Complex evolutionary history of translation Elongation Factor 2 and diphthamide biosynthesis in Archaea and parabasalids

Adrienne B. Narrowe[1,†], Anja Spang[2,3,†], Courtney W. Stairs[2], Eva F. Caceres[2], Brett J. Baker[4], Christopher S. Miller[1,‡,*] and Thijs J. G. Ettema[2,‡]

† These authors contributed equally to this work.

‡These authors contributed equally to this work.

*Author for correspondence: Christopher S. Miller, Department of Integrative Biology, University of Colorado Denver, Denver, CO, USA, 303-315-7665, chris.miller@ucdenver.edu

[1] Department of Integrative Biology, University of Colorado Denver, Denver, Colorado, USA

[2] Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden

[3] NIOZ, Royal Netherlands Institute for Sea Research, Department of Marine Microbiology and Biogeochemistry, and Utrecht University, AB Den Burg, The Netherlands

[4] Department of Marine Science, Marine Science Institute, University of Texas Austin, Port Aransas, TX, USA

Running title: EF-2 and diphthamide in archaea and eukaryotes

ABSTRACT

Diphthamide is a modified histidine residue which is uniquely present in archaeal and eukaryotic elongation factor 2 (EF-2), an essential GTPase responsible for catalyzing the coordinated translocation of tRNA and mRNA through the ribosome. In part due to the role of diphthamide in maintaining translational fidelity, it was previously assumed that diphthamide biosynthesis genes (*dph*) are conserved across all eukaryotes and archaea. Here, comparative analysis of new and existing genomes reveals that some archaea (i.e., members of the Asgard superphylum, *Geoarchaea*, and *Korarchaeota*) and eukaryotes (i.e., parabasalids) lack *dph*. In addition, while EF-2 was thought to exist as a single copy in archaea, many of these *dph*-lacking archaeal genomes encode a second EF-2 paralog missing key-residues required for diphthamide modification and for normal translocase function, perhaps suggesting functional divergence linked to loss of diphthamide biosynthesis. Interestingly, some Heimdallarchaeota previously suggested to be most closely related to the eukaryotic ancestor maintain *dph* genes and a single gene encoding canonical EF-2. Our findings reveal that the ability to produce diphthamide, once thought to be a universal feature in archaea and eukaryotes, has been lost multiple times during evolution, and suggest that anticipated compensatory mechanisms evolved independently.

Key words: Asgard, Korarchaeota, Trichomonas, metagenomics, EF-2

INTRODUCTION

Elongation factor 2 (EF-2) is a critical component of the translational machinery that interacts with both the small and large ribosomal subunits. EF-2 functions at the decoding center of the ribosome, where it is necessary for the translocation of messenger RNA and associated tRNAs (Spahn, et al. 2004). Archaeal and eukaryotic EF-2, as well as the homologous bacterial EF-G, are members of the highly conserved translational GTPase protein superfamily (Atkinson 2015). Gene duplications and subsequent neo-functionalizations have been inferred for eukaryotic EF-2 (eEF-2), with the identification of the spliceosome component Snu114 (Fabrizio, et al. 1997), and Ria1, a 60S ribosomal subunit biogenesis factor (Becam, et al. 2001). Bacterial EF-G is involved in both translocation and ribosome recycling and has undergone multiple duplications, including sub-functionalizations separating the translocation and ribosome recycling functions (Suematsu, et al. 2010; Tsuboi, et al. 2009) as well as neo-functionalizations including roles in back-translocation(Qin, et al. 2006), translation termination (Freistroffer, et al. 1997), regulation(Li, et al. 2014) and tetracycline resistance (Donhofer, et al. 2012). However, to date, archaea were thought to encode only a single essential protein within this superfamily, i.e. archaeal EF-2 (aEF-2) (Atkinson 2015).

Unlike bacterial EF-Gs, archaeal and eukaryotic EF-2s contain a post-translationally modified amino acid which is synthesized upon the addition of a 3-amino-3-carboxypropyl (ACP) group to a conserved histidine residue and its subsequent modification to diphthamide by the concerted action of 3 (in archaea) to 7 enzymes (in eukaryotes)(de Crécy-Lagard, et al. 2012; Schaffrath, et al. 2014). While diphthamide is perhaps best known as the target site of bacterial ADP-ribosylating toxins (Iglewski, et al. 1977; Jorgensen, et al. 2008) and as required for sensitivity to the antifungal sordarin (Botet, et al. 2008), its exact role remains a subject of

64 investigation. Yeast mutants incapable of synthesizing diphthamide have a higher rate of
65 translational frame shifts, suggesting that this residue plays a critical role in reading frame
66 fidelity during translation (Ortiz, et al. 2006). Furthermore, structural studies of eEF-2 using
67 high-resolution Cryo-EM have indicated that diphthamide interacts directly with codon-
68 anticodon bases in the translating ribosome, and facilitates translocation by displacing ribosomal
69 decoding bases (Anger, et al. 2013; Murray, et al. 2016). In addition, diphthamide has been
70 proposed to play a role in the regulation of translation, as it represents a site for reversible
71 endogenous ADP-ribosylation (Schaffrath, et al. 2014), and in the selective translation of certain
72 genes in response to cellular stress (Argüelles, et al. 2014). Given its anticipated role at the core
73 of the translational machinery, it is not surprising that, with the sole exception of *Korarchaeum*
74 *cryptofilum* (de Crécy-Lagard, et al. 2012; Elkins, et al. 2008), the diphthamide biosynthetic
75 pathway is universally conserved in all archaea and eukaryotes. Indeed, while not strictly
76 essential, loss of diphthamide biosynthesis has been shown to result in growth defects in yeast
77 (Kimata and Kohno 1994; Ortiz, et al. 2006) and some archaea (Blaby, et al. 2010), and is either
78 lethal or causes severe developmental abnormalities in mammals (Liu, et al. 2006; Webb, et al.
79 2008; Yu, et al. 2014).

80 In the current study, we explore the evolution and function of EF-2 and of diphthamide
81 biosynthesis genes using genomic data from novel major archaeal lineages that were recently
82 discovered using metagenomics and single-cell genomics approaches (Adam, et al. 2017; Hug, et
83 al. 2016; Spang, et al. 2017). In particular, we report the presence of EF-2 paralogs in many
84 archaeal genomes belonging to the Asgard archaea, *Korarchaeota* and Bathyarchaeota (Evans, et
85 al. 2015; He, et al. 2016; Lazar, et al. 2016; Meng, et al. 2014; Spang, et al. 2015; Zaremba-
86 Niedzwiedzka, et al. 2017) and the unexpected absence of diphthamide biosynthesis genes in

several archaea and in parabaslid eukaryotes. Our findings reveal a complex evolutionary history of EF-2 and diphthamide biosynthesis genes, and point to novel mechanisms of translational regulation in several archaeal lineages. Finally, our results are compatible with scenarios in which eukaryotes evolved from an Asgard-related ancestor (Spang, et al. 2015; Zaremba-Niedzwiedzka, et al. 2017) and suggest the presence of a diphthamidated EF-2 in this lineage.

MATERIALS AND METHODS

Sampling and sequencing of ABR Loki- and Thorarchaeota.

Sampling, DNA extraction, library preparation and sequencing was produced as described in (Zaremba-Niedzwiedzka, et al. 2017). We chose the four deepest samples, at 125 and 175 cm below sea-floor (MM3/PM3 and MM4/PM4 respectively), as they showed highest lokiarchaeal diversity in a maximum likelihood phylogeny of 5 to 15 ribosomal proteins (RP15) encoded on the same contig (Zaremba-Niedzwiedzka, et al. 2017). Adapters and low quality bases were trimmed using Trimmomatic version 0.32 with the following parameters: PE -phred33 ILLUMINACLIP:NexteraPE-PE.fa:2:30:10:1:true LEADING:3 TRAILING:6 SLIDINGWINDOW:4:15 MINLEN:36 (Bolger, et al. 2014).

Assembly of ABR Loki- and Thorarchaeota.

Samples from the same depth were assembled together using IDBA-UD (Peng, et al. 2012) (version 1.1.1-384, --maxk 124 -r <MERGED_READS>) producing four different assemblies (S1:MM1/PM1, S2:MM2/PM2, S3:MM3/PM3, S4:MM4/PM4). Assemblies S3 and S4 were particularly interesting as they showed the highest lokiarchaeal diversity. However, some lokiarchaeal members showed highly fragmented contigs, probably due to the low abundances of

these organisms. In an attempt to produce longer contigs we co-assembled those reads coming from Asgard archaea members in the samples MM3, PM3, MM4 and PM4. Asgard archaea reads were identified using Clark (version 1.2.3, -m 0) (Ounit, et al. 2015) and Bowtie2 (version 2.2.4, default parameters) (Langmead and Salzberg 2012) against a customized Asgard archaea database. Classified reads were extracted and co-assembled using SPAdes (version v.3.9.0, --careful) (Bankevich, et al. 2012).

In brief, the Asgard database was composed of Asgard genomes publicly available on February 2017. Clark does not perform well when organisms present in the samples of interest are not highly similar to the ones present in the provided database. To increase the classification sensitivity, we included in our database low-quality Asgard MAGs (with highly fragmented contigs) generated from assemblies S3 and S4, using CONCOCT (Alneberg, et al. 2014). Coverage profiles required by CONCOCT were estimated using kallisto (version 0.43.0, quant --plaintext) (Bray, et al. 2016). All available samples from the same location (MM1, PM1, MM2, PM2, MM3, PM3, MM4, PM4) were used and mapped independently against the assemblies S3 and S4. For each assembly, MAGs were reconstructed using two different minimum contig length thresholds (2000 and 3000 bp). We used the number of containing clusters of ribosomal proteins (ribocontigs) as a proxy to estimate the microbial diversity present in the community. The maximum number of clusters (-c option in CONCOCT) was estimated by calculating approximately 2.5 times the estimated number of species in the sample (Johannes Alneberg, personal communication), resulting into 900 and 600 for S3 and S4, respectively. Potential Asgard archaea bins were identified based on the presence of ribocontigs classified as Asgard archaea and were included in the database.

133 ***Binning of ABR Loki- and Thorarchaeota.***

134 Several binning tools with different settings were run independently: CONCOCT_2000: version
135 0.4.0, --read_length 200 and minimum contig length of 2000. CONCOCT_3000: version 0.4.0, --
136 read_length 200 and minimum contig length of 3000. In both cases, coverage files were created
137 mapping all 8 samples against the co-assembly using kallisto. MaxBin2: version 2.2.1, -
138 min_contig_length 2000 -markerset 40 –plotmarker (Wu, et al. 2016). The 8 samples were
139 mapped against the co-assembly using Bowtie2. Coverage was estimated using the getabund.pl
140 script provided. MyCC_4mer: 4mer -t 2000 (Lin and Liao 2016). MyCC_56mer: 56mer -t 2000.
141 Both coverage profiles were obtained as the authors described in their manual.
142 The results of those 5 binning methods were combined into a consensus: contigs were assigned
143 to bins if they had been classified as the same organism by at least 3 out of 5 methods. The
144 resulting bins were manually inspected and cleaned further using mmgenome (Albertsen, et al.
145 2013). Completeness and redundancy was computed using CheckM (Parks, et al. 2015).
146
147 ***Sampling and sequencing of OWC Thorarchaeota.***
148 Eight soil samples were collected from the Old Woman Creek (OWC) National Estaurine
149 Research Reserve and DNA was extracted as described previously (Narrowe, et al. 2017).
150 Library preparation and five lanes of Illumina HiSeq 2x125 bp sequencing followed standard
151 operating procedures at the US DOE Joint Genome Institute (GOLD study ID Gs0114821).
152 Sample M3-C4-D3 had replicate extraction, library preparation, and two lanes of sequencing
153 performed, and reads were combined before downstream analysis. For 3 additional samples (M3-
154 C4-D4, O3-C3-D3, O3-C3-D4) one lane of sequencing was performed. For the other 4 samples
155 (M3-C5-D1, M3-C5-D2, M3-C5-D3, M3-C5-D4) DNA was sheared to 300bp with a Covaris

S220, metagenomic sequencing libraries were prepared using the Nugen Ovation Ultralow Prep kit, and all four samples were multiplexed on one lane of Illumina HiSeq 2x125 sequencing at the University of Colorado Denver Anschutz Medical Campus Genomics and Microarray Core.

Assembly and binning of OWC Thorarchaeota.

For initial assembly of the 5 full-lane sequencing runs, adapter removal, read filtering and trimming were completed using BBDuk (sourceforge.net/projects/bbmap) ktrim=r, minlen=40, minlenfraction=0.6,
mink=11 tbo, tpe k=23, hdist=1 hdist2=1 ftm=5 , maq=8, maxns=1, minlen=40, minlenfraction=0.6, k=27, hdist=1, trimq=12, qtrim=rl. Filtered reads were assembled using megahit (Li, et al. 2015) version 1.0.6 with --k-list 23,43,63,83,103,123.
The individual metagenome from the O3-C4-D3 sample was binned using Emergent Self-Organizing Maps (ESOM)(Dick, et al. 2009) of tetranucleotide frequency (5kb contigs, 3kb windows). BLAST hits of predicted proteins identified a Thorarchaeota population bin. All scaffolds containing a window in this bin were used as a mapping reference and reads from the 9 OWC libraries were mapped to this bin using bbsplit with default parameters (sourceforge.net/projects/bbmap). The mapped reads were reassembled using SPAdes version 3.9.0 with --careful -k 21,33,55,77,95,105,115,125 (Bankevich, et al. 2012). Finally, the reads which were input to the reassembly were mapped to the assembled scaffolds using Bowtie 2 (Langmead and Salzberg 2012) to generate a coverage profile which was used to manual identify bins using Anvi'o (Eren, et al. 2015). Proteins were predicted using prodigal (Hyatt, et al. 2010) and searched against UniRef90 release 11-2016 (Suzek, et al. 2015), with the taxonomy of best blast hits used to validate contigs as probable Thorarchaeota. Contigs having no top hit to the

publicly available Thorarchaeota genomes were manually examined and removed if they could be assigned to another genome bin in the larger metagenomic assembly. Genome completeness and contamination was estimated using CheckM (Parks, et al. 2015).

Identification of diphthamide biosynthesis genes and EF-2 homologs in eukaryotes and archaea.

The EGGNOG members dataset (available at http://eggnogdb.embl.de/#/app/downloads) was surveyed for sequences corresponding to the following clusters of orthologous groups (COG): EF-2, COG0480; DPH1/DPH2, COG1736; DPH3, COG5216; DPH4, COG0484; DPH5, COG1798; DPH6, COG2102; and DPH7, ENOG4111MMJ. For genomes not represented in EGGNOG, we manually inspected publicly available genomes as indicated by 'orthology assignment source' (Supplementary File S1). Similarly, an in-house arCOG dataset, modeled after the publicly available arCOGs from Makarova et al. (Makarova, et al. 2015),was queried for the corresponding COG distribution in relevant archaeal genomes. Finally, aEF-2 and aEF-2p genes in Thorarchaeota OWC Bin 2,3 and 5 were identified using HMMER: version 3.1b2, hmmsearch --cut-tc (Eddy 2011) against PFAM models PF00679 (EF-G_C) and PF03764 (EFG_IV). Conserved synteny surrounding the Thorarchaoeta aEF-2p gene was used to further search for partial aEF-2p genes. In addition, all contigs with matching HMM hits to *dph2* and *dph5* in the full OWC assembly were manually examined for potential Thorarchaeal *dph* genes; none were identified.

Phylogenetic analyses

Elongation factor 2: EF-2 and EF-2 paralogs of Asgard archaea, Koarchaeota and Bathyarchaeota were aligned with a representative set of archaeal, bacterial EF-2 and eukaryotic EF-2, EFL1 and snRNP homologs using mafft-linsi (Katoh and Standley 2013). Subsequently, poorly aligned ends were removed manually before the alignments were trimmed with trimAl 5% (Capella-Gutierrez, et al. 2009), yielding 871 aligned amino acid positions. Maximum likelihood analyses were performed using IQ-tree using the mixture model LG+C60+R4+F, which was selected among the C-series models based on its Bayesian information criterion score by the built-in model test implemented in IQ-tree. Branch supports were assessed using ultrafast bootstrap approximation as well as with single branch test (-alrt option).

Diphthamide biosynthesis proteins Dph1/Dph2 (IPR016435; arCOG04112) and Dph5 (IPR004551; arCOG04161): Both Dph1 and Dph2 as well as Dph5 homologs of a representative set of eukaryotes were aligned with archaeal Dph1/2 and Dph5 homologs, respectively. Several DPANN genomes contain two genes encoding the CTD and NTD of Dph1/2 (Fig. 1, Supplementary File S1) such that Dph1/2 homologs of these organisms had to be concatenated prior to aligning Dph1/2 sequences. Alignments were performed using mafft-linsi and trimmed with BMGE (Criscuolo and Gribaldo 2010) using the blossum 30 matrix and setting the entropy to 0.55. This resulted in final alignments of 170 (Dph1/2) and 221 (Dph5). Maximum likelihood analyses were performed using IQ-tree (Nguyen, et al. 2015) with the mixture models resulting in the lowest BIC: LG+C50+R+F (Dph1/2) and LG+C60+R+F (Dph5), respectively. Branch supports were assessed using ultrafast bootstrap approximation (Hoang, et al. 2018) as well as with the single branch test (-alrt flag).

Concatenated ribosomal proteins: A phylogenetic tree of co-localized ribosomal proteins was performed using the rp15 pipeline as described previously (Zaremba-Niedzwiedzka, et al. 2017).

In brief, archaeal ribosomal proteins encoded in the r-protein gene cluster (requiring a minimum of 11 ribosomal proteins) were aligned with mafft-linsi, trimmed with trimAl using the -gappyout option, concatenated and subjected to maximum likelihood analyses using IQ-tree with the LG+C60+R4+F model chosen based on best BIC score as described above. Branch supports were assessed using ultrafast bootstrap approximation as well as with the single branch test (-alrt option) in IQTREE.

Structural modeling of EF-2 homologs.

Structural models of a/eEF-2 genes and paralogs were generated using the i-Tasser standalone package version 5.1 (Yang, et al. 2015), and visualized and analyzed using UCSF Chimera version 1.11.12 (Pettersen, et al. 2004). The best structural hits to the PDB for each sequence's top-scoring model were identified using COFACTOR (Roy, et al. 2012). The *Drosophila melanogaster eEF-2* structure in complex with the ribosome (PDB:4V6W) was used as a structural reference to which all models were superimposed (aligned) using Chimera's MatchMaker.

Loop motif logos of EF-2 homologs

e/aEF-2 and paralog sequences which were used to generate the EF-2 tree were clustered at 90% amino acid identity using CD-HIT: version 4.6, -c 0.9 -n 5 (Fu, et al. 2012) and the sequence alignment was filtered to retain only cluster centroids. The conserved loop sequences were extracted from the filtered EF-2 alignment using Jalview version 2.10.1 (Waterhouse, et al. 2009), verified by cross-referencing to the structural models, and sequence logos generated on

246 cluster centroids only using WebLogo: version 2.8.2 (weblogo.berkeley.edu) (Crooks, et al.
247 2004).
248
249 ***Accession Numbers***
250 Taxonomy and accession numbers for all genes analyzed in this study are listed in
251 Supplementary File S1.
252
253 **RESULTS**
254 **Most Asgard archaea, *Korarchaeota* and *Geoarchaea* as well as parabasalids lack**
255 **diphthamide synthesis genes**
256 It was previously assumed that EF-2 of all eukaryotes and Archaea was uniquely characterized
257 by the presence of diphthamide. To examine if this assumption is still valid when taking into
258 account recently sequenced genomes, we surveyed 337 archaeal and 168 eukaryotic genomes
259 (File S1) for each of the three known archaeal (de Crécy-Lagard, et al. 2012) and seven
260 eukaryotic (Su, et al. 2012a; Su, et al. 2012b; Uthman, et al. 2013) *dph* genes. While most
261 archaeal genomes encode clear *dph* homologues, we failed to detect the diphthamide
262 biosynthesis genes in a large diversity of metagenome-assembled genomes (MAGs) of
263 uncultured archaea, including newly assembled MAGs analyzed for this study (Fig. 1,
264 Supplementary Fig. S1, Supplementary File S1). In particular, our analyses showed that, as
265 reported for *K. cryptophilum* (de Crécy-Lagard, et al. 2012; Elkins, et al. 2008), all *Korarchaeota*
266 and *Geoarchaea* as well as nearly all members of the Asgard archaea lack the conserved archaeal
267 diphthamide biosynthesis genes *dph1/2*, *dph5* and *dph6*. As an exception, Asgard archaea related
268 to the Heimdallarchaeote LC3 clade were found to encode the complete archaeal diphthamide

biosynthetic pathway (Fig. 1). Genes coding for Dph5 and Dph6 could not be detected in two Bathyarchaeota draft genomes (RBG_13_46_16b and SG8_32_3). However, it is unclear whether these two genomes are in the process of losing *dph* biosynthesis genes or whether the absence of *dph5* and *dph6* genes is due to the incompleteness of these draft genomes. We also surveyed 168 eukaryotic genomes and high-quality transcriptomes, including those lineages that have undergone drastic genome reduction, such as microsporidians (Corradi, et al. 2010), diplomonads (Morrison, et al. 2007), and degenerate nuclei (i.e., nucleomorphs) of secondary plastids in cryptophytes (Lane, et al. 2007) (Supplementary File S1) for *dph* gene homologs. We detected *dph* homologues in all eukaryotic genomes and transcriptomes except for parabasalid protists, including animal pathogens such as *Trichomonas vaginalis, Tritrichomonas foetus* and *Dientamoeba fragilis* (Supplementary File S1). Unless these archaea and parabasalids possess alternative, yet undiscovered diphthamide biosynthesis pathways, these findings suggest that their cognate EF-2 lacks the modified diphthamide residue. As a peculiarity, while the Dph1/2 protein is encoded by a single fusion gene in seemingly all archaea, we found that in several members of the DPANN archaea(Castelle, et al. 2015; Rinke, et al. 2013) this protein is encoded by two genes that separately code for the N- and C-terminal domains. To our knowledge, this is the first systematic report of the widespread absence of diphthamide biosynthesis in diverse eukaryotes and archaea.

Various archaeal genomes that lack diphthamide biosynthesis genes encode an EF-2 paralog

To shed light into the implications of the potential lack of diphthamide in members of the Asgard archaea and *Korarchaeota*, we performed detailed analyses of eukaryotic and archaeal EF-2

292 homologs (Fig. 1). First, we found that the draft genomes of most Asgard archaea, some
293 *Korarchaeota* (Kor 1 and 3), and a few Bathyarchaeota encode two distantly related EF-2
294 paralogs. In contrast, the genomes of *K. cyptophilum* and two novel marine *Korarchaeota* (Kor 2
295 and 4) and Heimdallarchaeota LC2 and LC3 as well as *Geoarchaea* do not encode an EF-2
296 paralog. Given that the Heimdallarchaeota LC2 genome was estimated to be only 70-79 %
297 complete (Zaremba-Niedzwiedzka, et al. 2017), and based on phylogenetic analyses (see below),
298 we consider it possible that this genome might encode an as-yet unassembled aEF-2 paralog. The
299 presence of paralogous aEF-2 in most Asgard archaea and some *Korarchaeota* genomes
300 corresponds with the absence of diphthamide synthesis genes (Fig. 1 and 2). Yet, even though
301 the genomes of *K. cryptophilum*, Kor 2, Kor 4, and *Geoarchaea* as well as of Heimdallarchaeote
302 LC2 lack *dph* genes, they do not encode an EF-2 paralog. In all other archaeal genomes,
303 including that of Heimdallarchaeote LC3, the absence of an EF-2 paralog correlates with the
304 presence of *dph* genes.
305
306 **Archaea with two EF-2 family proteins encode only one *bona fide* EF-2**
307 We next addressed whether residues and structural motifs shown to be necessary for canonical
308 translocation were conserved in the various EF-2 and EF-2 paralogs. Domain IV of EF-2,
309 representing the anticodon mimicry domain, is critical for facilitating concerted translocation of
310 tRNA and mRNA (Ortiz, et al. 2006; Rodnina, et al. 1997). This domain includes three loops
311 that extend out from the body of EF-2 and interact with the decoding center of the ribosome. The
312 first of these three loops (HxDxxHRG) (canonical residue positions are numbered according to
313 sequence associated with *D. melanogaster* structural model PDB 4V6W (Anger, et al. 2013))
314 contains the site of the diphthamide modified histidine, H701, and is highly conserved across

315 archaea and eukaryotes (Ortiz, et al. 2006; Zhang, et al. 2008). High conservation is also seen in
316 a second adjacent loop (SPHKHN) in the a/eEF-2 domain IV (S581-N586), which contains a
317 lysine residue (K584) that interacts directly with the tRNA at the decoding center, and is itself
318 positioned by a stacking interaction between P582 and H585 (Murray, et al. 2016). The third
319 loop appears to stabilize the diphthamide loop, partially via a salt-bridge formed between a
320 nearby glutamate residue (E660) and R702 in the diphthamide loop (Anger, et al. 2013). Both of
321 these residues are highly conserved among archaea and eukaryotes.

322 Our analyses reveal that the sequence motifs in these loops are also strictly conserved
323 among the EF-2 family proteins of the Heimdallarchaeote LC3 lineage, *Geoarchaea*, as well as
324 in those *Korarchaeota* and Bathyarchaeota that lack an EF-2 paralog (Fig. 3, Supplementary Fig.
325 S2a). Notably, this conservation is seen irrespective of the presence or absence of *dph* genes in
326 those genomes. However, most *bona fide* EF-2 of parabasalids (which lack *dph* genes), possesses
327 a glycine to asparagine mutation at residue 703 (Fig. 3, Supplementary Fig. S2b, Supplementary
328 Fig. S3a), which may compensate for the lack of the diphthamide residue by contributing an
329 amide group (Fig. 3, Supplementary Fig. S3b).

330 In contrast, in those Asgard archaea and *Korarchaeota* (Kor 1/3 clade) that encode two
331 EF-2 family proteins, even within the bona fide EF-2 copy, these domain IV motifs show
332 reduced conservation. In the diphthamide loop, R702 is universally replaced by a threonine
333 residue. In 21 of 22 aEF-2 proteins, there is a correlated mutation of E660 to either arginine or
334 lysine (Supplementary Fig. S4). Structural homology modeling suggested that these correlated
335 mutations likely prevent unfavorable electrostatic interactions between domain IV loops, and
336 maintain stabilization of the diphthamide loop (Supplementary Fig. S4). While G703 is
337 conserved in most EF-2s of archaea, all Lokiarchaeota (except Lokiarchaeota CR_4), encode

338 either a serine or a glutamine at this site (Fig. 3, Supplementary Fig. S2a). Furthermore, analysis
339 of the second loop (S581-N586) revealed additional crucial mutations in the EF-2 of these
340 archaea; notably, K584 is not conserved (Fig. 3, Supplementary Fig. S2a). Despite these
341 modifications which correlate with the presence of an EF-2 paralog in these archaea, there is still
342 evidence for strong selection pressure maintaining many of the key conserved residues in these
343 domain IV motifs, including H701, the target site of diphthamide modification (Fig. 3,
344 Supplementary Fig. S2a).

345 In contrast, our analyses of the multiple sequence alignment and structural models
346 suggest that the paralogous EF-2 (aEF-2p) proteins encoded by these archaea lack conservation
347 in the stabilizing second loop (SPHKHN) as well as the first diphthamide loop (HxDxxHRG),
348 including H701 (Fig. 3). Based on predicted fold conservation in domains I and II, and the
349 overall conservation of the five sequence motifs (G1-G5) characterizing GTPase superfamily
350 proteins (Atkinson 2015), aEF-2p likely maintains GTPase activity (Supplementary Fig. S5).
351 However, given the apparent lack of conservation in key domain IV loops, it is unlikely that
352 aEF-2p proteins can serve as functional translocases in protein translation.

353

354 **EF-2 homologs of archaea experienced complex evolutionary history**

355 To resolve the evolutionary history of EF-2, we performed phylogenetic analyses of archaeal EF-
356 2 (aEF-2) and aEF-2p, bacterial EF-G and eukaryotic EF-2 family proteins, i.e. EF-2, Ria1 (or
357 Elongation factor like, EFL1) and Snu114 (or U5 small nuclear ribonucleoprotein, snRNP/ U5-
358 116kD) (Fig. 2) (Atkinson 2015). First, our analyses revealed that sequences from all non-LC3
359 Asgard archaea and the Kor-1 and -3 marine *Korarchaeota* formed two distinct clades, one of
360 which contains canonical aEF-2 proteins (as defined by conservation of the domain IV loop

known to interact with the ribosomal decoding center during translocation) while the other cluster comprises aEF-2p (Fig. 2). However, the phylogenetic placement of these protein clades relative to each other and within the phylogenetic backbone is not fully resolved due to lack of statistical support. This might be caused by modified (accelerated) evolutionary rates that appear to characterize the evolution of aEF-2 and aEF-2p in lineages that encode a paralog, as indicated by increased relative branch lengths for both the aEF-2 and aEF-2p clades (Fig. 2, Supplementary Files S2 and S3).

Secondly, bathyarchaeal EF-2 homologs were also found to form two separate clades. One of these clades is placed within the TACK superphylum, and includes both canonical bathyarchaeal EF-2s as well as potential paralogs (i.e., RBG_13_46_16b and SG8-32-3). In contrast, the second clade is only comprised of two sequences (i.e., RBG_13_46_16b and AD8-1), and is placed as a sister group of all TACK, Asgard and eukaryotic EF-2 homologs (Fig. 2). In spite of this deep placement in the phylogenetic analyses, the second clade is comprised of the canonical EF-2 homologs of Bathyarchaeota genomes RBG_13_46_16b and AD8-1, based on analysis of key domain IV residues. Currently, only the most complete of the latter two draft genomes, RBG_13_46_16b, contains an aEF-2 paralog. Therefore, the current data is insufficient to resolve the puzzling pattern of EF-2 evolution in the Bathyarchaeota phylum.

Finally, in our analysis, eEF-2, Ria1 and Snu114 were found to form a highly supported monophyletic group that emerged as a sister group to the aEF-2 proteins encoded by the genomes comprising the Heimdallarchaeote LC3 clade (LC3 and B3).

Close inspection of the EF-2 sequence alignment revealed that eukaryotic and LC3 EF-2 homologs share common indels to the exclusion of all other archaeal EF-2 family protein sequences (Supplementary Fig. S6, Supplementary Fig. S7). Notably, these highly conserved

384 indels were found to be encoded by the genomic bins of two distantly related members of the
385 Heimdallarchaeota LC3 lineage, which were independently assembled and binned from
386 geographically distinct metagenomes (Spang, et al. 2015; Zaremba-Niedzwiedzka, et al. 2017).
387 This refutes recently raised claims stating that these indels in Heimdallarchaeote LC3 may be the
388 results of contamination from eukaryotes (Da Cunha, et al. 2017) while supporting the sister-
389 relationship of eukaryotes and Asgard archaea (Eme, et al. 2017; Spang, et al. ; Spang, et al.
390 2015; Zaremba-Niedzwiedzka, et al. 2017). In addition, despite the low sequence identity of
391 39%, the high-confidence modeled structure of Heimdallarchaeote LC3 EF-2 was highly similar
392 to *Drosophila melanogaster* eEF-2 (RMSD (root-mean-square deviation) 1.3Å across all 796
393 residues to *D. melanogaster* structural model PDB 4V6W (Anger, et al. 2013); Supplementary
394 File S1). By comparison, the Heimdallarchaeaote AB-125 model aligns less confidently to the
395 Drosophila EF-2 structure (RMSD 16.4Å). The observed phylogenetic topology and the presence
396 of the full complement of *dph* biosynthesis genes in LC3 genomes (Figs. 1 and 2), support an
397 evolutionary scenario in which Heimdallarchaeote LC3 and eukaryotes share a common ancestry
398 with EF-2 being vertically inherited from this archaeal ancestor.
399

400 **DISCUSSION**

401 The use of metagenomic approaches has led to an expansion of genomic data from a large
402 diversity of previously unknown archaeal and bacterial lineages and has changed our perception
403 of the tree of life, microbial metabolic diversity and evolution, as well as the origin of eukaryotes
404 (Brown, et al. 2015; Castelle, et al. 2015; Hug, et al. 2016; Parks, et al. 2017; Spang, et al. 2015;
405 Zaremba-Niedzwiedzka, et al. 2017). Since most of what is known about archaeal informational
406 processing machineries is based on a few model organisms, we aimed to use the expansion of

407 genomic data to investigate key elements of the translational machinery - EF-2 and
408 diphthamidylation - across the tree of life.

409 Our analyses of archaeal EF-2 family proteins and the distribution of diphthamide
410 biosynthesis genes have revealed unusual features of the core translation machinery in several
411 archaeal lineages. These findings negate two long-held assumptions regarding the archaeal and
412 eukaryotic translation machineries, with both functional and evolutionary implications. First, we
413 show that diphthamide modification is not universally conserved across Archaea and eukaryotes.
414 Second, we demonstrate that, much like Bacteria and eukaryotes (Atkinson 2015), the archaeal
415 EF-2 protein family has undergone several gene duplication events, presumably coupled to
416 functional differentiation of EF-2 paralogs, throughout archaeal evolution.

417 The evolution of archaeal diphthamide biosynthesis and EF-2 is especially intriguing in
418 the context of eukaryogenesis. Recent findings based on comparative genomics indicate that
419 eukaryotes evolved from a symbiosis between an alphaproteobacterium with an archaeal host
420 that shares a most recent common ancestor with extant members of the Asgard archaea, possibly
421 a Heimdallarchaeota-related lineage (Spang, et al. 2015; Zaremba-Niedzwiedzka, et al. 2017).
422 Our study adds additional data to support this scenario by revealing close sequence and predicted
423 structural similarity of canonical EF-2 proteins of the Heimdallarchaeote LC3 lineage and
424 eukaryotic EF-2 proteins, including shared indels. Furthermore, phylogenetic analyses of EF-2
425 family proteins reveals that EF-2 of the Heimdallarchaeote LC3 lineage forms a monophyletic
426 group with EF-2 family proteins of eukaryotes, and therefore suggests that the archaeal ancestor
427 of eukaryotes was equipped with an EF-2 protein similar to the homologs found in this lineage.
428 The subsequent evolution of the eukaryotic EF-2 family appears to have included at least two
429 ancient duplication events leading to Ria1 and Snu114. Importantly, the presence of

characteristic eukaryotic indels in EF-2 of all members of the Heimdallarchaeote LC3 lineage further strengthens this hypothesis and underlines that concerns raised about the quality of these genomic bins (Da Cunha, et al. 2017) are unjustified (Spang, et al.).

In addition, the LC3 clade also represents the sole group within the Asgard archaea that is characterized by the presence of the full complement of archaeal diphthamide biosynthesis pathway genes. However, while phylogenetic analyses of Dph1/2 show weak support for a sister-relationship between Heimdallarchaeota and eukaryotes, eukaryotic Dph5 appears to be most closely related to homologs of Woesearchaeaota (Supplementary Fig. S8, Supplementary File S3), an archaeal lineage belonging to the proposed DPANN superphylum (Castelle, et al. 2015; Rinke, et al. 2013; Williams, et al. 2017), comprising various additional lineages with putative symbiotic and/or parasitic members (reviewed in Spang et al. (Spang, et al. 2017)). Notably, a previous study has also revealed an affiliation of some eukaryotic tRNA synthetases with DPANN archaea (Furukawa, et al. 2017). Given that several DPANN lineages infect or closely associate with other archaeal lineages, they may exchange genes with their hosts frequently, as was shown for *Nanoarchaeum equitans* and its crenarchaeal host *Ignicoccus hospitalis* (Podar, et al. 2008). Following a similar reasoning, the archaeal ancestor of eukaryotes (i.e. a relative of the Asgard archaea) may have acquired genes (e.g. *dph5*) from an ancestral DPANN/Woesearchaeota symbiont. However, prospective analyses and generation of genomic data from additional members of the Asgard and DPANN archaea are necessary to test this hypothesis and to clarify the evolutionary history of the origin of diphthamide biosynthesis genes in eukaryotes.

Furthermore, our findings have practical implications for studies that involve phylogenetic and metagenomic analyses. Previously, EF-2 has been widely used as a

phylogenetic marker, in both single-gene (Baldauf, et al. 1996; Elkins, et al. 2008; Hashimoto and Hasegawa 1996; Iwabe, et al. 1989), and multiple-gene alignments of universal single copy genes [(Guy, et al. 2014; Raymann, et al. 2015; Williams, et al. 2012), and others] to assess the relationships between Archaea, Bacteria and eukaryotes. However, the presence of paralogs of EF-2 in various Archaea and eukaryotes suggest that EF-2 should be excluded from such datasets. In addition, EF-2, Dph1/2, and Dph5 are part of single-copy marker gene sets regularly used to estimate genome completeness and purity of archaeal metagenomic bins (Parks, et al. 2015; Wu and Scott 2012). The presence of duplicated aEF-2 gene families, the absence of *dph* genes in most Asgard archaea, *Geoarchaea* and *Korarchaeota*, and the presence of two split genes for Dph1/2 in DPANN makes these genes unsuited as marker genes, and should hence be excluded from marker gene sets used to assess genome completeness.

The observed absence of *dph* biosynthesis genes in various Archaea as well as parabasalids is surprising given that diphthamide was previously thought to be a conserved feature across Archaea and eukaryotes (Schaffrath, et al. 2014), and critical for ensuring translational fidelity (Ortiz, et al. 2006). While we currently cannot rule out the possibility that *dph*-lacking archaea and parabasalids perform the multi-step process of diphthamidylation using a set of yet-unknown enzymes, future proteomics studies will be needed to conclusively rule out the presence of diphthamide in these taxa. Yet, it is more likely that these groups have evolved a different mechanism or mechanisms to fulfill the proposed roles of diphthamide in translation.

Many of the *dph*-lacking archaeal genomes encode two paralogs of the aEF-2 gene. Despite the apparent absence of diphthamide, our sequence and structural modeling analyses imply that these dipthamide-deficient aEF-2 proteins are likely under strong selective pressure to maintain translocase function. In contrast, analyses of the aEF-2p suggest that, while this paralog

476 is a member of the translational GTPase superfamily, aEF-2p is unlikely to function in the same
477 manner as canonical aEF-2. In fact, the complete lack of sequence conservation in aEF-2p key
478 domain IV loop residues indicates that these paralogs are not likely to act as translocases (Fig. 3,
479 Supplementary Fig. S2a) (Ortiz, et al. 2006; Rodnina, et al. 1997) and instead perform alternative
480 roles. For instance, it seems possible that aEF-2p may compensate for the absence of
481 diphthamide in at least some *dph*-lacking lineages. However, other functions for aEF-2p such as
482 error-correcting back-translocation or ribosome recycling also seem possible, given the observed
483 sub- and neo-functionalizations seen in eukaryotic and bacterial EF-2/EF-G paralogs (Qin, et al.
484 2006; Tsuboi, et al. 2009). Alternatively, given proposed regulation of translation via ADP-
485 ribosylation of diphthamide (Schaffrath, et al. 2014) and a role of diphthamide in responding to
486 oxidative stress (Argüelles, et al. 2014; Argüelles, et al. 2013), aEF-2p could perform another,
487 yet unknown role in translation regulation.

488 Currently, the consequences for the absence of *dph* biosynthesis genes in parabasalids
489 and in several Archaea remain unclear. Future studies could gain insight into such questions by
490 studying translation in the genetically tractable parabasalid *Trichomonas vaginalis,* whose cell
491 biology and metabolism has been extensively studied. In addition, acquisition of additional
492 sequencing data or enrichment cultures from members of the Asgard superphylum,
493 *Korarchaeota*, and other novel archaeal lineages will lead to a better understanding of the
494 evolution and function of EF-2 family proteins, and the absence of *dph* biosynthesis genes.
495

496 **ACKNOWLEDGEMENTS**

497 We thank Jordan Angle, Kay Stefanik, Rebecca Daly, and Kelly Wrighton for assistance with
498 sampling of OWC sediments, and Felix Homa for computational support. Sequencing of OWC

499 metagenomes was conducted in part by the U.S. Department of Energy Joint Genome Institute, a
500 DOE Office of Science User Facility that is supported by the Office of Science of the U.S.
501 Department of Energy under Contract No. DE-AC02-05CH11231. Sequencing of Aarhus bay
502 metagenomes was performed by the National Genomics Infrastructure sequencing platforms at
503 the Science for Life Laboratory at Uppsala University, a national infrastructure supported by the
504 Swedish Research Council (VR-RFI) and the Knut and Alice Wallenberg Foundation. We thank
505 the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) at
506 Uppsala University and the Swedish National Infrastructure for Computing (SNIC) at the PDC
507 Center for High-Performance Computing for providing computational resources. This work was
508 supported by grants of the European Research Council (ERC Starting grant 310039-
509 PUZZLE_CELL), the Swedish Foundation for Strategic Research (SSF-FFL5) and the Swedish
510 Research Council (VR grant 2015-04959) to T.J.G.E.. C.W.S. is supported by a European
511 Molecular Biology Organisation long-term fellowship (ALTF-997-2015) and the Natural
512 Sciences and Engineering Research Council of Canada postdoctoral research fellowship (PDF-
513 487174-2016).

514

515 **REFERENCES**

516 Adam PS, Borrel G, Brochier-Armanet C, Gribaldo S 2017. The growing tree of Archaea: new
517 perspectives on their diversity, evolution and ecology. ISME J 11: 2407-2425. doi:
518 10.1038/ismej.2017.122
519 Albertsen M, et al. 2013. Genome sequences of rare, uncultured bacteria obtained by differential
520 coverage binning of multiple metagenomes. Nat Biotechnol 31: 533-538. doi: 10.1038/nbt.2579
521 Alneberg J, et al. 2014. Binning metagenomic contigs by coverage and composition. Nat
522 Methods 11: 1144-1146. doi: 10.1038/nmeth.3103

523 Anger AM, et al. 2013. Structures of the human and Drosophila 80S ribosome. Nature 497: 80-
524 85. doi: 10.1038/nature12104
525 Argüelles S, Camandola S, Cutler RG, Ayala A, Mattson MP 2014. Elongation factor 2
526 diphthamide is critical for translation of two IRES-dependent protein targets, XIAP and FGF2,
527 under oxidative stress conditions. Free Radic Biol Med 67: 131-138. doi:
528 10.1016/j.freeradbiomed.2013.10.015
529 Argüelles S, et al. 2013. Molecular control of the amount, subcellular location, and activity state
530 of translation elongation factor 2 in neurons experiencing stress. Free Radic Biol Med 61: 61-71.
531 doi: 10.1016/j.freeradbiomed.2013.03.016
532 Atkinson GC 2015. The evolutionary and functional diversity of classical and lesser-known
533 cytoplasmic and organellar translational GTPases across the tree of life. BMC Genomics 16: 78.
534 doi: 10.1186/s12864-015-1289-7
535 Baldauf SL, Palmer JD, Doolittle WF 1996. The root of the universal tree and the origin of
536 eukaryotes based on elongation factor phylogeny. Proc Natl Acad Sci U S A 93: 7749-7754.
537 Bankevich A, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to
538 single-cell sequencing. J Comput Biol 19: 455-477. doi: 10.1089/cmb.2012.0021
539 Becam AM, Nasr F, Racki WJ, Zagulski M, Herbert CJ 2001. Ria1p (Ynl163c), a protein similar
540 to elongation factors 2, is involved in the biogenesis of the 60S subunit of the ribosome in
541 Saccharomyces cerevisiae. Mol Genet Genomics 266: 454-462. doi: 10.1007/s004380100548
542 Blaby IK, et al. 2010. Towards a systems approach in the genetic analysis of archaea:
543 Accelerating mutant construction and phenotypic analysis in Haloferax volcanii. Archaea 2010:
544 426239. doi: 10.1155/2010/426239
545 Bolger AM, Lohse M, Usadel B 2014. Trimmomatic: a flexible trimmer for Illumina sequence
546 data. Bioinformatics 30: 2114-2120. doi: 10.1093/bioinformatics/btu170
547 Botet J, Rodriguez-Mateos M, Ballesta JP, Revuelta JL, Remacha M 2008. A chemical genomic
548 screen in Saccharomyces cerevisiae reveals a role for diphthamidation of translation elongation
549 factor 2 in inhibition of protein synthesis by sordarin. Antimicrob Agents Chemother 52: 1623-
550 1629. doi: 10.1128/AAC.01603-07
551 Bray NL, Pimentel H, Melsted P, Pachter L 2016. Near-optimal probabilistic RNA-seq
552 quantification. Nat Biotechnol 34: 525-527. doi: 10.1038/nbt.3519

553 Brown CT, et al. 2015. Unusual biology across a group comprising more than 15% of domain
554 Bacteria. Nature 523: 208-211. doi: 10.1038/nature14486
555 Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T 2009. trimAl: a tool for automated
556 alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25: 1972-1973. doi:
557 10.1093/bioinformatics/btp348
558 Castelle CJ, et al. 2015. Genomic expansion of domain archaea highlights roles for organisms
559 from new phyla in anaerobic carbon cycling. Curr Biol 25: 690-701. doi:
560 10.1016/j.cub.2015.01.014
561 Corradi N, Pombert JF, Farinelli L, Didier ES, Keeling PJ 2010. The complete sequence of the
562 smallest known nuclear genome from the microsporidian Encephalitozoon intestinalis. Nat
563 Commun 1: 77. doi: 10.1038/ncomms1082
564 Criscuolo A, Gribaldo S 2010. BMGE (Block Mapping and Gathering with Entropy): a new
565 software for selection of phylogenetic informative regions from multiple sequence alignments.
566 BMC Evol Biol 10: 210. doi: 10.1186/1471-2148-10-210
567 Crooks GE, Hon G, Chandonia JM, Brenner SE 2004. WebLogo: a sequence logo generator.
568 Genome Res 14: 1188-1190. doi: 10.1101/gr.849004
569 Da Cunha V, Gaia M, Gadelle D, Nasir A, Forterre P 2017. Lokiarchaea are close relatives of
570 Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. PLoS Genet 13:
571 e1006810. doi: 10.1371/journal.pgen.1006810
572 de Crécy-Lagard V, Forouhar F, Brochier-Armanet C, Tong L, Hunt JF 2012. Comparative
573 genomic analysis of the DUF71/COG2102 family predicts roles in diphthamide biosynthesis and
574 B12 salvage. Biol Direct 7: 32. doi: 10.1186/1745-6150-7-32
575 Dick GJ, et al. 2009. Community-wide analysis of microbial genome sequence signatures.
576 Genome Biol 10: R85. doi: 10.1186/gb-2009-10-8-r85
577 Donhofer A, et al. 2012. Structural basis for TetM-mediated tetracycline resistance. Proc Natl
578 Acad Sci U S A 109: 16900-16905. doi: 10.1073/pnas.1208037109
579 Eddy SR 2011. Accelerated Profile HMM Searches. PLoS Comput Biol 7: e1002195. doi:
580 10.1371/journal.pcbi.1002195
581 Elkins JG, et al. 2008. A korarchaeal genome reveals insights into the evolution of the Archaea.
582 Proceedings of the National Academy of Sciences 105: 8102-8107. doi:
583 10.1073/pnas.0801980105

584 Eme L, Spang A, Lombard J, Stairs CW, Ettema TJG 2017. Archaea and the origin of
585 eukaryotes. Nature Reviews Microbiology 15: 711. doi: 10.1038/nrmicro.2017.133
586 Eren AM, et al. 2015. Anvi'o: an advanced analysis and visualization platform for 'omics data.
587 PeerJ 3: e1319. doi: 10.7717/peerj.1319
588 Evans PN, et al. 2015. Methane metabolism in the archaeal phylum Bathyarchaeota revealed by
589 genome-centric metagenomics. Science 350: 434-438. doi: 10.1126/science.aac7745
590 Fabrizio P, Laggerbauer B, Lauber J, Lane WS, Luhrmann R 1997. An evolutionarily conserved
591 U5 snRNP-specific protein is a GTP-binding factor closely related to the ribosomal translocase
592 EF-2. EMBO J 16: 4092-4106. doi: 10.1093/emboj/16.13.4092
593 Freistroffer DV, Pavlov MY, MacDougall J, Buckingham RH, Ehrenberg M 1997. Release
594 factor RF3 in E.coli accelerates the dissociation of release factors RF1 and RF2 from the
595 ribosome in a GTP-dependent manner. EMBO J 16: 4126-4133. doi: 10.1093/emboj/16.13.4126
596 Fu L, Niu B, Zhu Z, Wu S, Li W 2012. CD-HIT: accelerated for clustering the next-generation
597 sequencing data. Bioinformatics 28: 3150-3152. doi: 10.1093/bioinformatics/bts565
598 Furukawa R, Nakagawa M, Kuroyanagi T, Yokobori SI, Yamagishi A 2017. Quest for Ancestors
599 of Eukaryal Cells Based on Phylogenetic Analyses of Aminoacyl-tRNA Synthetases. J Mol Evol
600 84: 51-66. doi: 10.1007/s00239-016-9768-2
601 Guy L, Saw JH, Ettema TJ 2014. The archaeal legacy of eukaryotes: a phylogenomic
602 perspective. Cold Spring Harb Perspect Biol 6: a016022. doi: 10.1101/cshperspect.a016022
603 Hashimoto T, Hasegawa M 1996. Origin and early evolution of eukaryotes inferred from the
604 amino acid sequences of translation elongation factors 1alpha/Tu and 2/G. Adv Biophys 32: 73-
605 120.
606 He Y, et al. 2016. Genomic and enzymatic evidence for acetogenesis among multiple lineages of
607 the archaeal phylum Bathyarchaeota widespread in marine sediments. Nature Microbiology 1:
608 16035. doi: 10.1038/nmicrobiol.2016.35
609 Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS 2018. UFBoot2: Improving the
610 Ultrafast Bootstrap Approximation. Mol Biol Evol 35: 518-522. doi: 10.1093/molbev/msx281
611 Hug LA, et al. 2016. A new view of the tree of life. Nature Microbiology 1. doi: Artn 16048
612 10.1038/Nmicrobiol.2016.48
613 Hyatt D, et al. 2010. Prodigal: prokaryotic gene recognition and translation initiation site
614 identification. BMC Bioinformatics 11: 119. doi: 10.1186/1471-2105-11-119

615 Iglewski BH, Liu PV, Kabat D 1977. Mechanism of action of Pseudomonas aeruginosa exotoxin
616 Aiadenosine diphosphate-ribosylation of mammalian elongation factor 2 in vitro and in vivo.
617 Infect Immun 15: 138-144.
618 Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T 1989. Evolutionary relationship of
619 archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes.
620 Proc Natl Acad Sci U S A 86: 9355-9359.
621 Jorgensen R, et al. 2008. Cholix toxin, a novel ADP-ribosylating factor from Vibrio cholerae. J
622 Biol Chem 283: 10671-10678. doi: 10.1074/jbc.M710008200
623 Katoh K, Standley DM 2013. MAFFT multiple sequence alignment software version 7:
624 improvements in performance and usability. Mol Biol Evol 30: 772-780. doi:
625 10.1093/molbev/mst010
626 Kimata Y, Kohno K 1994. Elongation factor 2 mutants deficient in diphthamide formation show
627 temperature-sensitive cell growth. J Biol Chem 269: 13497-13501.
628 Lane CE, et al. 2007. Nucleomorph genome of Hemiselmis andersenii reveals complete intron
629 loss and compaction as a driver of protein structure and function. Proc Natl Acad Sci U S A 104:
630 19908-19913. doi: 10.1073/pnas.0707419104
631 Langmead B, Salzberg SL 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:
632 357-359. doi: 10.1038/nmeth.1923
633 Lazar CS, et al. 2016. Genomic evidence for distinct carbon substrate preferences and ecological
634 niches of Bathyarchaeota in estuarine sediments. Environ Microbiol 18: 1200-1211. doi:
635 10.1111/1462-2920.13142
636 Li D, Liu CM, Luo R, Sadakane K, Lam TW 2015. MEGAHIT: an ultra-fast single-node
637 solution for large and complex metagenomics assembly via succinct de Bruijn graph.
638 Bioinformatics 31: 1674-1676. doi: 10.1093/bioinformatics/btv033
639 Li L, et al. 2014. Ribosomal elongation factor 4 promotes cell death associated with lethal stress.
640 MBio 5: e01708. doi: 10.1128/mBio.01708-14
641 Lin HH, Liao YC 2016. Accurate binning of metagenomic contigs via automated clustering
642 sequences using information of genomic signatures and marker genes. Sci Rep 6: 24175. doi:
643 10.1038/srep24175
644 Liu S, et al. 2006. Dph3, a small protein required for diphthamide biosynthesis, is essential in
645 mouse development. Mol Cell Biol 26: 3835-3841. doi: 10.1128/MCB.26.10.3835-3841.2006

646 Makarova KS, Wolf YI, Koonin EV 2015. Archaeal Clusters of Orthologous Genes (arCOGs):
647 An Update and Application for Analysis of Shared Features between Thermococcales,
648 Methanococcales, and Methanobacteriales. Life (Basel) 5: 818-840. doi: 10.3390/life5010818
649 Meng J, et al. 2014. Genetic and functional properties of uncultivated MCG archaea assessed by
650 metagenome and gene expression analyses. ISME J 8: 650-659. doi: 10.1038/ismej.2013.174
651 Morrison HG, et al. 2007. Genomic minimalism in the early diverging intestinal parasite Giardia
652 lamblia. Science 317: 1921-1926. doi: 10.1126/science.1143837
653 Murray J, et al. 2016. Structural characterization of ribosome recruitment and translocation by
654 type IV IRES. Elife 5. doi: 10.7554/eLife.13567
655 Narrowe AB, et al. 2017. High-resolution sequencing reveals unexplored archaeal diversity in
656 freshwater wetland soils. Environ Microbiol 19: 2192-2209. doi: 10.1111/1462-2920.13703
657 Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ 2015. IQ-TREE: a fast and effective
658 stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32: 268-
659 274. doi: 10.1093/molbev/msu300
660 Ortiz PA, Ulloque R, Kihara GK, Zheng H, Kinzy TG 2006. Translation elongation factor 2
661 anticodon mimicry domain mutants affect fidelity and diphtheria toxin resistance. J Biol Chem
662 281: 32639-32648. doi: 10.1074/jbc.M607076200
663 Ounit R, Wanamaker S, Close TJ, Lonardi S 2015. CLARK: fast and accurate classification of
664 metagenomic and genomic sequences using discriminative k-mers. BMC Genomics 16: 236. doi:
665 10.1186/s12864-015-1419-2
666 Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW 2015. CheckM: assessing the
667 quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome
668 Res 25: 1043-1055. doi: 10.1101/gr.186072.114
669 Parks DH, et al. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially
670 expands the tree of life. Nature Microbiology 2: 1533-1542. doi: 10.1038/s41564-017-0012-7
671 Peng Y, Leung HC, Yiu SM, Chin FY 2012. IDBA-UD: a de novo assembler for single-cell and
672 metagenomic sequencing data with highly uneven depth. Bioinformatics 28: 1420-1428. doi:
673 10.1093/bioinformatics/bts174
674 Pettersen EF, et al. 2004. UCSF Chimera--a visualization system for exploratory research and
675 analysis. J Comput Chem 25: 1605-1612. doi: 10.1002/jcc.20084

676 Podar M, et al. 2008. A genomic analysis of the archaeal system Ignicoccus hospitalis-
677 Nanoarchaeum equitans. Genome Biol 9: R158. doi: 10.1186/gb-2008-9-11-r158
678 Qin Y, et al. 2006. The highly conserved LepA is a ribosomal elongation factor that back-
679 translocates the ribosome. Cell 127: 721-733. doi: 10.1016/j.cell.2006.09.037
680 Raymann K, Brochier-Armanet C, Gribaldo S 2015. The two-domain tree of life is linked to a
681 new root for the Archaea. Proc Natl Acad Sci U S A 112: 6670-6675. doi:
682 10.1073/pnas.1420858112
683 Rinke C, et al. 2013. Insights into the phylogeny and coding potential of microbial dark matter.
684 Nature 499: 431-437. doi: 10.1038/nature12352
685 Rodnina MV, Savelsbergh A, Katunin VI, Wintermeyer W 1997. Hydrolysis of GTP by
686 elongation factor G drives tRNA movement on the ribosome. Nature 385: 37-41. doi:
687 10.1038/385037a0
688 Roy A, Yang J, Zhang Y 2012. COFACTOR: an accurate comparative algorithm for structure-
689 based protein function annotation. Nucleic Acids Res 40: W471-477. doi: 10.1093/nar/gks372
690 Schaffrath R, Abdel-Fattah W, Klassen R, Stark MJ 2014. The diphthamide modification
691 pathway from Saccharomyces cerevisiae--revisited. Mol Microbiol 94: 1213-1226. doi:
692 10.1111/mmi.12845
693 Spahn CM, et al. 2004. Domain movements of elongation factor eEF2 and the eukaryotic 80S
694 ribosome facilitate tRNA translocation. EMBO J 23: 1008-1019. doi: 10.1038/sj.emboj.7600102
695 Spang A, Caceres EF, Ettema TJG 2017. Genomic exploration of the diversity, ecology, and
696 evolution of the archaeal domain of life. Science 357. doi: 10.1126/science.aaf3883
697 Spang A, et al. Asgard archaea are the closest prokaryotic relatives of eukaryotes. PLoS Genetics
698 forthcoming.
699 Spang A, et al. 2015. Complex archaea that bridge the gap between prokaryotes and eukaryotes.
700 Nature 521: 173-+. doi: 10.1038/nature14447
701 Su X, et al. 2012a. YBR246W is required for the third step of diphthamide biosynthesis. J Am
702 Chem Soc 134: 773-776. doi: 10.1021/ja208870a
703 Su X, et al. 2012b. Chemogenomic approach identified yeast YLR143W as diphthamide
704 synthetase. Proc Natl Acad Sci U S A 109: 19983-19987. doi: 10.1073/pnas.1214346109

Suematsu T, et al. 2010. A bacterial elongation factor G homologue exclusively functions in ribosome recycling in the spirochaete Borrelia burgdorferi. Mol Microbiol 75: 1445-1454. doi: 10.1111/j.1365-2958.2010.07067.x

Suzek BE, et al. 2015. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31: 926-932. doi: 10.1093/bioinformatics/btu739

Tsuboi M, et al. 2009. EF-G2mt is an exclusive recycling factor in mammalian mitochondrial protein synthesis. Mol Cell 35: 502-510. doi: 10.1016/j.molcel.2009.06.028

Uthman S, et al. 2013. The amidation step of diphthamide biosynthesis in yeast requires DPH6, a gene identified through mining the DPH1-DPH5 interaction network. PLoS Genet 9: e1003334. doi: 10.1371/journal.pgen.1003334

Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189-1191. doi: 10.1093/bioinformatics/btp033

Webb TR, et al. 2008. Diphthamide modification of eEF2 requires a J-domain protein and is essential for normal development. J Cell Sci 121: 3140-3145. doi: 10.1242/jcs.035550

Williams TA, Foster PG, Nye TM, Cox CJ, Embley TM 2012. A congruent phylogenomic signal places eukaryotes within the Archaea. Proc Biol Sci 279: 4870-4879. doi: 10.1098/rspb.2012.1795

Williams TA, et al. 2017. Integrative modeling of gene and genome evolution roots the archaeal tree of life. Proc Natl Acad Sci U S A 114: E4602-E4611. doi: 10.1073/pnas.1618463114

Wu M, Scott AJ 2012. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics 28: 1033-1034. doi: 10.1093/bioinformatics/bts079

Wu YW, Simmons BA, Singer SW 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32: 605-607. doi: 10.1093/bioinformatics/btv638

Yang J, et al. 2015. The I-TASSER Suite: protein structure and function prediction. Nat Methods 12: 7-8. doi: 10.1038/nmeth.3213

Yu YR, You LR, Yan YT, Chen CM 2014. Role of OVCA1/DPH1 in craniofacial abnormalities of Miller-Dieker syndrome. Hum Mol Genet 23: 5579-5596. doi: 10.1093/hmg/ddu273

Zaremba-Niedzwiedzka K, et al. 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541: 353-358. doi: 10.1038/nature21031

736 Zhang Y, Liu S, Lajoie G, Merrill AR 2008. The role of the diphthamide-containing loop within
737 eukaryotic elongation factor 2 in ADP-ribosylation by Pseudomonas aeruginosa exotoxin A.
738 Biochem J 413: 163-174. doi: 10.1042/BJ20071083
739

740 **FIGURE LEGENDS**

741 **Figure 1 - Diphthamide biosynthesis genes are conserved across most eukaryotic and**
742 **archaeal lineages.** Eukaryotic and archaeal orthologues of diphthamide biosynthesis (DPH)
743 genes were retrieved from the publicly available EGGNOG and an in-house archaeal orthologues
744 (arCOG) datasets. Complete list of genomes surveyed can be found in Supplementary File S1
745 including reduced genomes from nucleomorphs (not shown on figure). Total number of genomes
746 surveyed are shown next to each group. Since Dph4 is a member of the large DNAJ-containing
747 protein family, we could not unequivocally identify this protein based on orthology alone and is
748 therefore excluded from the figure. $^{+}$No arCOG available for DPH3. *All eukaryotic genomes
749 are complete except five deeply-sequenced transcriptomes from Parabasalia; dark and light grey
750 circles indicate whether homologues were detected in more or less than 50% of the genomes
751 surveyed respectively; yellow circles indicate the absence of a detectable homologue; pink
752 circles indicate lack of conservation of the diphthamide modification motif; half-circles indicate
753 the presence of multiple copies of EF-2 with and without the conserved diphthamide
754 modification motif. 1 - Homologue detected in the original assembly (ABR_125(Zaremba-
755 Niedzwiedzka, et al. 2017)) but not in the reassembly (ABR16 genome); a closer inspection of
756 the contig revealed that it is chimeric and will thus be removed from the final bin; 2 -
757 Homologue detected in only one Lokiarchaeota assembly (AB_15); 3 - Several DPANN
758 genomes contain two proteins that encode the CTD and NTD of Dph1/2, respectively.
759

Figure 2 - The evolution of archaeal EF-2 family proteins. Phylogenetic tree of EF-2 family proteins based on maximum likelihood analyses of 871 aligned positions using IQ-tree. EF-2 of Bathyarchaeota grouping in an unexpected position or representing potential aEF-2p are shaded in orange. aEF-2 of Kor- and Asgard archaea are shaded in purple, while their aEF-2p are shaded in green. Highlighted amino acids show the conservation of key residues and black/white circles reveal the presence/absence of *dph* biosynthesis genes in the respective organisms/MAGs. Branch support values are based on ultrafast bootstrap approximation as well as single branch tests, respectively and are represented by differentially colored circles as detailed in the figure panel. Whenever branch support values were below 80 for any of the two methods, values have been removed and branches cannot be considered significantly supported.
Scale bar indicates the number of substitutions per site. Abr.; snRNP: U5 small nuclear ribonucleoprotein EFL1: elongation factor-like GTPase ; n.c.: not conserved; p.c.: partially conserved; n.d.: not determined.

Figure 3 - Predicted structure of Asgard archaea EF-2 and EF-2 paralogs
Structural modeling of representative EF-2 genes and paralogs compared to eukaryotic EF-2 structure shows conservation of overall EF-2 structure regardless of diphthamide synthesis capacity (top). The overall fold of two loops located at the tip of domain IV is conserved, but otherwise highly conserved sequence motifs in these loops are not conserved in DPH^- Asgard archaea and Korarchaea or in EF-2 paralogs (middle). Bottom panels show a close-up of the key residues from the motifs, highlighting that these residues are those positioned at the tip of the domain IV loops crucial for interaction with the decoding site in canonical EF-2 structures. Histidine residue that is the site of dph modification is starred.

Figure 1

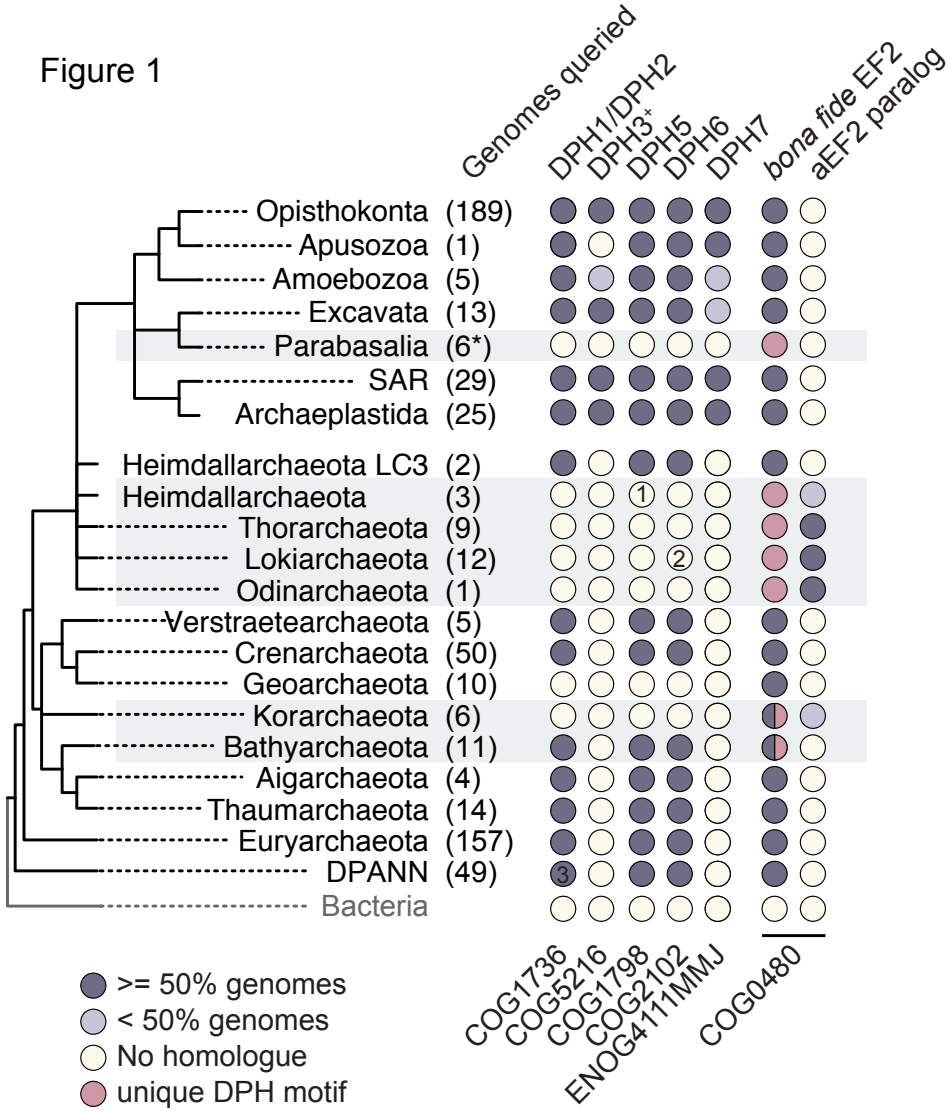

Genomes queried
DPH1/DPH2
DPH3+
DPH5
DPH6
DPH7
bona fide EF2
aEF2 paralog
Opisthokonta (189)
Apusozoa (1)
Amoebozoa (5)
Excavata (13)
Parabasalia (6*)
SAR (29)
Archaeplastida (25)
Heimdallarchaeota LC3 (2)
Heimdallarchaeota (3)
Thorarchaeota (9)
Lokiarchaeota (12)
Odinarchaeota (1)
Verstraetearchaeota (5)
Crenarchaeota (50)
Geoarchaeota (10)
Korarchaeota (6)
Bathyarchaeota (11)
Aigarchaeota (4)
Thaumarchaeota (14)
Euryarchaeota (157)
DPANN (49)
Bacteria
1
2
3
COG1736
COG5216
COG1798
COG2102
ENOG4111MMJ
COG0480
>= 50% genomes
< 50% genomes
No homologue
unique DPH motif

Figure 2

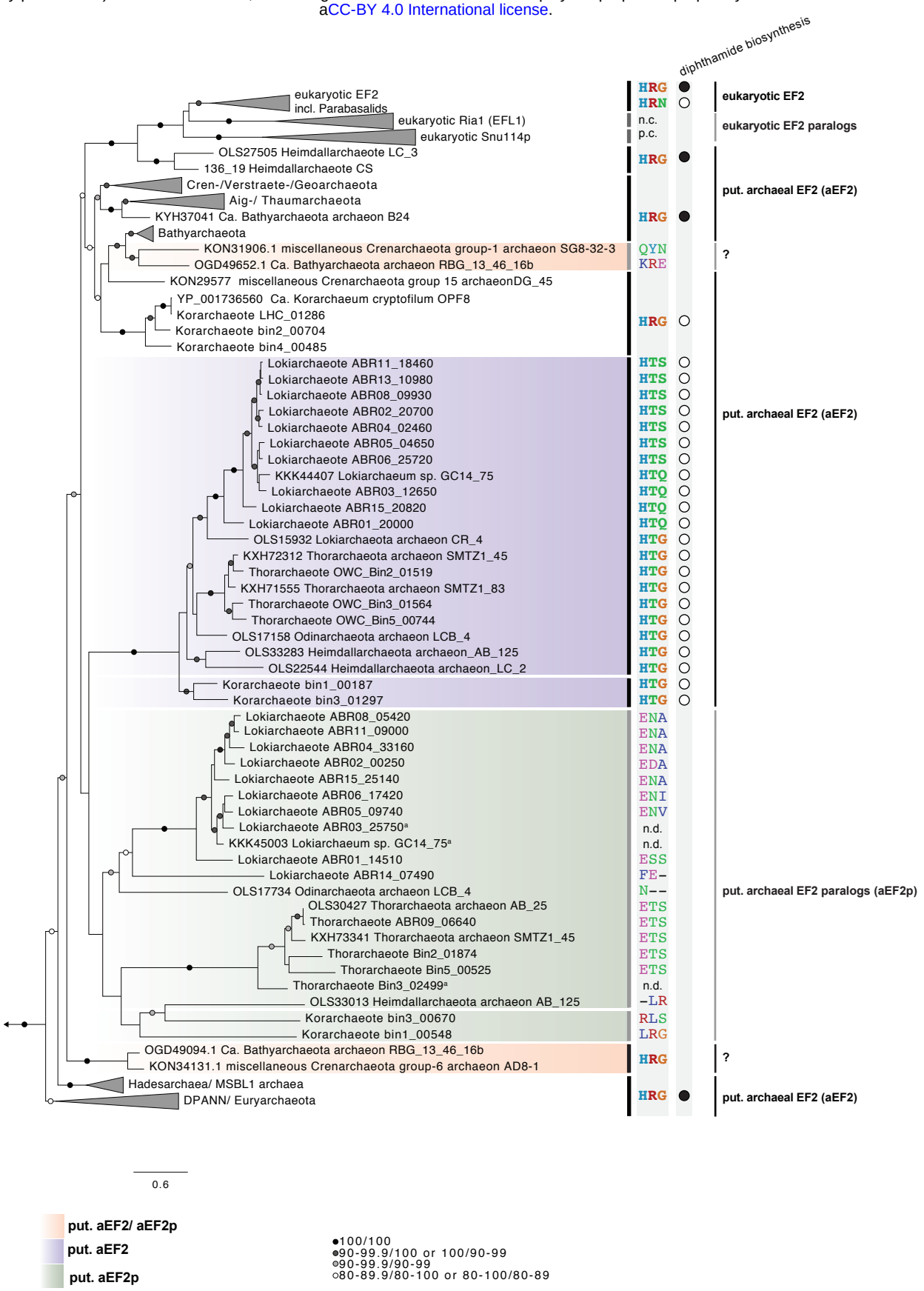


Figure 3

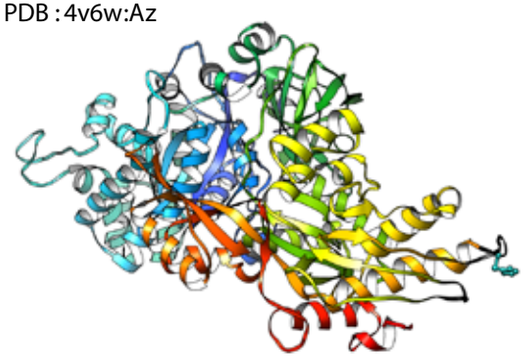
DPH+ Eukaryota canonical eEF-2

DPH+ Archaea canonical aEF-2

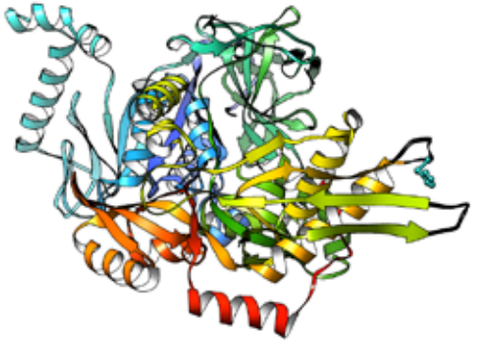
DPH- Asgard and Korarcheaum putative EF-2

DPH- Asgard and Korarchaeum EF-2 paralog

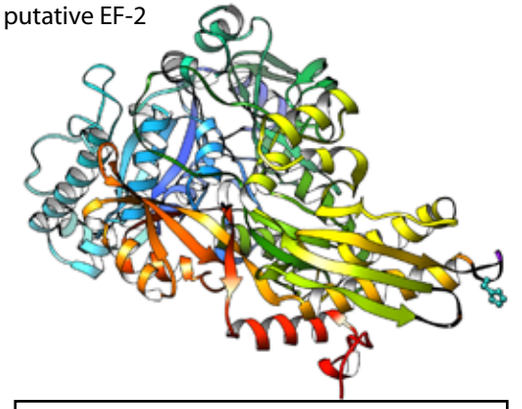
D. melanogaster eEF-2
PDB : 4v6w:Az



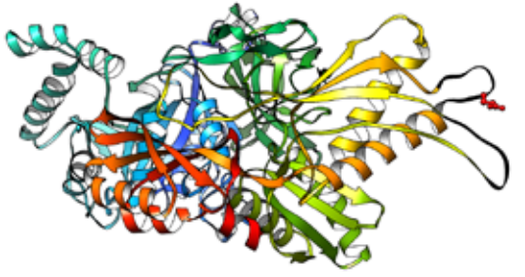
Hadesarchaea DG-33 aEF-2

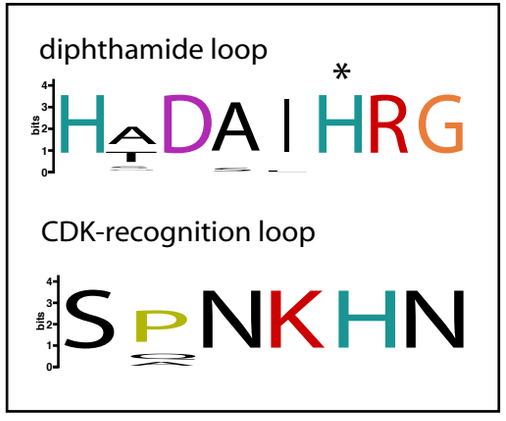


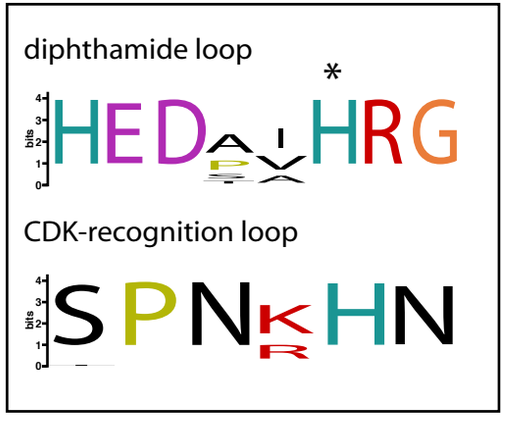
Thorarchaeota OWC-2
putative EF-2

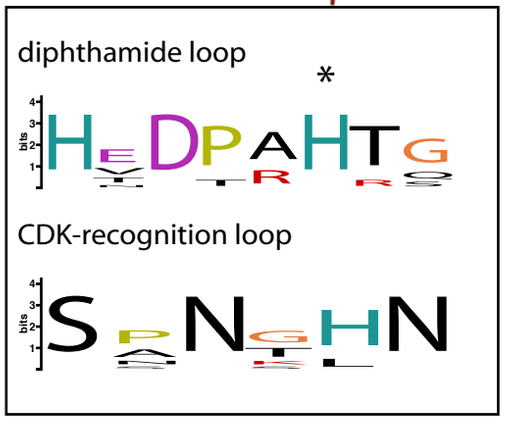


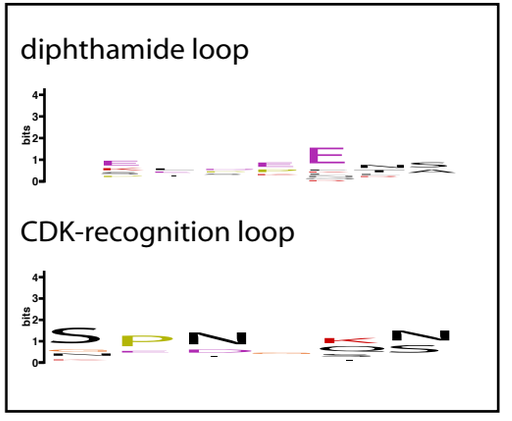
Thorarchaeota OWC-2
EF-2 paralog

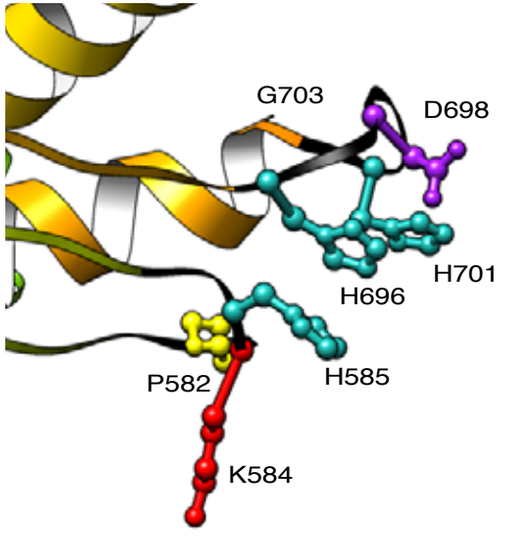


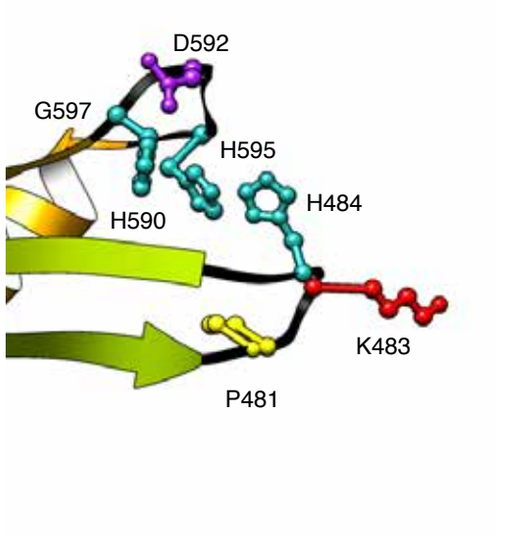


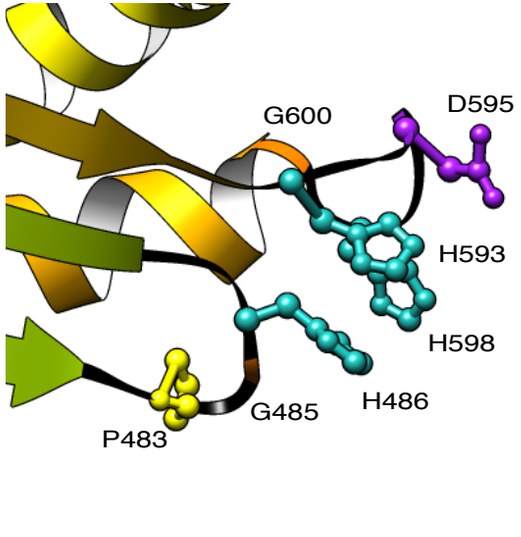


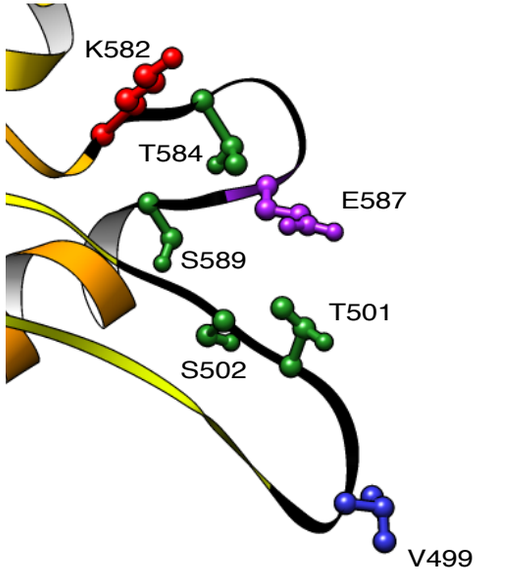












SUPPLEMENTARY INFORMATION

Supplementary Figures.pdf - Contains all supplementary figures referenced in the text.

Supplementary File S1 - Sheet 1: Distribution of diphthamide biosynthesis genes in archaea. Archaeal homologues and their corresponding accession numbers for each Dph gene are shown as retrieved from an in-house archaeal COG (arCOG) dataset. Total counts for each archaeal group of interest are shown. **Sheet 2: Distribution of diphthamide biosynthesis genes in eukaryotes.** Eukaryotic homologues and their corresponding accession numbers for each Dph gene are shown as retrieved from EGGNOG or manual inspection. Total counts for each eukaryotic supergroup are indicated in different colours. When appropriate, nucleomorph- or nucleus-encoded sequences are indicated. **Sheet 3: Structural modeling results including scoring of top structural model predicted by i-Tasser and best structural hit to that model from PDB.**

Supplementary File S2 - Trimmed alignment of EF-2 homologs that was used for phylogenetic analyses. The alignment was generated using mafft-LINSi (Katoh K, Standley DM. Mol Biol Evol. 30:772-80, 2013, doi:10.1093/molbev/mst010) and subjected to trimming with trimAL (5%) (Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. Bioinformatics 25:1972-3, 2009, doi:10.1093/bioinformatics/btp348) after the manual removal of poorly aligned ends. Please refer to methods section for more details.

806 **Supplementary File S3a - Newick file of concatenated ribosomal proteins phylogeny.** Please
807 refer to figure legend of Fig. S1 for more details.
808
809 **Supplementary File S3b - Newick file of EF-2 phylogeny presented in Figure 2.** Please refer
810 to figure legend of Fig. 2 for more details.
811
812 **Supplementary File S3c - Newick file of Dph1/2 phylogeny.** Please refer to figure legend of
813 Fig. S8a for more details.
814
815 **Supplementary File S3d - Newick file of Dph5 phylogeny.** Please refer to figure legend of Fig.
816 S8b for more details.
817
818

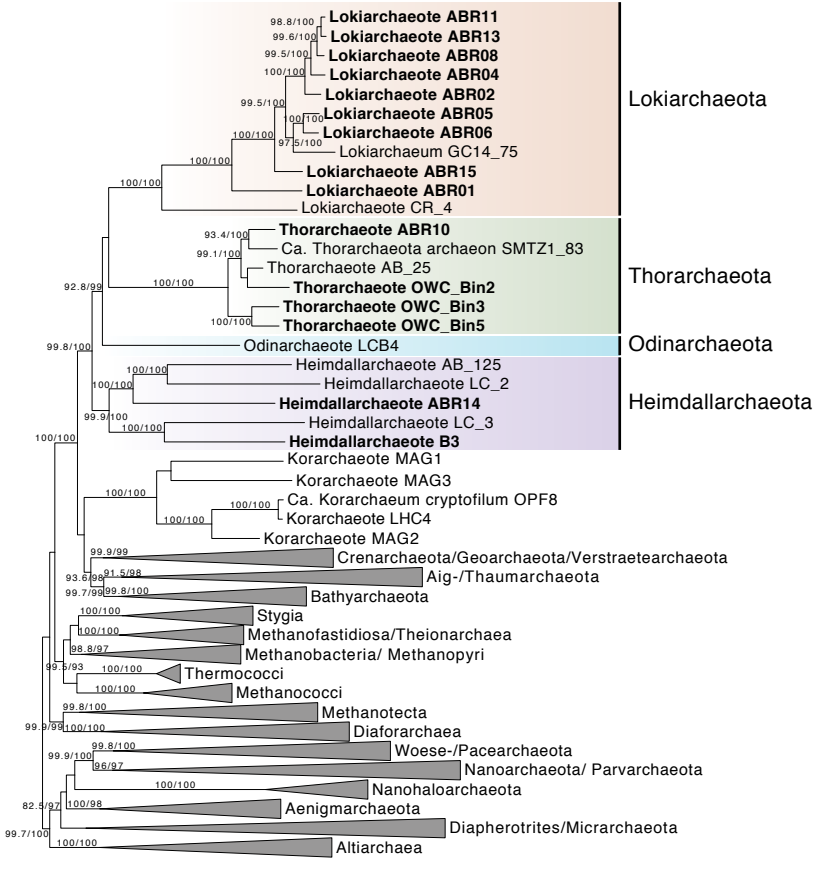
Supplementary Figures for:

Complex evolutionary history of translation Elongation Factor 2 and diphthamide biosynthesis in Archaea and parabasalids

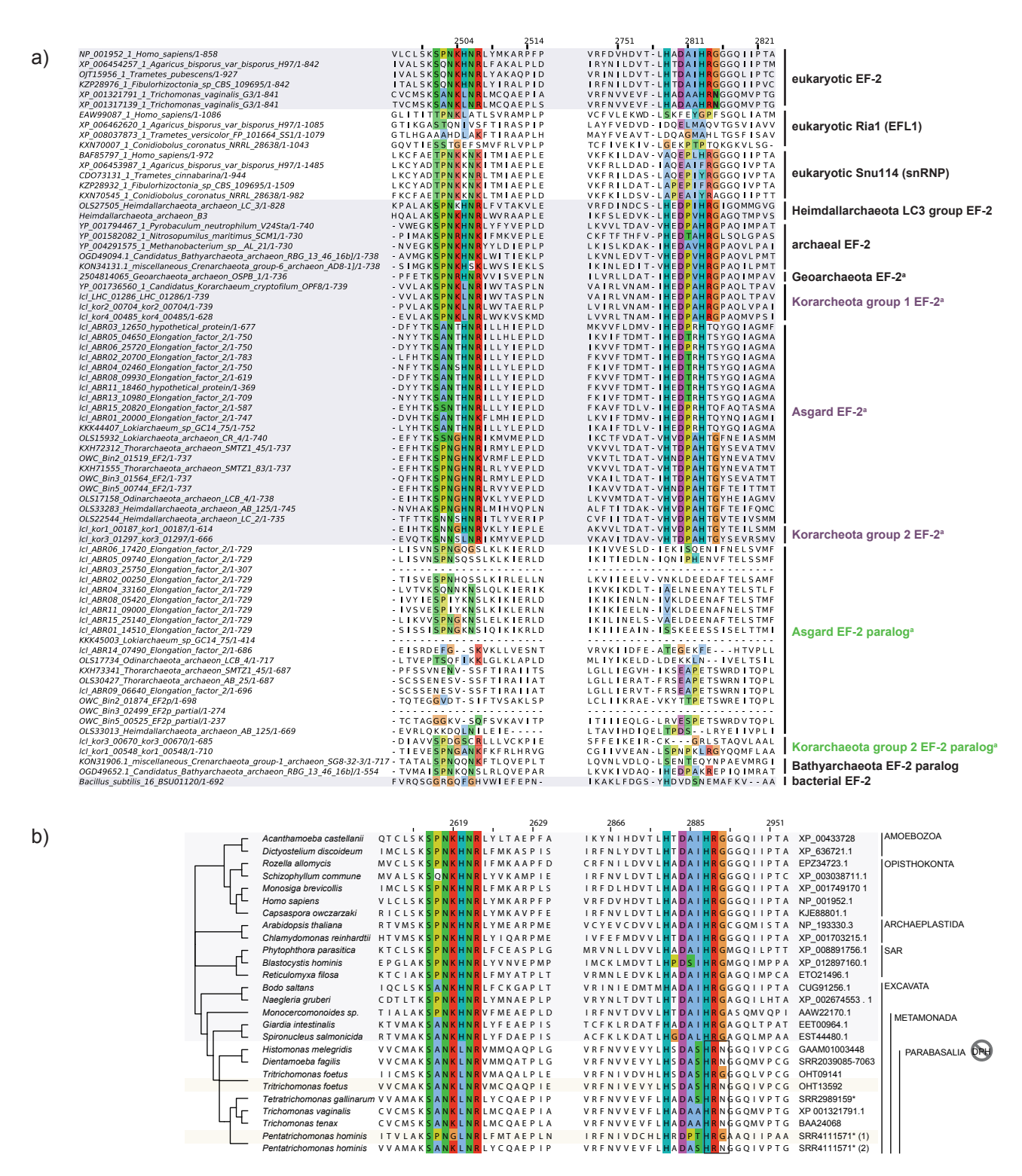
Adrienne B. Narrowe[1,†], Anja Spang[2,3,†], Courtney W. Stairs[2], Eva F. Caceres[2], Brett J. Baker[4], Christopher S. Miller[1,‡,*] and Thijs J. G. Ettema[2,‡]

† These authors contributed equally to this work.
‡These authors contributed equally to this work.
*Author for correspondence: Christopher S. Miller, Department of Integrative Biology, University of Colorado Denver, Denver, CO, USA, 303-315-7665, chris.miller@ucdenver.edu

[1] Department of Integrative Biology, University of Colorado Denver, Denver, Colorado, USA
[2] Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden
[3] NIOZ, Royal Netherlands Institute for Sea Research, Department of Marine Microbiology and Biogeochemistry, and Utrecht University, AB Den Burg, The Netherlands
[4] Department of Marine Science, Marine Science Institute, University of Texas Austin, Port Aransas, TX, USA


Lokiarchaeote ABR11
Lokiarchaeote ABR13
Lokiarchaeote ABR08
Lokiarchaeote ABR04
Lokiarchaeote ABR02
Lokiarchaeote ABR05
Lokiarchaeote ABR06
Lokiarchaeum GC14_75
Lokiarchaeote ABR15
Lokiarchaeote ABR01
Lokiarchaeote CR_4
Lokiarchaeota
Thorarchaeote ABR10
Ca. Thorarchaeota archaeon SMTZ1_83
Thorarchaeote AB_25
Thorarchaeote OWC_Bin2
Thorarchaeote OWC_Bin3
Thorarchaeote OWC_Bin5
Thorarchaeota
Odinarchaeote LCB4
Odinarchaeota
Heimdallarchaeote AB_125
Heimdallarchaeote LC_2
Heimdallarchaeote ABR14
Heimdallarchaeote LC_3
Heimdallarchaeote B3
Heimdallarchaeota
Korarchaeote MAG1
Korarchaeote MAG3
Ca. Korarchaeum cryptofilum OPF8
Korarchaeote LHC4
Korarchaeote MAG2
Crenarchaeota/Geoarchaeota/Verstraetearchaeota
Aig-/Thaumarchaeota
Bathyarchaeota
Stygia
Methanofastidiosa/Theionarchaea
Methanobacteria/ Methanopyri
Thermococci
Methanococci
Methanotecta
Diaforarchaea
Woese-/Pacearchaeota
Nanoarchaeota/ Parvarchaeota
Nanohaloarchaeota
Aenigmarchaeota
Diapherotrites/Micrarchaeota
Altiarchaea
0.2

Supplementary Fig. S1 - Phylogenetic analysis of at least 11 out of 15 concatenated archaeal ribosomal proteins (2416 AA) based on maximum likelihood analyses performed with IQ-tree. The diverse metagenome-assembled genomes (MAGs) belonging to Asgard archaea and included in our analyses are shaded in colors according to phylum. MAGs that were not part of the initial description of the Asgard superphylum (Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Backstrom D, Juzokaite L, Vancaester E, Seitz KW, Anantharaman K, Starnawski P, Kjeldsen KU, Stott MB, Nunoura T, Banfield JF, Schramm A, Baker BJ, Spang A, Ettema TJ. Nature 541:353-358, 2017, doi:10.1038/nature21031) are shown in boldface. Naming of the respective archaeal groups based on a recent suggestion by Adam et al. (ISMEJ 11:2407-2425,2017,doi:10.1038/ismej.2017.122). Branch support values are based on ultrafast bootstrap approximation as well as single branch tests, respectively. Scale bar indicates the number of substitutions per site.


a)
2504
2514
2751
2811
2821
NP_001952_1_Homo_sapiens/1-858
XP_006454257_1_Agaricus_bisporus_var_bisporus_H97/1-842
OJT15956_1_Trametes_pubescens/1-927
KZP28976_1_Fibulorhizoctonia_sp_CBS_109695/1-842
XP_001321791_1_Trichomonas_vaginalis_G3/1-841
XP_001317139_1_Trichomonas_vaginalis_G3/1-841
EAW99087_1_Homo_sapiens/1-1086
XP_006462620_1_Agaricus_bisporus_var_bisporus_H97/1-1085
XP_008037873_1_Trametes_versicolor_FP_101664_SS1/1-1079
KXN70007_1_Conidiobolus_coronatus_NRRL_28638/1-1043
BAF85797_1_Homo_sapiens/1-972
XP_006453987_1_Agaricus_bisporus_var_bisporus_H97/1-1485
CDO73131_1_Trametes_cinnabarina/1-944
KZP28932_1_Fibulorhizoctonia_sp_CBS_109695/1-1509
KXN70545_1_Conidiobolus_coronatus_NRRL_28638/1-982
OLS27505_Heimdallarchaeota_archaeon_LC_3/1-828
Heimdallarchaeota_archaeon_B3
YP_001794467_1_Pyrobaculum_neutrophilum_V24Sta/1-740
YP_001582082_1_Nitrosopumilus_maritimus_SCM1/1-730
YP_004291575_1_Methanobacterium_sp__AL_21/1-730
OGD49094_1_Candidatus_Bathyarchaeota_archaeon_RBG_13_46_16b]/1-738
KON34131_1_miscellaneous_Crenarchaeota_group-6_archaeon_AD8-1]/1-738
2504814065_Geoarchaeota_archaeon_OSPB_1/1-736
YP_001736560_1_Candidatus_Korarchaeum_cryptofilum_OPF8/1-739
lcl_LHC_01286_LHC_01286/1-739
lcl_kor2_00704_kor2_00704/1-739
lcl_kor4_00485_kor4_00485/1-628
lcl_ABR03_12650_hypothetical_protein/1-677
lcl_ABR05_04650_Elongation_factor_2/1-750
lcl_ABR06_25720_Elongation_factor_2/1-750
lcl_ABR02_20700_Elongation_factor_2/1-783
lcl_ABR04_02460_Elongation_factor_2/1-750
lcl_ABR08_09930_Elongation_factor_2/1-619
lcl_ABR11_18460_hypothetical_protein/1-369
lcl_ABR13_10980_Elongation_factor_2/1-709
lcl_ABR15_20820_Elongation_factor_2/1-587
lcl_ABR01_20000_Elongation_factor_2/1-747
KKK44407_Lokiarchaeum_sp_GC14_75/1-752
OLS15932_Lokiarchaeota_archaeon_CR_4/1-740
KXH72312_Thorarchaeota_archaeon_SMTZ1_45/1-737
OWC_Bin2_01519_EF2/1-737
KXH71555_Thorarchaeota_archaeon_SMTZ1_83/1-737
OWC_Bin3_01564_EF2/1-737
OWC_Bin5_00744_EF2/1-737
OLS17158_Odinarchaeota_archaeon_LCB_4/1-738
OLS33283_Heimdallarchaeota_archaeon_AB_125/1-745
OLS22544_Heimdallarchaeota_archaeon_LC_2/1-735
lcl_kor1_00187_kor1_00187/1-614
lcl_kor3_01297_kor3_01297/1-666
lcl_ABR06_17420_Elongation_factor_2/1-729
lcl_ABR05_09740_Elongation_factor_2/1-729
lcl_ABR03_25750_Elongation_factor_2/1-307
lcl_ABR02_00250_Elongation_factor_2/1-729
lcl_ABR04_33160_Elongation_factor_2/1-729
lcl_ABR08_05420_Elongation_factor_2/1-729
lcl_ABR11_09000_Elongation_factor_2/1-729
lcl_ABR15_25140_Elongation_factor_2/1-729
lcl_ABR01_14510_Elongation_factor_2/1-729
KKK45003_Lokiarchaeum_sp_GC14_75/1-414
lcl_ABR14_07490_Elongation_factor_2/1-686
OLS17734_Odinarchaeota_archaeon_LCB_4/1-717
KXH73341_Thorarchaeota_archaeon_SMTZ1_45/1-687
OLS30427_Thorarchaeota_archaeon_AB_25/1-687
lcl_ABR09_06640_Elongation_factor_2/1-696
OWC_Bin2_01874_EF2p/1-698
OWC_Bin3_02499_EF2p_partial/1-274
OWC_Bin5_00525_EF2p_partial/1-237
OLS33013_Heimdallarchaeota_archaeon_AB_125/1-669
lcl_kor3_00670_kor3_00670/1-685
lcl_kor1_00548_kor1_00548/1-710
KON31906.1_miscellaneous_Crenarchaeota_group-1_archaeon_SG8-32-3/1-717
OGD49652.1_Candidatus_Bathyarchaeota_archaeon_RBG_13_46_16b]/1-554
Bacillus_subtilis_16_BSU01120/1-692
eukaryotic EF-2
eukaryotic Ria1 (EFL1)
eukaryotic Snu114 (snRNP)
archaeal EF-2
Heimdallarchaeota LC3 group EF-2
Geoarchaeota EF-2a
Korarcheota group 1 EF-2a
Asgard EF-2a
Korarcheota group 2 EF-2a
Asgard EF-2 paralog a
Korarchaeota group 2 EF-2 paralog a
Bathyarchaeota EF-2 paralog
bacterial EF-2
b)
2619
2629
2866
2885
2951
Acanthamoeba castellanii
Dictyostelium discoideum
Rozella allomycis
Schizophyllum commune
Monosiga brevicollis
Homo sapiens
Capsaspora owczarzaki
Arabidopsis thaliana
Chlamydomonas reinhardtii
Phytophthora parasitica
Blastocystis hominis
Reticulomyxa filosa
Bodo saltans
Naegleria gruberi
Monocercomonoides sp.
Giardia intestinalis
Spironucleus salmonicida
Histomonas melegridis
Dientamoeba fagilis
Tritrichomonas foetus
Tritrichomonas foetus
Tetratrichomonas gallinarum
Trichomonas vaginalis
Trichomonas tenax
Pentatrichomonas hominis
Pentatrichomonas hominis
XP_00433728
XP_636721.1
EPZ34723.1
XP_003038711.1
XP_001749170 1
NP_001952.1
KJE88801.1
NP_193330.3
XP_001703215.1
XP_008891756.1
XP_012897160.1
ETO21496.1
CUG91256.1
XP_002674553 . 1
AAW22170.1
EET00964.1
EST44480.1
GAAM01003448
SRR2039085-7063
OHT09141
OHT13592
SRR2989159*
XP 001321791.1
BAA24068
SRR4111571* (1)
SRR4111571* (2)
AMOEBOZOA
OPISTHOKONTA
ARCHAEPLASTIDA
SAR
EXCAVATA
METAMONADA
PARABASALIA
DHE

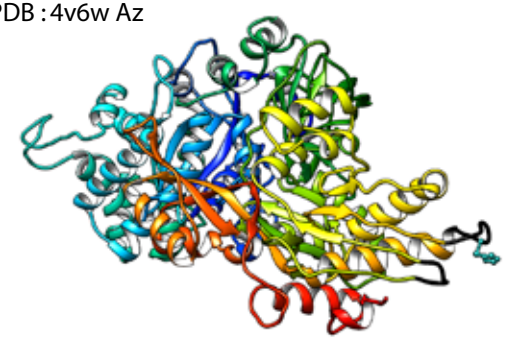
Supplementary Fig. S2 - Multiple sequence alignment of archaeal and eukaryotic EF-2 and EF-2 paralogs showing domain IV sequence motifs. (a) Multiple sequence alignment of a selected set of EF-2 from representative organisms, showing domain IV sequence motifs as in Figure 3. *Bona fide* EF-2 homologues are shaded in grey. Organisms lacking diphthamide biosynthesis genes are indicated with 'a'. (b) Diphthamide modification motifs are not conserved in parabasalid EF-2. EF-2 sequences were mined and aligned from representative genomes or transcriptomes from each of the major lineages of eukaryotes and diphthamide-interacting residues are colored. Here, we show a representative subset of eukaryotes, all surveyed genomes can be found in File S1. Eukaryotic relationships are shown with a schematic cladogram. *Bona fide* diphthamidylated EF-2 sequences are shaded in purple. Boxed region indicates the region that is not conserved in most parabasalids. Parabasalid EF-2 paralogs with unsubstituted diphthamide modification motifs are shaded in yellow. All parabasalids to not encoded diphthamide biosynthesis genes as indicated with the 'no DPH' icon. SAR, Stramenopilia, Alveolata, Rhizaria; DPH, diphthamide biosynthesis genes. The *Pentatrichomonas hominis* and *Tetratrichomonas gallinarum* sequences were retrieved by assembling the sequencing projects available at the indicated SRA accession numbers. Sequences and assembly are available upon request.

a

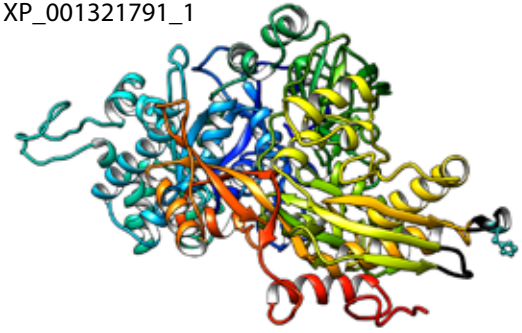
Dph+ Eukaryota canonical eEF-2

Dph- Trichomonas

D. melanogaster eEF-2
PDB : 4v6w Az



Trichomonas vaginalis G3
XP_001321791_1



diphthamide loop

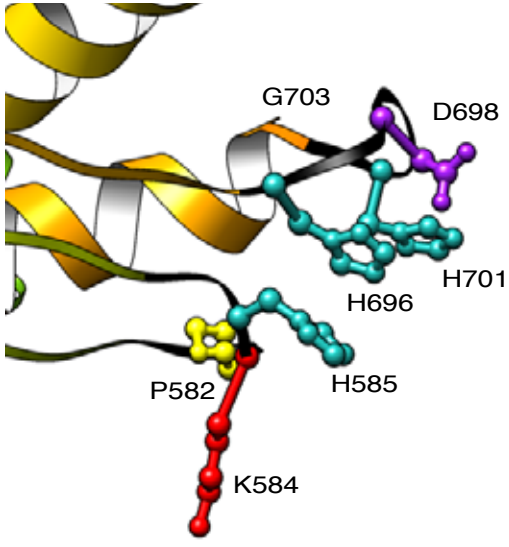
second loop

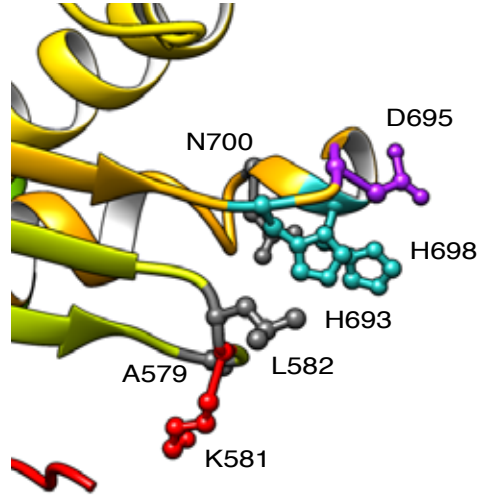
diphthamide loop

HADAAHRN

second loop

SANKLN

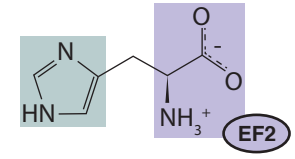




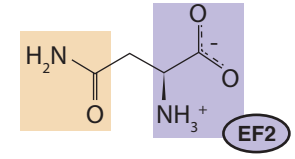
b

Trichomonas (HRN)

H: Histidine

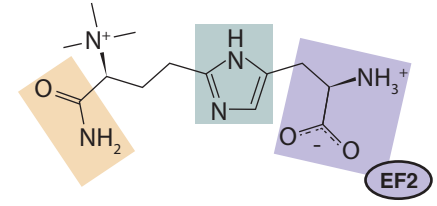


N: Asparagine

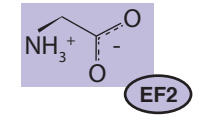


Canonical EF2 (H[dph]RG)

Diphthamide

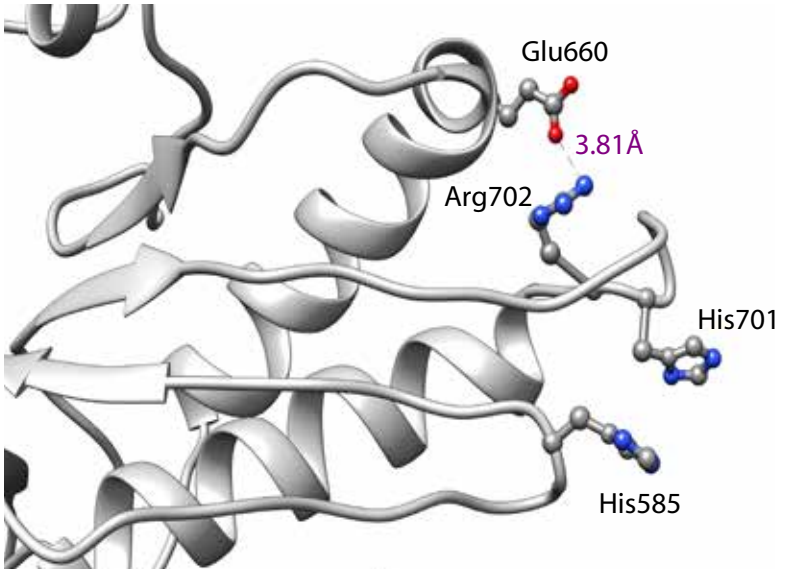


G: Glycine

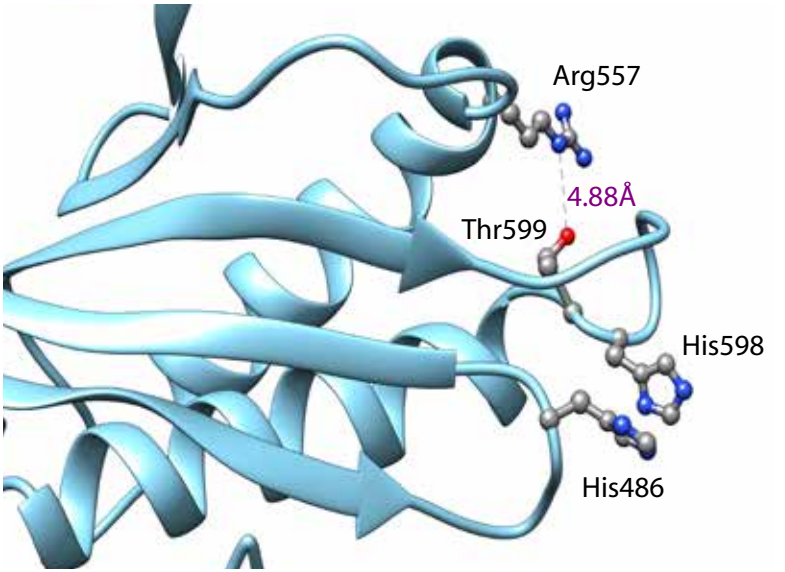


Supplementary Fig. S3 - EF-2 gene from Dph-lacking *Trichomonas vaginalis* shown aligned to *D. melanogaster* eEF-2 structure. (a) Panels are as in Figure 3. *T. vaginalis* EF-2 fits closely to *D. melanogaster* structure (RMSD of 1.589 Å across all 830 residues). While overall structure is maintained, certain key residues in domain IV loops are not conserved. (b) Structure of the three last amino acids comprising the diphthamide loop in EF-2 of *T. vaginalis* compared to canonical eukaryotic EF-2. The amino acids comprising the DRG motif of canonical EF2 (with D referring to diphthamide) have a backbone highly similar to the HRN motif of *T. vaginalis* (with the histidine being not modified to diphthamide). The mutation of the canonical G to N, which provides an amide group, may compensate for the lack of the modification of the histidine.

a



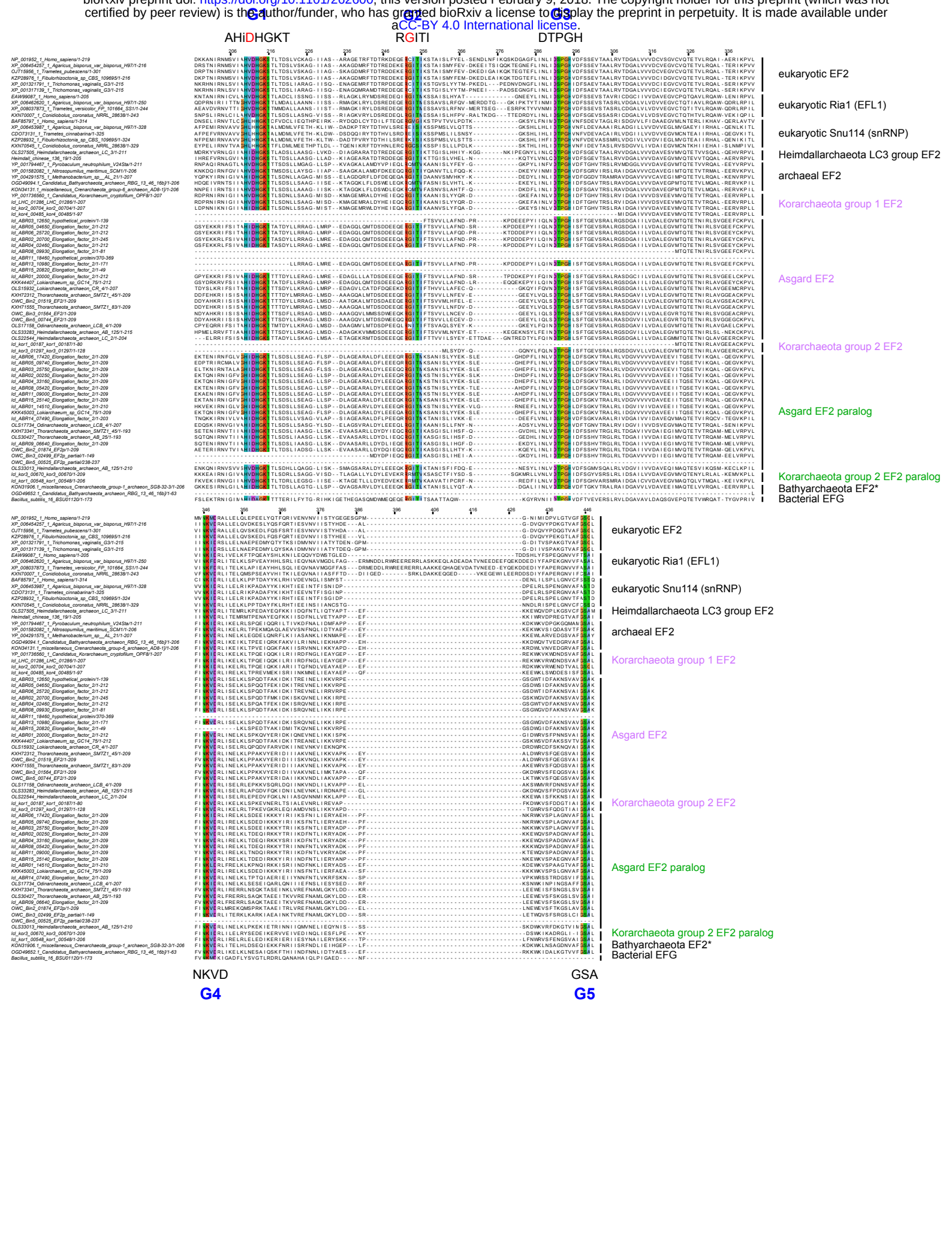
b



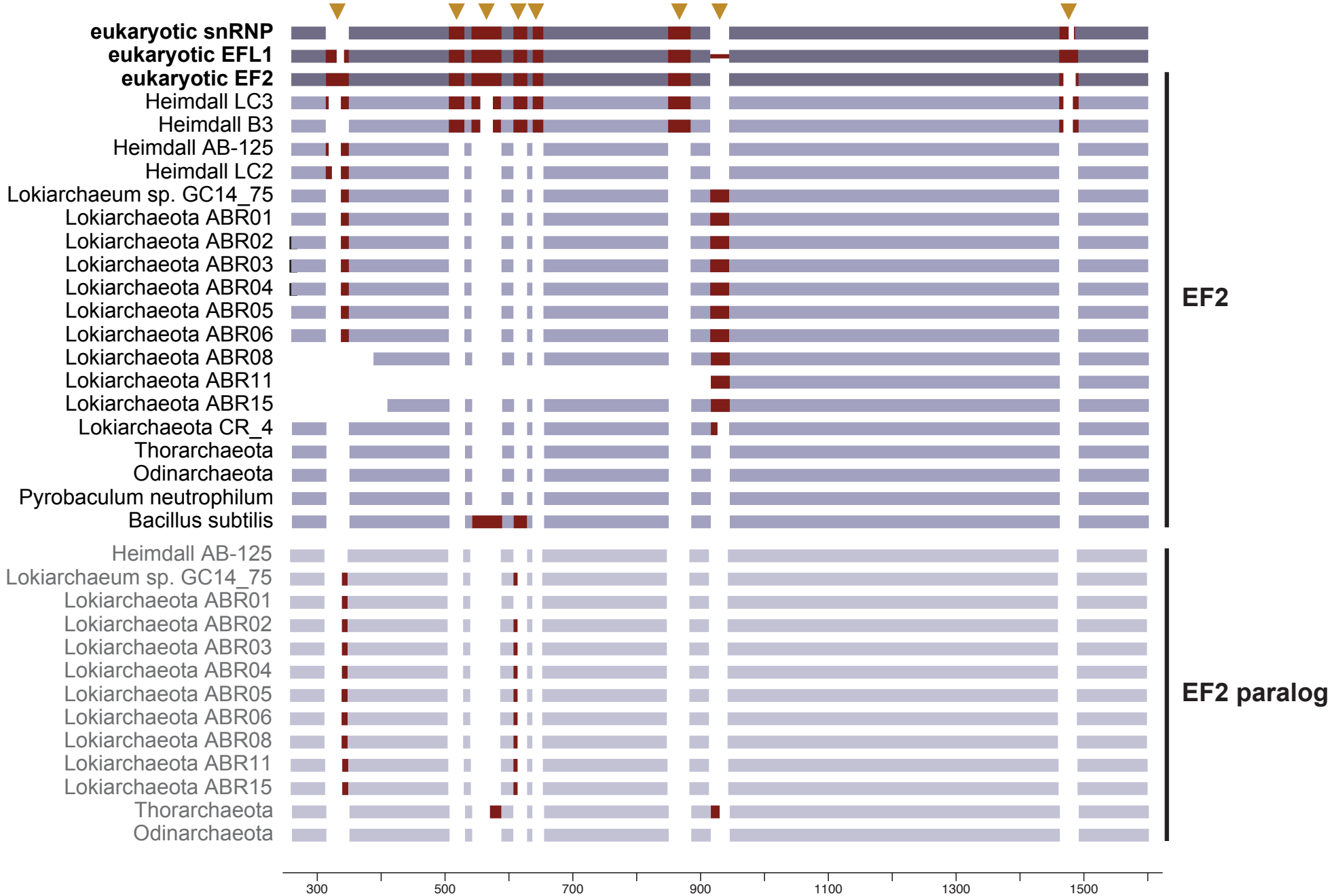
c

	Glu660	Arg702	Diphthamide synthesis pathway	EF2 paralog in genome
Eukaryotes	E	R	+	-
Archaea	E/D	R	+	-
Heimdallarchaeota LC3	E	R	+	-
Trichomonas	E	R	-	-
Korarchaeum 2/4/cryptofilum & Geoarchaea	E/D	R	-	-
Korarchaeum 1/3	R	T	-	+
Asgard	K/R	T	-	+
Heimdall AB125 OLS33283	N	T	-	+

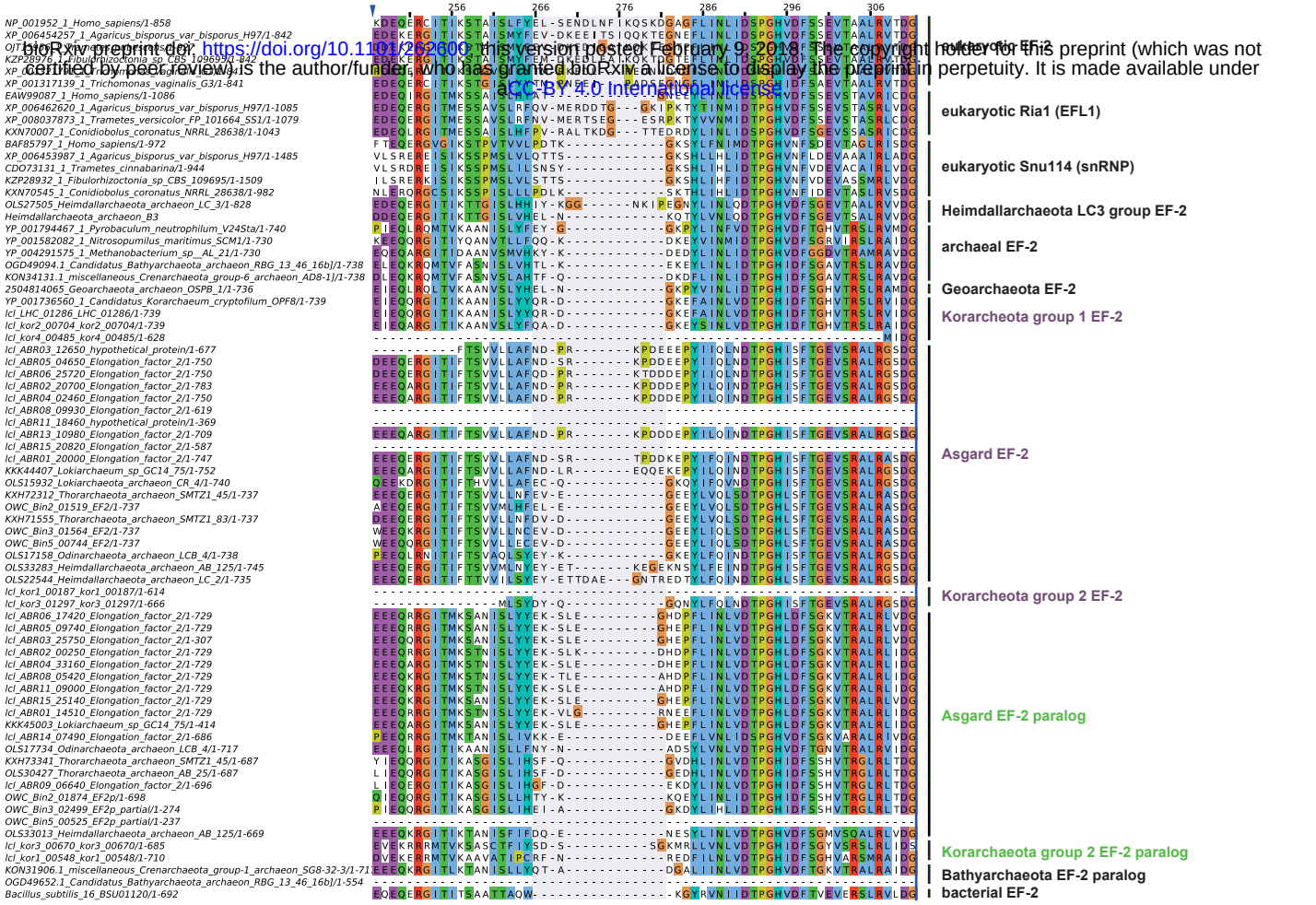
Supplementary Fig. S4 - A universally conserved EF-2 domain IV salt bridge is replaced by conserved correlated mutations in EF-2p containing genomes. (a) EF-2 from the *D. melanogaster* EF2 cryo-EM structure (Anger AM, Armache JP, Berninghausen O, Habeck M, Subklewe M, Wilson DN, Beckmann R. Nature 497:80-5, 2013, doi:10.1038/nature12104) shows that Glu660 and Arg702, which are universally conserved in all archaeal and eukaryotic genomes lacking aEF-2p (Figure 3), form a salt bridge that stabilizes the diphthamide-containing loop of domain IV. (b) Representative modeled EF-2 structure of Thorarchaeota OWC Bin 2, with correlated mutations to Arg557 and Thr599 highlighted. (c) Thr599 is conserved in all EF-2p-containing genomes, and the correlated mutation at the Arg557 position is almost always positive or polar.

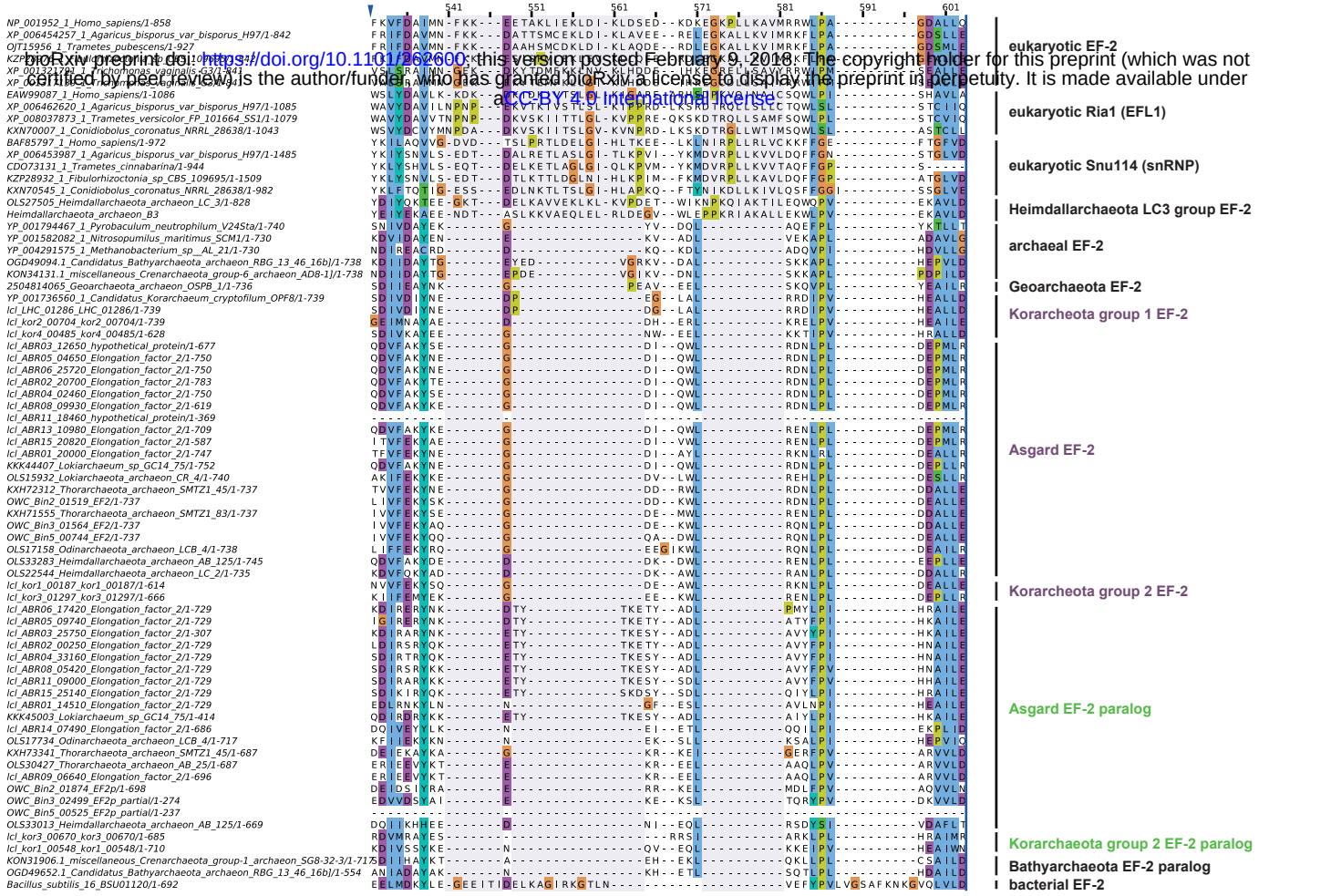


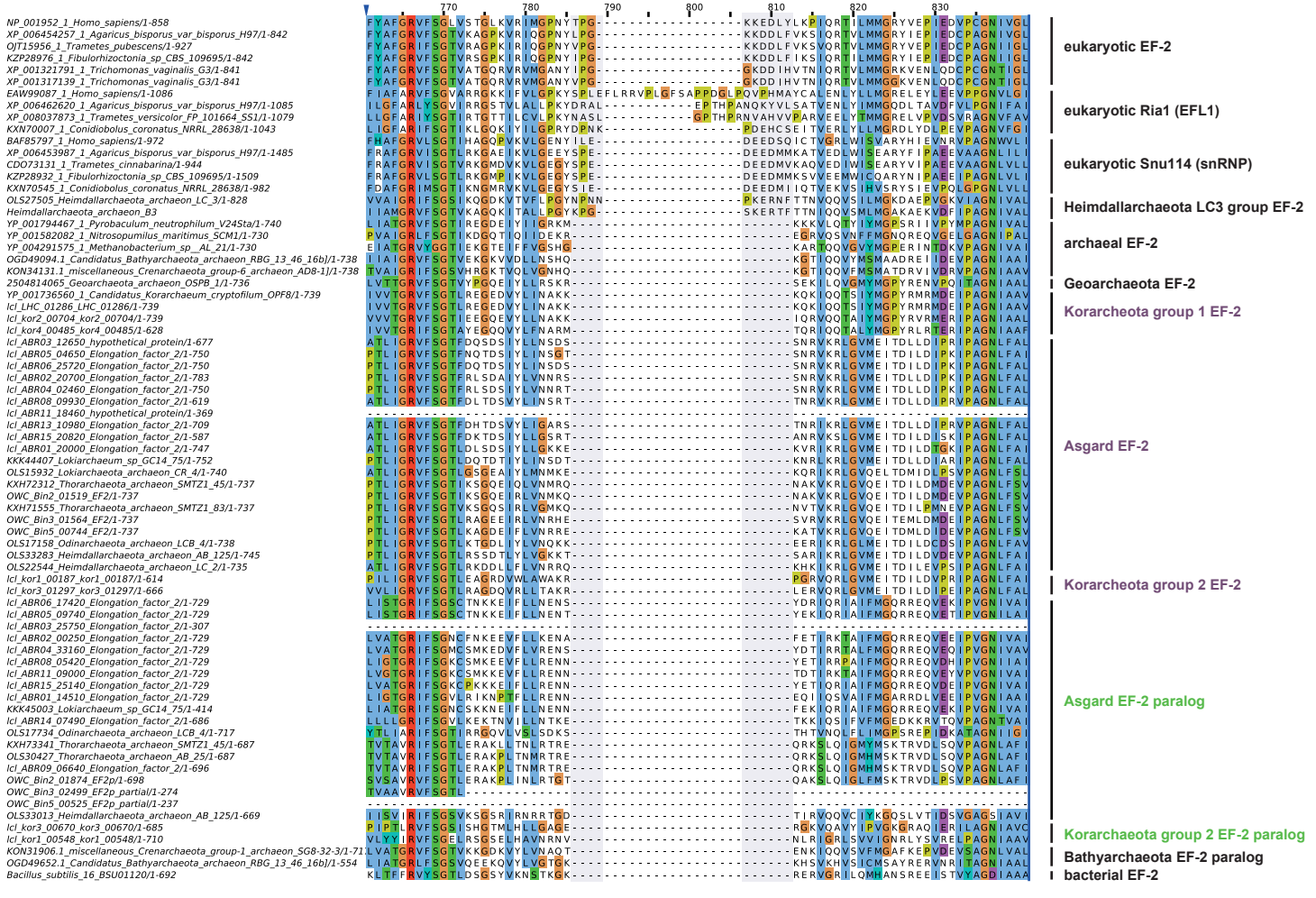
Supplementary Fig. S5: Multiple sequence alignment showing conservation of GTP binding region motifs in Asgard aEF-2 and EF-2 paralogs. Multiple sequence alignment of eEF-2, eEF-2 paralogs, aEF-2, aEF-2 paralogs and bacterial EF-G. Conserved GTP binding motifs G1 - G5 are shown in color in the alignment. Archaeal 60% consensus motif sequences as identified by Atkinson [BMC Genomics 16:78, 2015, doi:10.1186/s12864-015-1289-7] are shown outside the alignment and residues associated with cation binding are shown in red.

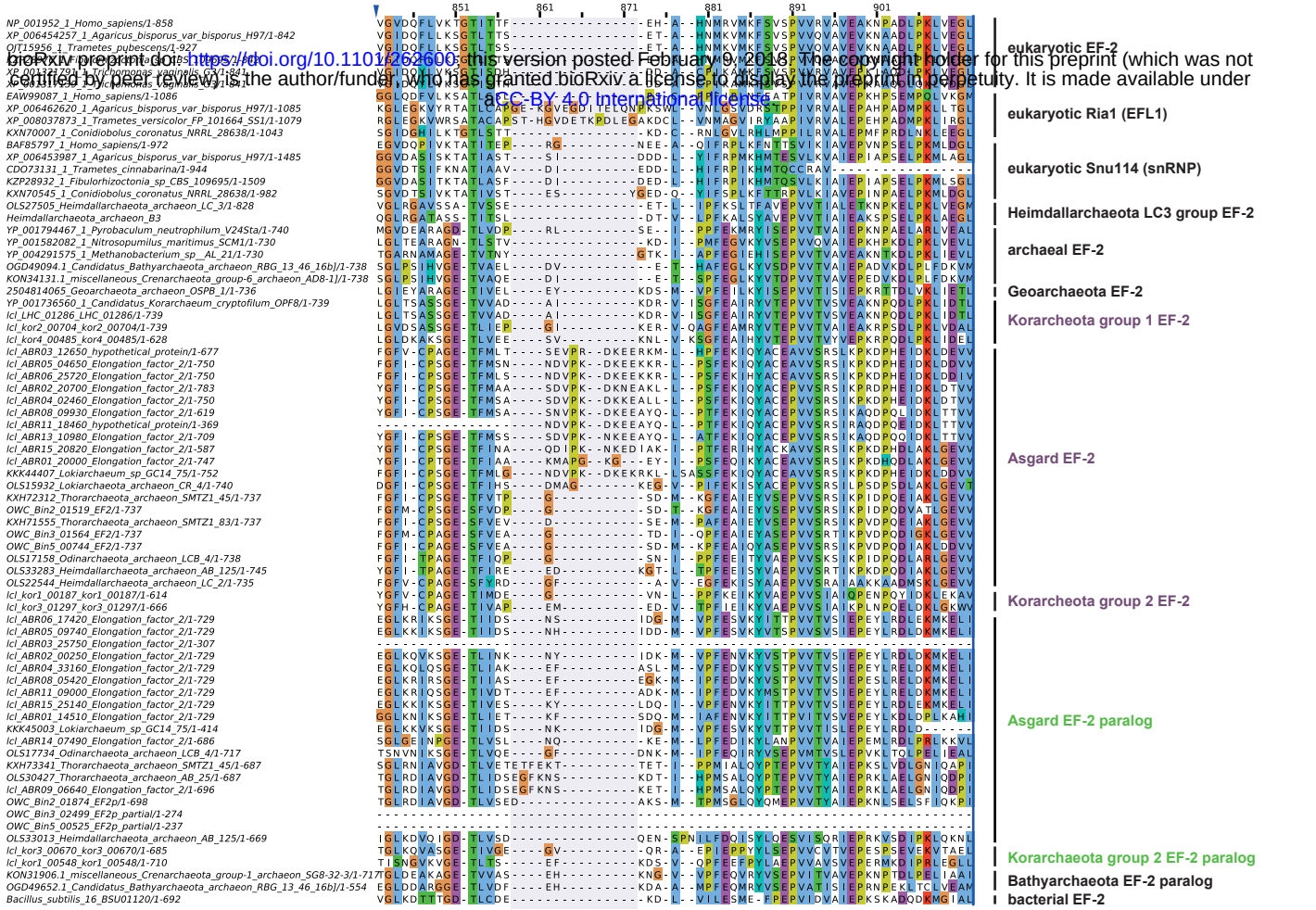

eukaryotic snRNP
eukaryotic EFL1
eukaryotic EF2
Heimdall LC3
Heimdall B3
Heimdall AB-125
Heimdall LC2
Lokiarchaeum sp. GC14_75
Lokiarchaeota ABR01
Lokiarchaeota ABR02
Lokiarchaeota ABR03
Lokiarchaeota ABR04
Lokiarchaeota ABR05
Lokiarchaeota ABR06
Lokiarchaeota ABR08
Lokiarchaeota ABR11
Lokiarchaeota ABR15
Lokiarchaeota CR_4
Thorarchaeota
Odinarchaeota
Pyrobaculum neutrophilum
Bacillus subtilis
EF2
Heimdall AB-125
Lokiarchaeum sp. GC14_75
Lokiarchaeota ABR01
Lokiarchaeota ABR02
Lokiarchaeota ABR03
Lokiarchaeota ABR04
Lokiarchaeota ABR05
Lokiarchaeota ABR06
Lokiarchaeota ABR08
Lokiarchaeota ABR11
Lokiarchaeota ABR15
Thorarchaeota
Odinarchaeota
EF2 paralog
300
500
700
900
1100
1300
1500

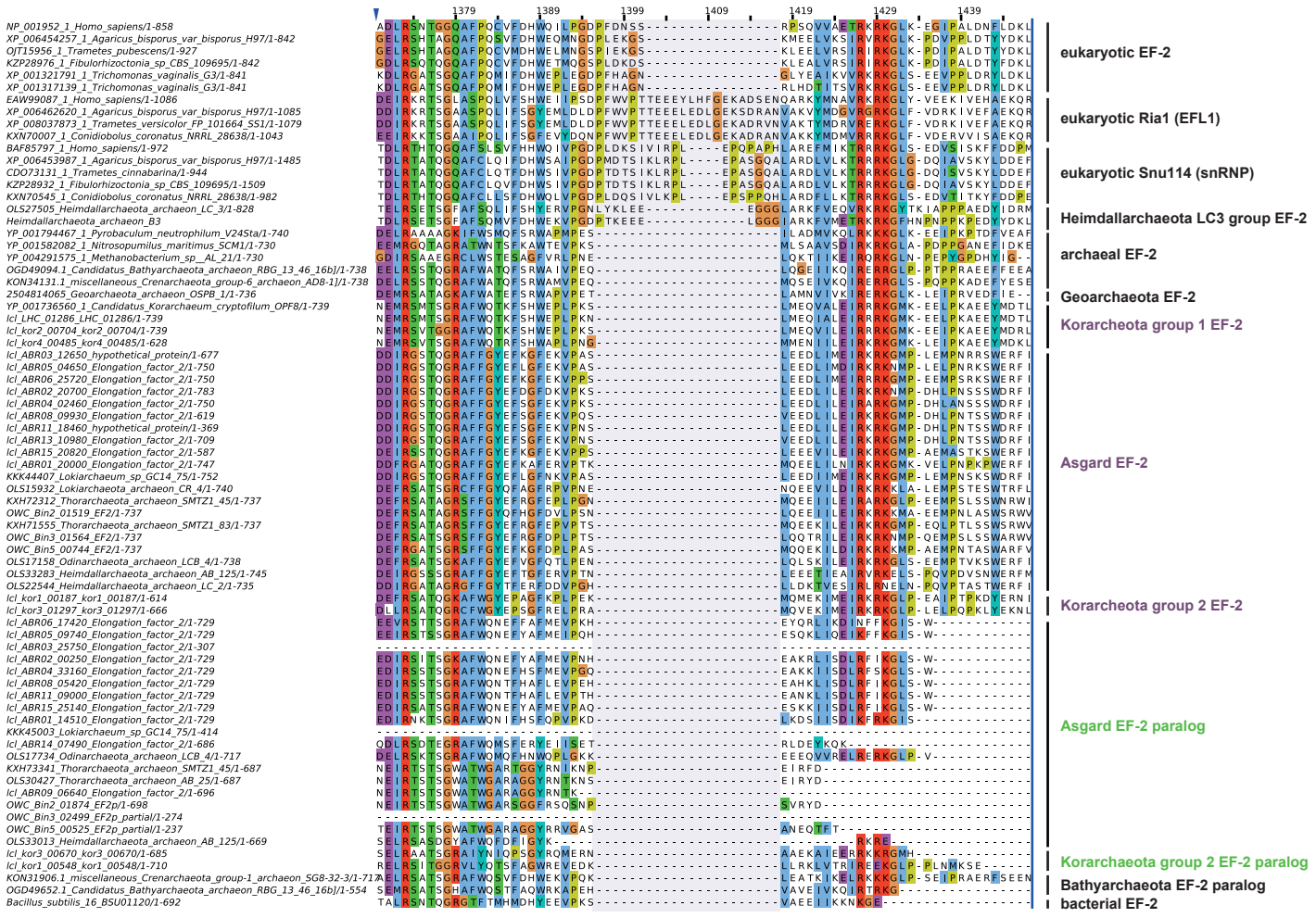
Supplementary Fig. S6: Schematic view of occurrence of indels in archaeal, eukaryotic and bacterial EF-2 and EF-2 paralogs. The cartoon is based on an alignment of EF-2 and EF-2 paralogs from a selected set of representative organisms, mainly comprising Asgard archaea and Eukaryotes. Canonical EF-2 sequences are represented by purple and EF-2 paralogs by light purple bars. Potential indels are shaded by red bars and indel positions are highlighted with orange triangles.



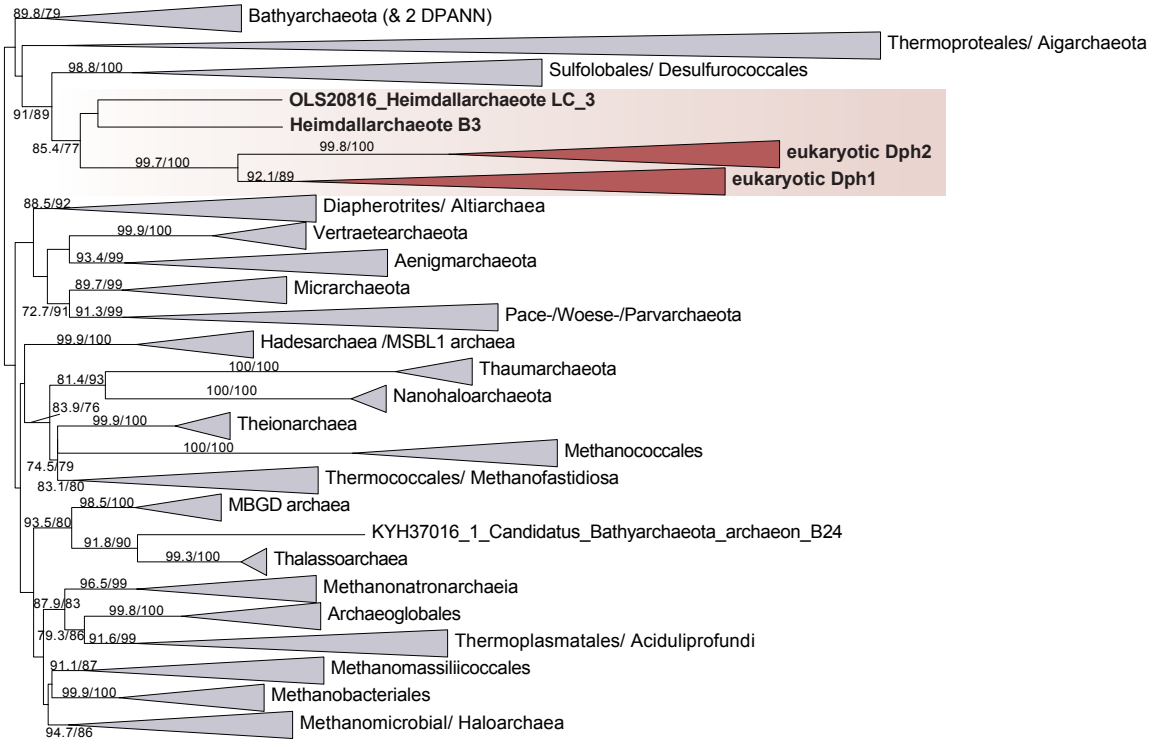


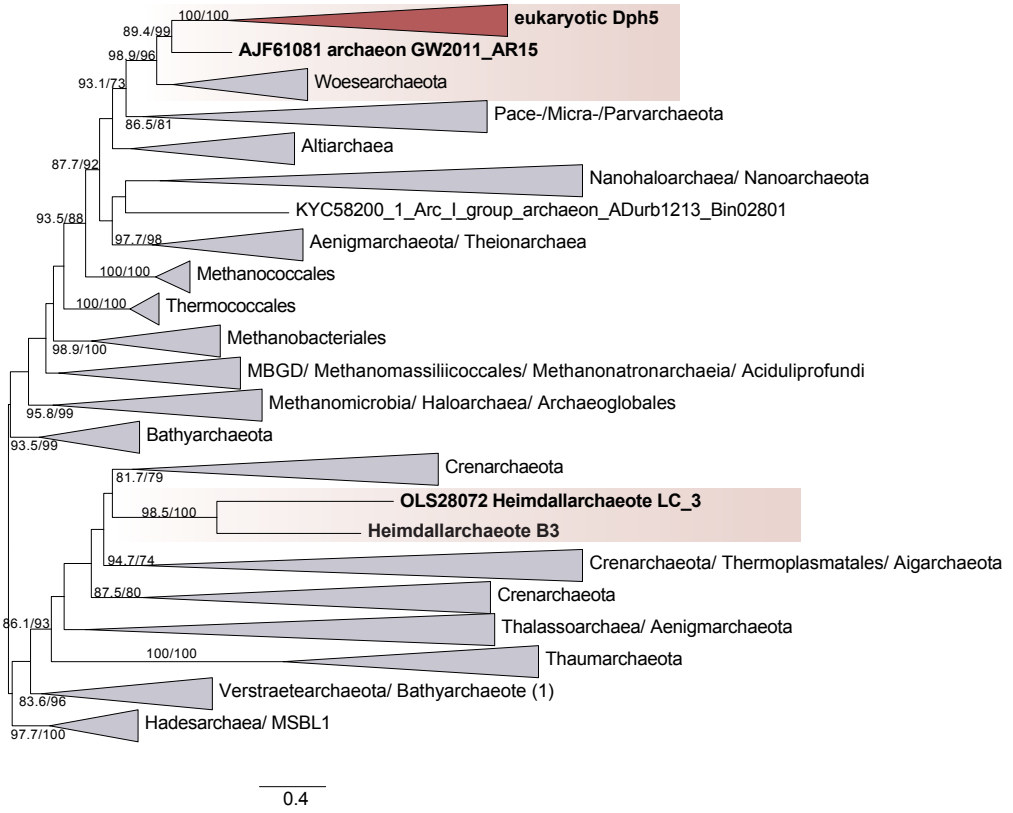






Supplementary Fig. S7: Multiple sequence alignment of occurrence of indels in archaeal, eukaryotic, and bacterial EF-2 and EF-2 paralogs. Selected characteristic indel regions derived from the multiple sequence alignment of a representative set of EF-2 homologs, which provided the basis for Fig. S6. Indel positions are shaded in light purple.


89.8/79
Bathyarchaeota (& 2 DPANN)
Thermoproteales/ Aigarchaeota
98.8/100
Sulfolobales/ Desulfurococcales
OLS20816_Heimdallarchaeote LC_3
91/89
Heimdallarchaeote B3
85.4/77
99.8/100
eukaryotic Dph2
99.7/100
92.1/89
eukaryotic Dph1
88.5/92
Diapherotrites/ Altiarchaea
99.9/100
Vertraetearchaeota
93.4/99
Aenigmarchaeota
89.7/99
Micrarchaeota
72.7/91
91.3/99
Pace-/Woese-/Parvarchaeota
99.9/100
Hadesarchaea /MSBL1 archaea
100/100
Thaumarchaeota
81.4/93
100/100
Nanohaloarchaeota
83.9/76
99.9/100
Theionarchaea
100/100
Methanococcales
74.5/79
83.1/80
Thermococcales/ Methanofastidiosa
98.5/100
MBGD archaea
93.5/80
KYH37016_1_Candidatus_Bathyarchaeota_archaeon_B24
91.8/90
99.3/100
Thalassoarchaea
96.5/99
Methanonatronarchaeia
87.9/83
99.8/100
Archaeoglobales
79.3/86
91.6/99
Thermoplasmatales/ Aciduliprofundi
91.1/87
Methanomassiliicoccales
99.9/100
Methanobacteriales
94.7/86
Methanomicrobial/ Haloarchaea
0.5


100/100
eukaryotic Dph5
89.4/99
AJF61081 archaeon GW2011_AR15
98.9/96
93.1/73
Woesearchaeota
Pace-/Micra-/Parvarchaeota
86.5/81
Altiarchaea
87.7/92
Nanohaloarchaea/ Nanoarchaeota
KYC58200_1_Arc_I_group_archaeon_ADurb1213_Bin02801
93.5/88
97.7/98
Aenigmarchaeota/ Theionarchaea
100/100
Methanococcales
100/100
Thermococcales
98.9/100
Methanobacteriales
MBGD/ Methanomassiliicoccales/ Methanonatronarchaeia/ Aciduliprofundi
95.8/99
Methanomicrobia/ Haloarchaea/ Archaeoglobales
93.5/99
Bathyarchaeota
Crenarchaeota
81.7/79
OLS28072 Heimdallarchaeote LC_3
98.5/100
Heimdallarchaeote B3
94.7/74
Crenarchaeota/ Thermoplasmatales/ Aigarchaeota
87.5/80
Crenarchaeota
86.1/93
Thalassoarchaea/ Aenigmarchaeota
100/100
Thaumarchaeota
83.6/96
Verstraetearchaeota/ Bathyarchaeote (1)
97.7/100
Hadesarchaea/ MSBL1
0.4

Supplementary Fig. S8: Maximum likelihood phylogenetic analyses of Dph1/2 (a) and Dph5 (b) performed using IQ-tree. Eukaryotic homologs were collapsed and are represented by dark red triangles, while homologs of Woesearchaeota and Heimdallarchaeota are shaded in light red. Both phylogenetic trees are unrooted. Values on branches refer to support values based on ultrafast bootstrap approximation as well as single branch tests. Whenever any of the two support values were lower than 70, bootstraps were removed. Scale bar indicates the number of substitutions per site.