Cas systems: a burst of class 2 and derived variants. Nat Rev Microbiol 18:67-
1096 83.
1097 88. Makarova KS, Wolf YI, Koonin EV. 2003. Potential genomic determinants of
1098 hyperthermophily. Trends Genet 19:172-176.
1099 89. Gudbergsdóttir SR, Menzel P, Krogh A, Young M, Peng X. 2016. Novel viral genomes
1100 identified from six metagenomes reveal wide distribution of archaeal viruses and
1101 high viral diversity in terrestrial hot springs. Environ Microbiol 18:863-874.
1102 90. Goldfarb T, Sberro H, Weinstock E, Cohen O, Doron S, Charpak-Amikam Y, Afik S, Ofir
1103 G, Sorek R. 2015. BREX is a novel phage resistance system widespread in microbial
1104 genomes. The EMBO Journal 34:169-183.
1105

1106

1107 ## Table Captions

1108

1109 Table 1. Overview of Bayesian phylogenetic analyses results for the five New Marker
1110 (NM54) alignments.

1111

1112 ## Figure Captions

1113

1114 **Fig. 1**. Phylogenetic tree corresponding to the Bayesian consensus tree (4 chains)
1115 reconstructed from the χ2-trimmed alignment (50% most heterogeneous sites
1116 removed; 7594 alignment positions) of 54 concatenated new marker proteins
1117 (NM54). *Ca*. Korarchaeota clades originating from hot springs and hydrothermal
1118 vents have a light and dark blue background shading, respectively. The two
1119 transitions between hydrothermal vents and hot springs are indicated with a star
1120 symbol. Scale bar indicates 0.1 substitutions per site.

1121

1122 **Fig. 2**. Overview of the Korarchaeota clade from Fig. 1. Internal nodes are annotated
1123 with the ancestor name/number code and the number of inferred coding sequences
1124 from the ALE analysis. Branches are annotated with bars representing the number of
1125 originations, transfers, duplications and losses. Bar heights in the legend correspond
1126 to 1000 such events each. Raw values can be found in Table S9. Terminal branches
1127 are labelled with the MAG name and accession number. Additional MAG statistics
1128 are given between square brackets: estimated genome size; GC content; number of
1129 CDS; estimated number of CDS.

1130

1131 **Fig. 3**. Ancestral gene content reconstruction of the Last Korarchaeota common
1132 ancestor (LKorCA). Genes were considered present if it was inferred by ALE with a
1133 copy number ≥ 0.3. For protein complexes, at least half the subunits had to be
1134 inferred as mentioned before.

1135

1136 **Fig. 4**. Ancestral gene content reconstruction of the Last *Ca*. Korarchaeum, *Ca*.
1137 Methanodesulfokora and *Ca*. Fumariihydrokora common ancestor (LKMFCA), which
1138 highlights the first transition from hydrothermal vents to hot springs. Genes were

1139 considered present if it was inferred by ALE with a copy number ≥ 0.3. For protein
1140 complexes, at least half the subunits had to be inferred as mentioned before.
1141
1142 **Fig. 5**. Ancestral gene content reconstruction of the Last *Ca*. Korarchaeum and *Ca*.
1143 Fumariihydrokora common ancestor (LKFCA), which highlights the second transition
1144 from hydrothermal vents to hot springs. Genes were considered present if it was
1145 inferred by ALE with a copy number ≥ 0.3. For protein complexes, at least half the
1146 subunits had to be inferred as mentioned before.
1147
1148 **Fig. 6**. Ancestral gene content reconstruction of the Last *Ca*. Methanodesulfokora
1149 common ancestor (LMCA). Genes were considered present if it was inferred by ALE
1150 with a copy number ≥ 0.3. For protein complexes, at least half the subunits had to be
1151 inferred as mentioned before.
1152

Table 1. Overview of Bayesian phylogenetic analyses for the five NM54 alignments.

	Untreated	SR4 recoded	Custom recoded	50% least heterogeneous sites (χ^2-trimmed)	50% fastest evolving sites trimmed
General alignment statistics					
No. of sites	15142	15142	15142	7594	7571
Model of evolution	LG+C60+F+G4	GTR+G4+C60SR4	LG+C60+F+G4	LG+C60+F+G4	LG+C60+F+G4
No. of informative sites	13862	11810	13427	6341	6291
Missing data (%)	8.80	8.80	8.80	8.59	8.51
PhyloBayes sampling statistics					
Chains used for optimal Maxdiff	4	2	2	4	2
Cycles	53599 - 54546	69504 - 69901	67722 - 68016	69911 - 70485	68356 - 70038
Burnin	10000	20000	20000	20000	20000
Step	30	30	30	30	30
Effective sampling size	>287	>15702	>16772	>3399	>1510
Optimal Maxdiff	1	0.049	0.217	1	0.053
Maxdiff 4 chains	1	0.397	1	1	1
Posterior predictive tests					
Maximum squared heterogeneity (p-value)	0	0	0	0.36 - 0.41	0
Maximum squared heterogeneity (z-score)	185.41 - 195.53	45.74 - 49.51	36.98 - 39.85	0.06 - 0.13	33.33 - 35.38
Mean squared heterogeneity (p-value)	0	0	0	1	0
Mean squared heterogeneity (z-score)	464.51 - 502.73	97.00 - 100.48	131.87 - 140.52	-7.38 - -7.15	73.08 - 76.04
Mean diversity per site (p-value)	0	0.47 - 0.50	0.04 - 0.06	0	0
Mean diversity per site (z-score)	13.32 - 13.87	0.03 - 0.09	1.56 - 1.68	7.87 – 8.01	6.81 - 7.20

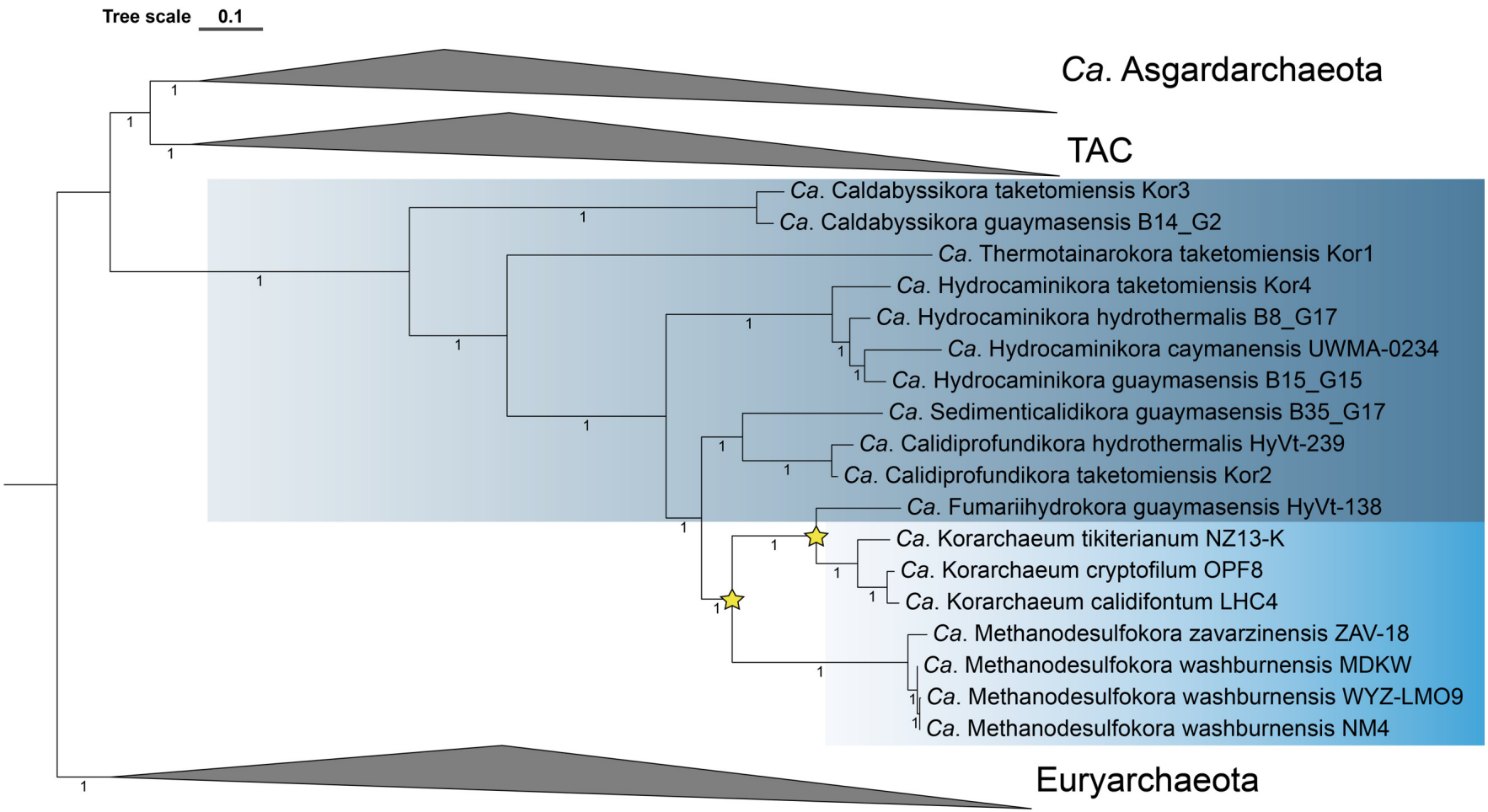

Fig. 1. Phylogenetic tree corresponding to the Bayesian consensus tree (4 chains) reconstructed from the χ2-trimmed alignment (50% most heterogeneous sites removed; 7594 alignment positions) of 54 concatenated new marker proteins (NM54). *Ca*. Korarchaeota clades originating from hot springs and hydrothermal vents have a light and dark blue background shading, respectively. The two

transitions between hydrothermal vents and hot springs are indicated with a star symbol. Scale bar indicates 0.1 substitutions per site.

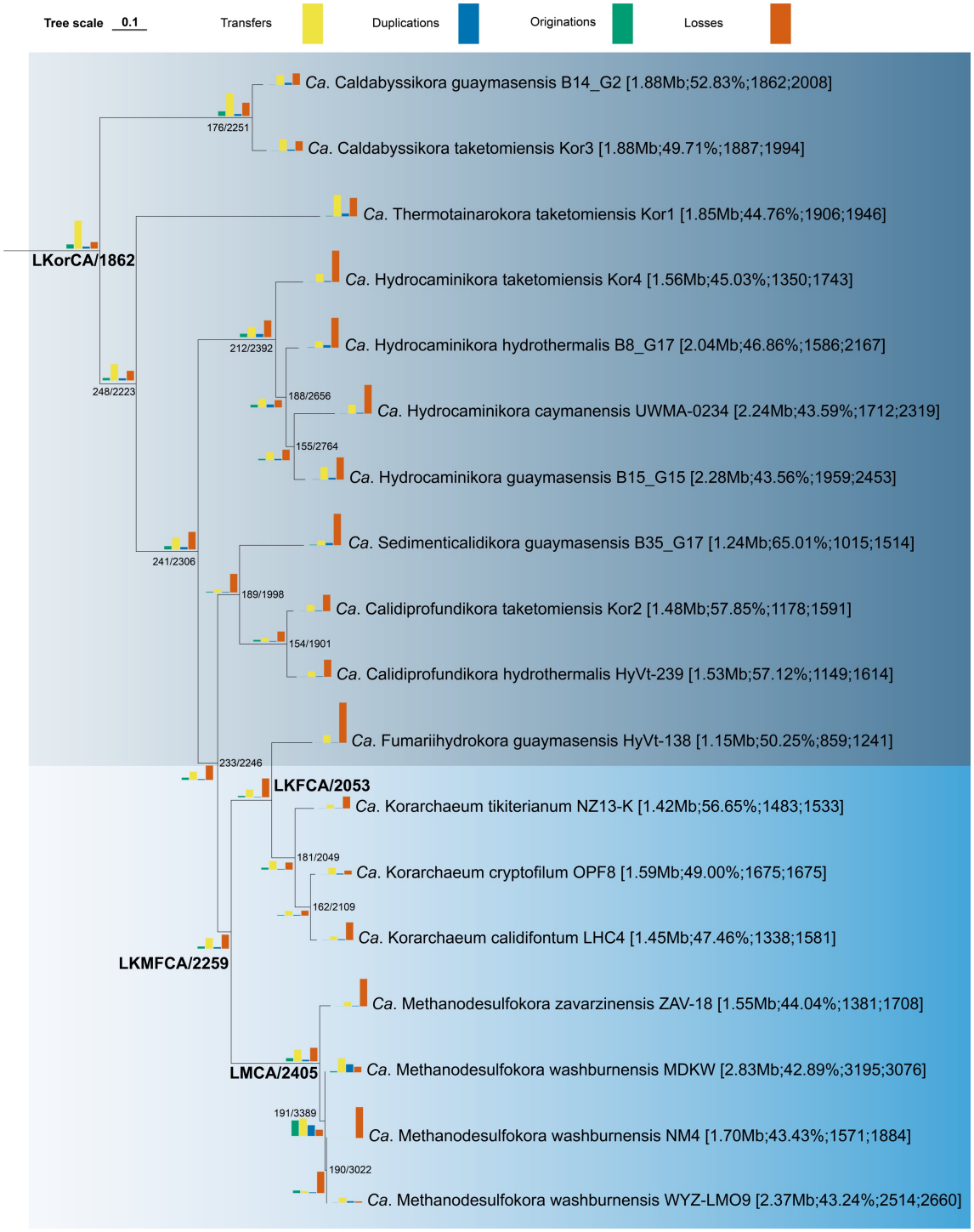

Fig. 2. Overview of the Korarchaeota clade from Fig. 1. Internal nodes are annotated with the ancestor name/number code and the number of inferred coding sequences from the ALE analysis. Branches are annotated with bars representing the number of originations, transfers, duplications and losses. Bar heights in the legend correspond to 1000 such events each. Raw values can be found in Table S9. Terminal branches

are labelled with the MAG name and accession number. Additional MAG statistics are given between square brackets: estimated genome size; GC content; number of CDS; estimated number of CDS.

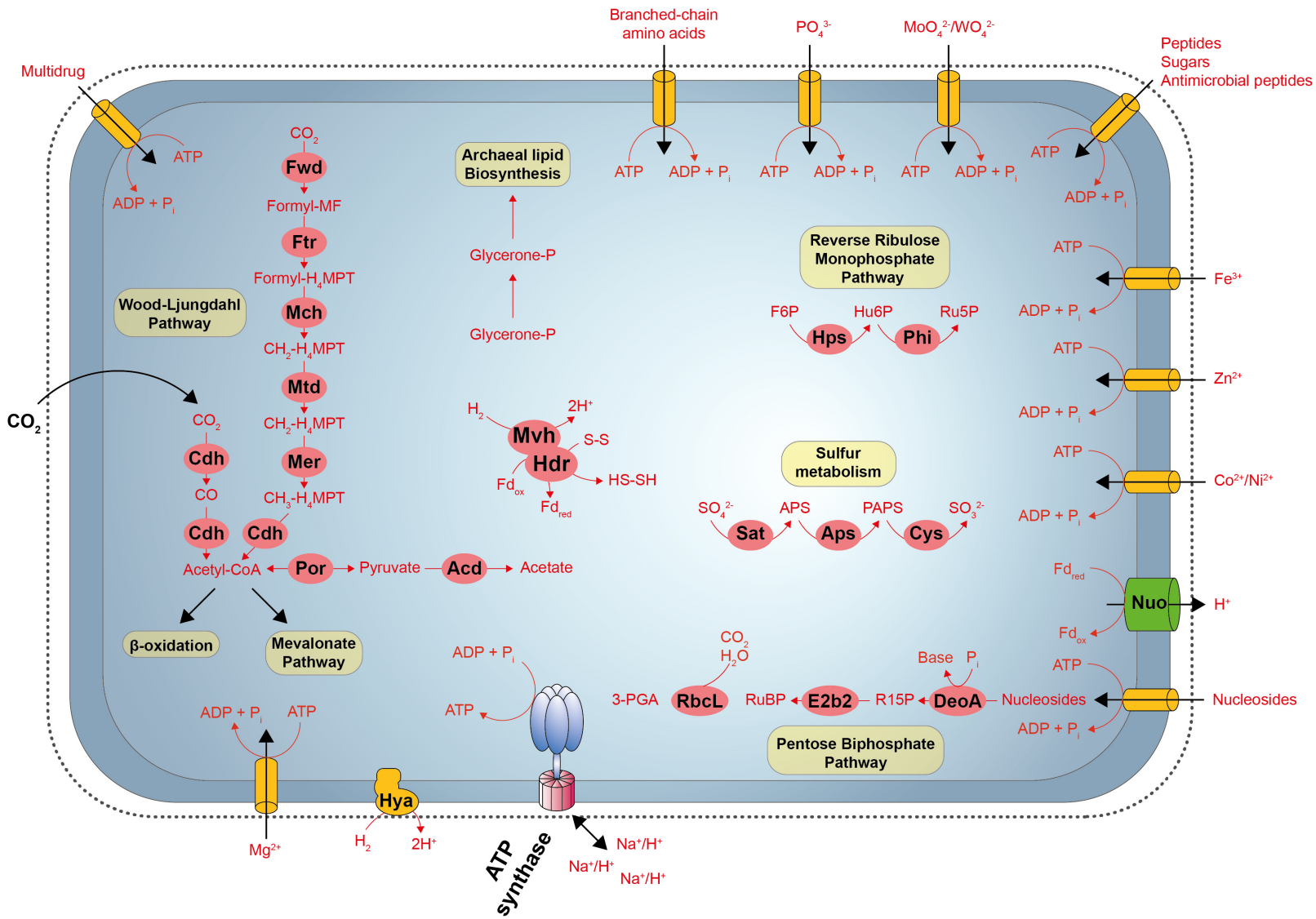
Multidrug
ATP
ADP + Pi
Branched-chain amino acids
PO4 3-
MoO4 2-/WO4 2-
Peptides
Sugars
Antimicrobial peptides
ATP
ADP + Pi
CO2
Fwd
Formyl-MF
Ftr
Formyl-H4MPT
Mch
CH2-H4MPT
Mtd
CH2-H4MPT
Mer
CH3-H4MPT
Wood-Ljungdahl Pathway
CO2
Cdh
CO
Cdh
Cdh
Acetyl-CoA
Por
Pyruvate
Acd
Acetate
β-oxidation
Mevalonate Pathway
Archaeal lipid Biosynthesis
Glycerone-P
Glycerone-P
H2
2H+
Mvh
S-S
Hdr
HS-SH
Fdox
Fdred
Reverse Ribulose Monophosphate Pathway
F6P
Hps
Hu6P
Phi
Ru5P
Sulfur metabolism
SO4 2-
Sat
APS
Aps
PAPS
Cys
SO3 2-
Fe3+
Zn2+
Co2+/Ni2+
Fdred
Nuo
H+
Fdox
CO2
H2O
3-PGA
RbcL
RuBP
E2b2
R15P
Base
Pi
DeoA
Nucleosides
Nucleosides
Pentose Biphosphate Pathway
ADP + Pi
ATP
Mg2+
Hya
H2
2H+
ATP synthase
Na+/H+
Na+/H+
Na+/H+

Fig. 3. Ancestral gene content reconstruction of the Last Korarchaeota common ancestor (LKorCA). Genes were considered present if it was inferred by ALE with a copy number ≥ 0.3. For protein complexes, at least half the subunits had to be inferred as mentioned before.

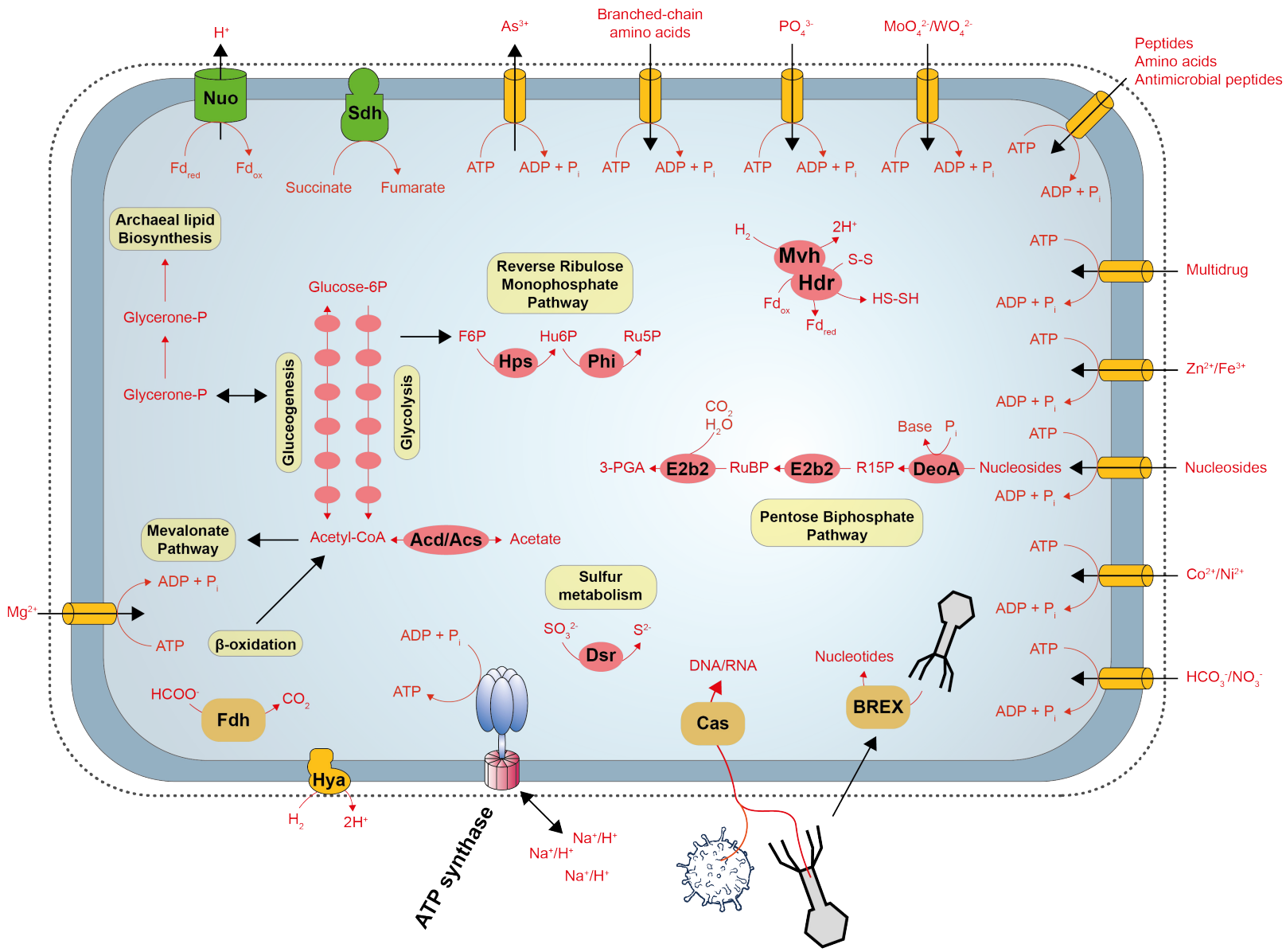
H+
Nuo
Fdred
Fdox
Sdh
Succinate
Fumarate
As3+
ATP
ADP + Pi
Branched-chain amino acids
PO43-
MoO42-/WO42-
Peptides
Amino acids
Antimicrobial peptides
Archaeal lipid Biosynthesis
Glycerone-P
Glucose-6P
Gluceogenesis
Glycolysis
Reverse Ribulose Monophosphate Pathway
F6P
Hu6P
Ru5P
Hps
Phi
H2
2H+
Mvh
Hdr
S-S
HS-SH
Multidrug
Zn2+/Fe3+
CO2
H2O
Base
3-PGA
E2b2
RuBP
R15P
DeoA
Nucleosides
Pentose Biphosphate Pathway
Mevalonate Pathway
Acetyl-CoA
Acd/Acs
Acetate
Co2+/Ni2+
Mg2+
β-oxidation
Sulfur metabolism
SO32-
S2-
Dsr
HCO3-/NO3-
HCOO-
Fdh
CO2
DNA/RNA
Cas
Nucleotides
BREX
Hya
ATP synthase
Na+/H+

Fig. 4. Ancestral gene content reconstruction of the Last *Ca*. Korarchaeum, *Ca*. Methanodesulfokora and *Ca*. Fumariihydrokora common ancestor (LKMFCA), which highlights the first transition from hydrothermal vents to hot springs. Genes were considered present if it was inferred by ALE with a copy number ≥ 0.3. For protein complexes, at least half the subunits had to be inferred as mentioned before.

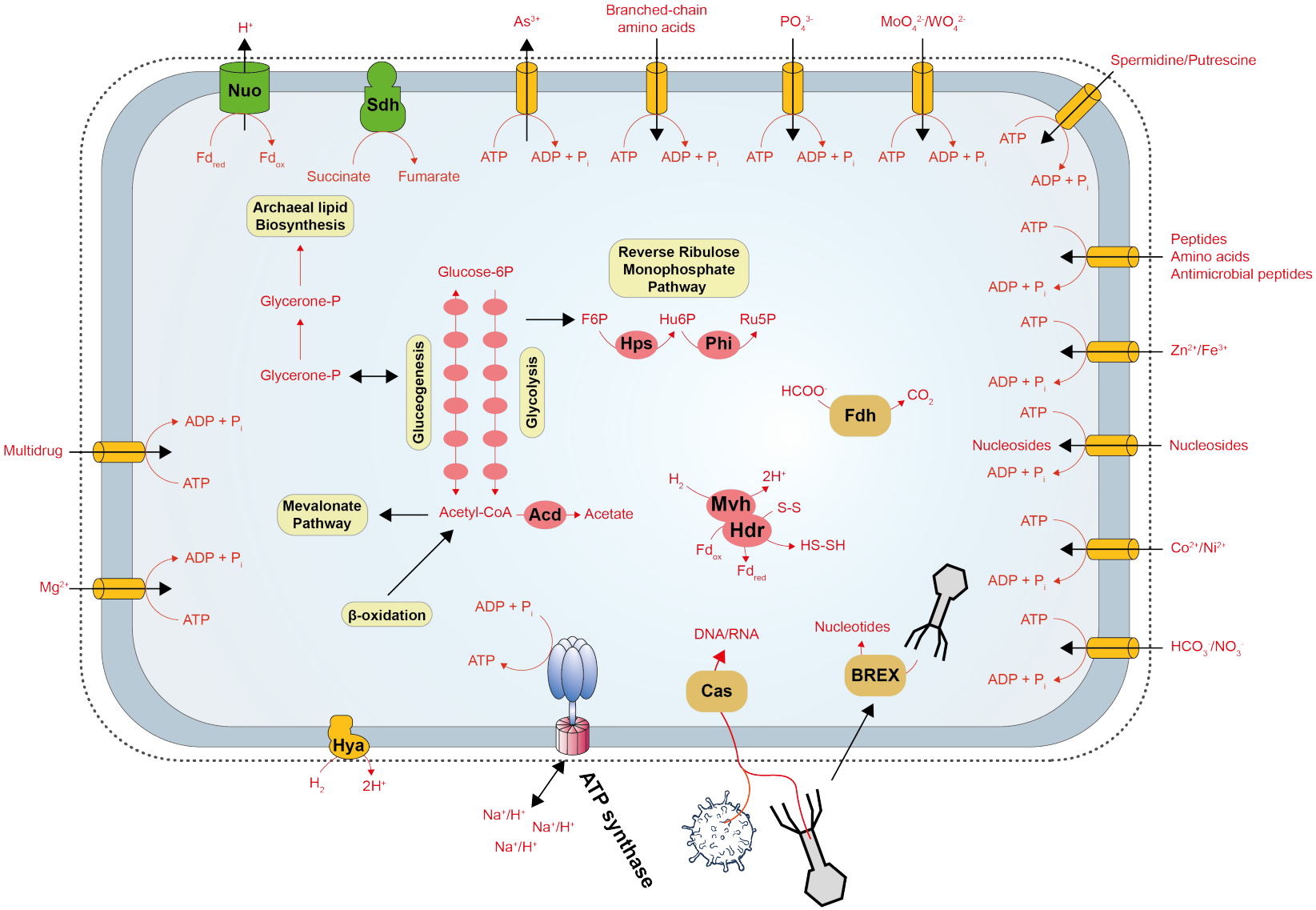
H+
Nuo
Fd red
Fd ox
Sdh
Succinate
Fumarate
As3+
Branched-chain amino acids
PO4 3-
MoO4 2-/WO4 2-
ATP
ADP + Pi
Spermidine/Putrescine
Archaeal lipid Biosynthesis
Glycerone-P
Glucose-6P
Reverse Ribulose Monophosphate Pathway
F6P
Hps
Hu6P
Phi
Ru5P
Gluceogenesis
Glycolysis
Peptides
Amino acids
Antimicrobial peptides
Zn2+/Fe3+
HCOO-
Fdh
CO2
Nucleosides
Multidrug
Mevalonate Pathway
Acetyl-CoA
Acd
Acetate
H2
2H+
Mvh
Hdr
S-S
HS-SH
Co2+/Ni2+
Mg2+
β-oxidation
DNA/RNA
Nucleotides
Cas
BREX
HCO3-/NO3-
Hya
Na+/H+
ATP synthase

Fig. 5. Ancestral gene content reconstruction of the Last *Ca*. Korarchaeum and *Ca*. Fumariihydrokora common ancestor (LKFCA), which highlights the second transition from hydrothermal vents to hot springs. Genes were considered present if it was inferred by ALE with a copy number ≥ 0.3. For protein complexes, at least half the subunits had to be inferred as mentioned before.

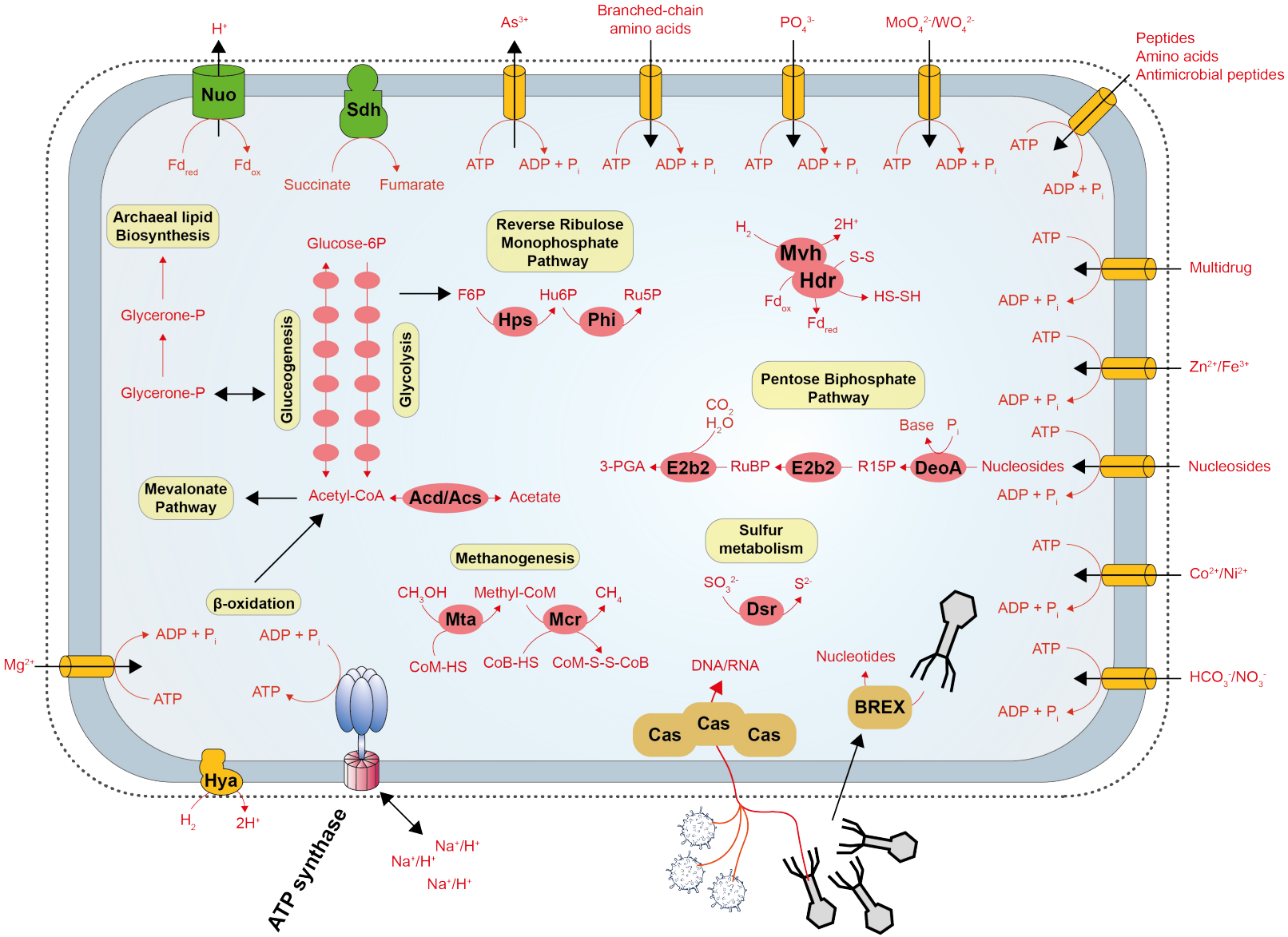
H+
Nuo
Fdred
Fdox
Sdh
Succinate
Fumarate
As3+
ATP
ADP + Pi
Branched-chain amino acids
PO43-
MoO42-/WO42-
Peptides
Amino acids
Antimicrobial peptides
Archaeal lipid Biosynthesis
Glycerone-P
Glucose-6P
Gluceogenesis
Glycolysis
Reverse Ribulose Monophosphate Pathway
F6P
Hps
Hu6P
Phi
Ru5P
H2
2H+
Mvh
Hdr
S-S
HS-SH
Multidrug
Zn2+/Fe3+
Pentose Biphosphate Pathway
CO2
H2O
Base
Pi
3-PGA
E2b2
RuBP
R15P
DeoA
Nucleosides
Mevalonate Pathway
Acetyl-CoA
Acd/Acs
Acetate
Sulfur metabolism
SO32-
S2-
Dsr
Methanogenesis
CH3OH
Methyl-CoM
CH4
Mta
Mcr
CoM-HS
CoB-HS
CoM-S-S-CoB
Co2+/Ni2+
β-oxidation
Mg2+
DNA/RNA
Nucleotides
BREX
HCO3-/NO3-
Cas
Hya
H2
2H+
ATP synthase
Na+/H+

Fig. 6. Ancestral gene content reconstruction of the Last *Ca*. Methanodesulfokora common ancestor (LMCA). Genes were considered present if it was inferred by ALE with a copy number ≥ 0.3. For protein complexes, at least half the subunits had to be inferred as mentioned before.