Complex Evolutionary History of Translation Elongation Factor 2 and Diphthamide Biosynthesis in Archaea and Parabasalids

Adrienne B. Narrowe1,†, Anja Spang2,3,†, Courtney W. Stains3, Eva F. Caceres3, Brett J. Baker4, Christopher S. Miller1,*,†, and Thijs J.G. Ettema3,†

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

†These authors contributed equally to this work.
*Corresponding author: E-mail: chris.miller@ucdenver.edu.
Accepted: July 26, 2018
Data deposition: New sequencing for this project has been deposited in GenBank under the accession numbers listed in supplementary file S1, Supplementary Material online.

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, diphthamide.

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 neofunctionalizations 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 (Bécam et al. 2001). Bacterial EF-G is involved in both translocation and ribosome recycling and has undergone multiple duplications (Atkinson and Baldauf 2011; Atkinson 2015), including subfunctionalizations separating the translocation and ribosome recycling functions (Tsuboi et al. 2009; Suematsu et al. 2010). Several more ancient duplications have also been identified in bacteria; these duplications
have led to neofunctionalizations including roles in termina-
tion (Freistroffer et al. 1997), ribosome biogenesis (Gibbs and
Fredrick 2018), in tetracycline resistance (Donhofer et al.
2012), and roles for which no function has yet been deter-
dined (Margus et al. 2011). However, to date, archaea were
thought to encode only a single essential protein within this
family, that is, archaeal EF-2 (aEF-2) (Atkinson 2015).

Unlike bacterial EF-Gs, archaean and eukaryotic EF-2s con-
tain a posttranslationally modified amino acid which is syn-
thesized upon the addition of a 3-amino-3-carboxypropyl
(ACP) group to a conserved histidine residue and its subse-
quent modification to diphthamide by the concerted action of
three (in archaea) to seven enzymes (in eukaryotes) (de Cr
cy-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 investi-
gation. Yeast mutants incapable of synthesizing diphthamide
have a higher rate of translational frame shifts, suggesting
that this residue plays a critical role in reading frame fidelity
during translation (Ortiz et al. 2006). Furthermore, structural
studies of eEF-2 using high-resolution Cryo-EM have indicated
that diphthamide interacts directly with codon–anticodon
bases in the translating ribosome, and facilitates translocation
by displacing ribosomal decoding bases (Anger et al. 2013;
Murray et al. 2016). In addition, diphthamine has been pro-
posed to play a role in the regulation of translation, as it
represents a site for reversible endogenous ADP-ribosylation
(Schaffrath et al. 2014), and in the selective translation of
certain genes in response to cellular stress (Arguelles et al.
2014). Given its anticipated role at the core of the transla-
tional machinery, it is not surprising that, with the sole excep-
tion of Korarchaeum cryptofilum (Elkins et al. 2008; de Crécy-
Lagard et al. 2012), the diphthamide biosynthetic pathway is
universally conserved in all archaea and eukaryotes. Indeed,
while not strictly essential, loss of diphthamide biosynthesis
has been shown to result in growth defects in yeast (Kimata
and Kohn 1994; Ortiz et al. 2006) and some archaea (Blaby
et al. 2010), and is either lethal or causes severe development-
al abnormalities in mammals (Liu et al. 2006; Webb et al.
2008; Yu et al. 2014).

In the current study, we explore the evolution and function of
EF-2 and of diphthamide biosynthesis genes using genomic
data from novel major archaean lineages that were recently
discovered using metagenomics and single-cell genomics
approaches (Hug et al. 2016; Adam et al. 2017; Spang et
al. 2017). In particular, we report the presence of EF-2
paralogs in many archaean genomes belonging to the
Asgard archaea, Korarchaeota and Batharchaeota (Meng
et al. 2014; Evans et al. 2015; Spang et al. 2015; He et al.
and the unexpected absence of diphthamide biosynthesis
genes in several archaea and in parabasalid 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 archaean lin-
eages. 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

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 con-
tigs, probably due to the low abundances of these organisms.
In an attempt to produce longer contigs, we coassembled
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) (Ounite
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 coassembled 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 pro-
vided 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 (Ahnberg 2014). Coverage profiles
required by CONCOCT were estimated using kallisto (version
0.43.0, quant –plaintext) (Bray et al. 2016). All available sam-
pies 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 (2,000 and 3,000 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 \( \sim 2.5 \) times the estimated number of species in the sample (Alneberg J, personal communication), resulting in 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.

Binning of ABR Loki- and Thorarchaeota

Several binning tools with different settings were run independently: CONCOCT_2000: version 0.4.0, \(-\text{read\_length} 200\) and minimum contig length of 2,000. CONCOCT_3000: version 0.4.0, \(-\text{read\_length} 200\) and minimum contig length of 3,000. In both cases, coverage files were created by mapping all eight samples against the coassembly using kallisto. MaxBin2: version 2.2.1, \(-\text{min\_contig\_length} 2000\) -markerset 40 –plotmarker (Wu et al. 2016). The eight samples were mapped against the coassembly using Bowtie2. Coverage was estimated using the getabund.pl script provided. MyCC_4mer: 4mer -t 2000 (Lin and Liao 2016). MyCC_56mer: 56mer -t 2000. Both coverage profiles were obtained as the authors described in their manual.

The results of those five binning methods were combined into a consensus: contigs were assigned to bins if they had been classified as the same organism by at least three out of five methods. The resulting bins were manually inspected and cleaned further using mmgenome (Albertsen 2013). Completeness and redundancy was computed using CheckM (Parks et al. 2015).

Sampling and Sequencing of OWC Thorarchaeota

Eight soil samples were collected from the Old Woman Creek (OWC) National Estuarine Research Reserve and DNA was extracted as described previously (Narow et al. 2017). Library preparation and five lanes of Illumina HiSeq 2x125 bp sequencing followed standard operating procedures at the US DOE Joint Genome Institute (GOLD study ID Gs0114821). Sample M3-C4-D3 had replicate extraction, library preparation, and two lanes of sequencing performed, and reads were combined before downstream analysis. For three additional samples (M3-C4-D4, O3-C3-D3, O3-C3-D4) one lane of sequencing was performed. For the other four samples (M3-C5-D1, M3-C5-D2, M3-C5-D3, M3-C5-D4) DNA was sheared to 300 bp 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 five 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, hdist1 = 1 hdist2 = 1 ftm = 5, maq = 8, maxns = 1, minlen = 40, minlenfraction = 0.6, k = 27, hdist = 1, trimq = 12, qtrim=r. 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 (5 kb contigs, 3 kb 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 nine OWC libraries were mapped to this bin using bbmap 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 Bowtie2 (Langmead and Salzberg 2012) to generate a coverage profile which was used to manually 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

For eukaryotes, the precomputed eggNOG members data set (available at http://eggnogdbl.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. The eggNOG database consists of nonsupervised orthologous groups that comprises 2,031 eukaryotic and prokaryotic organisms (Huerta-Cepas et al. 2016). Trees, alignments, profiles and functional annotations are available and can be explored through their website or downloaded in bulk, which makes it a very useful resource for comparative analyses. A complete list of eukaryotic genomes surveyed can be found in supplementary file S1, workbook 2,
Supplementary Material online. For genomes not represented in eggNOG, we manually searched these COGs against publicly available genomes, as indicated by “orthology assignment source” in supplementary file S1, Supplementary Material online. We also manually searched for COGs and BLAST homologs to S. cerevisiae EF-2 and diphthamide biosynthesis genes in a number of transcriptome sequencing projects for the following parabasalids: Histomonas meleagridis (NCBI short read archive: SRR553451), Dientamoeba fragilis (SRR2039085), Tetraichromonas gallinarum (short read archive: SRR2989159), Tritrichomonas foetus (Bioproject: PRJNA345179), Trichomonas tenax (clone library: D78481), and Pentatrichomonas hominis (short read archive: SRR4111571).

Similarly, an in-house arCOG data set, 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 as previously described (Spang et al. 2015) (supplementary file S1, Supplementary Material online). For the Thorarchaeota OWC Bin 2, 3, and 5 assemblies, in order to exclude the possibility that Thorarchaeota dph2 and dph5 genes might exist but were not correctly binned in the Thorarchaeota genomes, all contigs with matching HMM hits to dph2 and dph5 in the full OWC assembly were manually examined for potential Thorarchaeal dph genes. Homology searches and phylogenetic analyses of all dph2 and dph5 genes identified in the entire metagenome confirmed their membership in non-Thorarchaeota clades known to be represented in the metagenomic samples, and adjacent genes on all dph2- and dph5-containing contigs also had homology to known non-Thorarchaeota taxa.

As archaea only possess Dph2, Dph5, and Dph6 homologs (de Crécy-Lagarde et al. 2012; Schaffrath et al. 2014), there is no archaeal arCOG for Dph3. To confirm that there are no Dph3 homologs in archaea, including the newly described archaea, we used COG5216 to search the archaeal genomes listed in supplementary file S1, Supplementary Material online. Dph4 contains a DnaJ domain (COG0484) and an additional CSL zinc-finger domain which distinguishes Dph4 from DnaJ (Liu et al. 2004). To confirm that there are no Dph4 homologs in archaea, including the newly described archaea, we searched all genomes listed in supplementary file S1, Supplementary Material online, using PFAMs for DnaJ (PF00226) and ZF-CSL (PF05207). While eukaryotic Dph4 genes contain both these domains, our searches identified no archaeal genes containing hits to both these domains.

To verify that we had identified all EF-2 homologs in the set of archaeal genomes, we also constructed a full-length EF-2/aEF-2p HMM from all archaeal EF-2 homologs included in supplementary file S1, Supplementary Material online. We searched this hmm against the set of Swiss-Prot reviewed sequences of LepA, TypA/BipA, SeB, aIF5B, and EF1α to determine the e-value at which this HMM begins to detect ancient, distant paralogs of EF-2 (1e-24). We then searched all archaeal proteomes listed in supplementary table S1, Supplementary Material online, using this HMM and the identified e-value cutoff to identify any additional candidate EF-2 paralogs. We detected no additional aEF-2p proteins, and the handful of protein sequences with hits below 1e-24 which were not monophyletic with the archaeal EF-2 and aEF-2p proteins appear to be derived from misbinned contigs (Arc I group archaeon U11s0528-Bin89: KYC51594, KYC51595) or possible misassemblies (Candidatus Altarchaeales: ODS41826, ODS42854).

Construction of Multiple Sequence Alignments and Phylogenetic Analyses

EF-2 Alignment and Phylogeny

EF-2 and EF-2 paralogs of Asgard archaea, Koarchaeota, and Bathyarchaeota identified as described earlier were aligned with representative sets of archaeal aEF-2, bacterial EF-G and eukaryotic eEF-2, EFL1 and Snu114 homologs using mafft-linsi (Katoh and Standley 2013). Sequence data sets were iteratively refined upon inspection of alignments and phylogenetic trees to remove sequences that were highly partial or emerged on single long branches and therefore could cause phylogenetic artefacts. Additionally, for the composite Lokiarchaeum GC14_75 genome bin (Spang et al. 2015), which is comprised of 1.5 closely related strains, only one nonredundant homolog each of aEF-2 and aEF-2p was retained, while partial redundant copies of aEF-2 (p) were excluded from the final analysis.

Alignments were viewed in Jalview and annotated in Adobe Illustrator. For subsequent phylogenetic analyses, alignments were trimmed using BMGE (Criscuolo and Gribaldo 2010) (blossom 30, entropy score of 0.55) yielding 620 positions both in the alignment with and without bacterial outgroup. Maximum likelihood analyses were performed using IQ-tree using the mixture model LG+C60+i+G+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. In each case, branch supports were assessed using ultrafast bootstrap approximation as well as with single branch test (-alrt option).

Extended EF-2 Family Protein Phylogeny

In order to get an unbiased view of the phylogenetic relationship of EF-2 homologs of archaea with respect to all eukaryotic and bacterial homologs, all sequences from cellular organisms (NCBI taxid 131567) assigned to the homologous superfamily “EF-G domain III/V-like” (IPR005225) and encoding the signature domain “Small GTP-binding protein domain” (IPR005225) were downloaded from uniprot (www.uniprot.org; last accessed May 8, 2018). These IPR domains capture all major lineages within the EF-2 family.
proteins, including archaeal aEF-2 and aEF-2p homologs, eu-
karyotic EF-2, EFL1, and Snu114 homologs, bacterial EF-G, as
well as bacterial tetracycline resistance proteins (Tet), GTP-
binding protein TypA (TypA/BipA), Peptide chain release fac-
tor 3 (RF3), LepA (Elongation factor 4), and various EF-G
homologs (such as EF-G1, EF-G2, and EF-GII). This set of pro-
tein sequences (94,830) was filtered to keep only sequences
without X’s and that contained between 500 and 1,200 posi-
tions (98% of all sequences) to exclude poor quality and par-
tial sequences and to prevent misalignments to rare
insertions. In addition, archaeal, bacterial, and eukaryotic se-
quence sets were separately clustered with CD-HIT (Fu et al.
2012) using a sequence similarity cutoff of 63% (for bacterial
and eukaryotic homologs) and 90% (for archaeal homologs).
Finally, all data sets were combined, with selected eukaryotic
EF-2 homologs and homologs from the novel members of the
Asgard archaea and Korarchaeota added. Because these
sequences comprise various different protein families with
different domains and domain architectures they could not
be reliable aligned using mafft. Therefore, and based on the
fact that these sequences share the characteristic GTP-binding
domain, we used the hmm-profile of PF00009 (i.e., GTP-
binding elongation factor family, EF-Tu/EF-1A subfamily) as
seed for an hmm-alignment. The alignment was trimmed us-
ing BMGE (Cricuzzo and Gribaldo 2010) with blossom 30
and an entropy score of 0.6. Subsequent phylogenetic anal-
yses were performed using FastTree (LG, gamma). Several
rounds of iterative refinements were performed to be able
to remove poorly aligned sequences and/or extremely long
branches in the resulting tree. The final sequence alignment
consisted of 3,270 sequences and 138 aligned sites.

Phylogeny of Diphthamide Biosynthesis Proteins (Dph1/
Dph2 |PR016435; arCOG04112] and Dph5 |PR004551;
arCOG04161])

Both Dph1 and Dph2 as well as Dph5 homologs of a rep-
sentative 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 and supplementary file S1, Supplementary Material
online) 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 (Cricuzzo and Gribaldo 2010) using the blossom
30 matrix and setting the entropy to 0.55. This resulted in final
alignments of 170 (Dph1/2) and 221 (Dph5). Maximum like-
lihood analyses were performed using IQ-tree (Nguyen et al.
2015) with the mixture models selected among the C-series
models based on its Bayesian information criterion score by
the built-in model test implemented in IQ-tree: LG+C50+R+F
(Dph1/2) and LG+C60+R+F (Dph5), re-
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
colocalized 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+R+F model chosen based on best BIC score as
described earlier. Branch supports were assessed using ultra-
fast bootstrap approximation as well as with the single branch
(-alrt option) in IQ-tree.

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 using UCSF Chimera version
1.11.12 (Petterson 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). Briefly, the nonredundant
PDB database provided by I-TASSER was used for threading
up to 20 template structures per target in the initial constraint
generation steps before structure assembly and refinement,
and for searching for structural homologs of the best model.
A recently published crystal structure of aEF-2 was also added to
this PDB database (Tanzawa et al. 2018).

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 se-
quence 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 cluster centroids only using WebLogo: version 2.8.2 (weblogo.
berkeley.edu) (Crooks et al. 2004).

Accession Numbers

Taxonomy and accession numbers for all genes analyzed in
this study are listed in supplementary file S1, Supplementary
Material online.

Results

Most Asgard Archaea, Korarchaeota, and Geoarchaea as
Well as Parabasalids, Lack Diphthamide Synthesis Genes

It was previously assumed that EF-2 of all eukaryotes and
Archaea was uniquely characterized by the presence of
diphthamide. To examine if this assumption is still valid when taking into account recently sequenced genomes, we surveyed 337 archaeal and 168 eukaryotic genomes (supplementary file S1, Supplementary Material online) for each of the three known archaeal (de Crècy-Lagard et al. 2012) and seven eukaryotic (Su, Chen, et al. 2012; Su, Lin, et al. 2012; Uthman et al. 2013) dph genes. While most archaeal genomes encode clear dph homologues, we failed to detect the diphthamide biosynthesis genes in a large diversity of metagenome-assembled genomes (MAGs) of uncultured archaea, including newly assembled MAGs analyzed for this study (fig. 1 and supplementary fig. S1 and file S1, Supplementary Material online). In particular, our analyses showed that, as reported for K. cryptophilum (Elkins et al. 2008; de Crècy-Lagard et al. 2012), all Korarchaeota and Geoarchaea as well as nearly all members of the Asgard archaea lack the conserved archaeal diphthamide biosynthesis genes dph1/2, dph5, and dph6. As an exception, Asgard...
archaea related 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 Batharchaeota 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 (Lame et al. 2007) (supplementary file S1, Supplementary Material online) 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, Trichomonas foetus, and Dientamoeba fragilis (supplementary file S1, Supplementary Material online). 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 (Rinke et al. 2013; Castelle et al. 2015) 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 on 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 homologs (fig. 1). First, we found that the draft genomes of most Asgard archaea, some Korarchaeota (Kor 1 and 3), and a few Batharchaeota encode two distinctly related EF-2 paralogs. In contrast, the genomes of K. cystophillum and two novel marine Korarchaeota (Kor 2 and 4) and Heimdallarchaeote LC2 and LC3 as well as Geoarchaeae do not encode an EF-2 paralog. Given that the Heimdallarchaeote LC2 genome was estimated to be only 70–79% complete (Zaremba-Niedzwiedzka et al. 2017), and based on phylogenetic analyses (see below), we consider it possible that this genome might encode an as-yet unassembled aEF-2 paralog. The presence of paralogous aEF-2 in most Asgard archaea and some Korarchaeota genomes corresponds with the absence of diphthamide synthesis genes (figs. 1 and 2). Yet, even though the genomes of K. cystophillum, Kor 2, Kor 4, and Geoarchaeae as well as of Heimdallarchaeote LC2 lack dph genes, they do not encode an EF-2 paralog. In all other archaeal genomes, including that of Heimdallarchaeote LC3, the absence of an EF-2 paralog correlates with the presence of dph genes.

Archaea with Two EF-2 Family Proteins Encode Only One Bona Fide EF-2

We next addressed whether residues and structural motifs shown to be necessary for canonical translocation were conserved in the various EF-2 and EF-2 paralogs. Domain IV of EF-2, representing the anticodon mimicry domain, is critical for facilitating concerted translocation of tRNA and mRNA (Rodnina et al. 1997; Ortiz et al. 2006). This domain includes three loops that extend out from the body of EF-2 and interact with the decoding center of the ribosome. The first of these three loops (HxDxxHRG) (canonical residue positions are numbered according to sequence associated with D. melanogaster structural model PDB 4V6W; Anger et al. 2013) contains the site of the diphthamide modified histidine, H701, and is highly conserved across archaea and eukaryotes (Ortiz et al. 2006; Zhang et al. 2008). High conservation is also seen in a second adjacent loop (SPHKHN) in the a/eEF-2 domain IV (S581-N586), which contains a lysine residue (K584) that interacts directly with the tRNA at the decoding center, and is itself positioned by a stacking interaction between PS82 and H585 (Murray et al. 2016). The third loop appears to stabilize the diphthamide loop, partially via a salt-bridge formed between a nearby glutamate residue (E660) and R702 in the diphthamide loop (Anger et al. 2013). Both of these residues are highly conserved among archaea and eukaryotes.

Our analyses reveal that the sequence motifs in these loops are also strictly conserved among the bona fide canonical EF-2 family proteins of the Heimdallarchaeote LC3 lineage, Geoarchaeae, as well as in those Korarchaeota and Batharchaeota that lack an EF-2 paralog (fig. 3 and supplementary fig. S2a, Supplementary Material online). Notably, this conservation is seen irrespective of the presence or absence of dph genes in those genomes. However, most canonical EF-2 of parabasalids (which lack dph genes), possesses a glycine to asparagine mutation at residue 703 (fig. 3 and supplementary figs. S2b and S3a, Supplementary Material online), which may compensate for the lack of the diphthamide residue by contributing an amide group (fig. 3 and supplementary fig. S3b, Supplementary Material online).

In contrast, in those Asgard archaea and Korarchaeota (Kor 1/3 clade) that encode two EF-2 family proteins (aEF-2 and aEF-2p), the canonical aEF-2 copies contain domain IV motifs with reduced conservation. In these genomes, R702 of the diphthamide loop is universally replaced by a threonine residue in aEF-2. In 22 of 22 aEF-2 proteins, there is a correlated mutation of E660 to either arginine or lysine (supplementary fig. S4, Supplementary Material online). Structural homology modeling suggested that these correlated mutations likely prevent unfavorable electrostatic interactions between
EF-2 and Diphthamide in Archaea and Eukaryotes

Fig. 2.—The evolution of archaeal EF-2 family proteins. Rooted phylogenetic tree of EF-2 family proteins based on maximum likelihood analyses of 620 aligned positions using IQ-tree with the LG + C60 + F + G mixture model. The tree was rooted by outgroup rooting with bacterial EF2-family proteins. 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. Sequences from members of the Asgard archaea are labeled in bold. 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 <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. Snu114, US small nuclear ribonucleoprotein; EFL1, elongation factor-like GTPase; n.c., not conserved; p.c., partially conserved; n.d., not determined; a, partial sequences, which therefore lack information regarding the key residues.
domain IV loops, and maintain stabilization of the diphthamide loop (supplementary fig. S4, Supplementary Material online). While G703 is conserved in most aEF-2 of archaea, all Lokiarchaeota (except Lokiarchaeota CR_4) aEF-2 encode either a serine or a glutamine at this site (fig. 3 and supplementary fig. S2a, Supplementary Material online). Furthermore, analysis of the second loop (S581-N586) revealed additional crucial mutations in the canonical EF-2 of these archaea; notably, K584 is not conserved (fig. 3 and supplementary fig. S2a, Supplementary Material online). Despite these aEF-2 modifications which correlate with the presence of an aEF-2p paralog in these archaea, there is still evidence for strong selection pressure maintaining many of the key conserved residues in these domain IV motifs, including H701, the target site of diphthamide modification (fig. 3 and supplementary fig. S2a, Supplementary Material online).

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

**EF-2 Homologs of Archaea Experienced Complex Evolutionary History**

To resolve the evolutionary history of EF-2, we performed phylogenetic analyses of archaeal EF-2 (aEF-2) and aEF-2p, utilizing different sets of bacterial and eukaryotic homologs.
The placement of aEF-2 family proteins in a tree comprising an extensive set of bacterial EF-G (e.g., Tet, TypA/BipA, RF3, LepA, Elongation factor 4, and EF-G1, EF-G2, and EF-GII) and eukaryotic EF-2 family proteins (i.e., EF-2, Ria1 [or Elongation factor like, EFL1] and Snu114 [or U5 small nuclear ribonucleoprotein, snRNP/U5-116kD]) (supplementary fig. S6, Supplementary Material online) (Atkinson 2015) confirmed that both aEF-2 and aEF-2p of Asgard archaea, Korarchaeota and Bathyarchaeota are part of a monophyletic clade, which includes canonical archaeal EF-2 homologs as well as all eukaryotic EF-2 family proteins (supplementary fig. S6, Supplementary Material online).

To improve the phylogenetic resolution, we subsequently analyzed a smaller set of archaeal and eukaryotic EF2 family proteins with (fig. 2) and without (supplementary fig. S7, Supplementary Material online) a bacterial outgroup. Phylogenetic analyses revealed that canonical aEF-2 homologs (as defined by conservation of the domain IV loop known to interact with the ribosomal decoding center during translocation) from all non-Heimdallarchaeote LC3 Asgard archaea and the Kor-1 and -3 marine Korarchaeota formed a highly supported clade (fig. 2; support 100/100). In contrast, their aEF2 paralogs comprise two separate clades, only one of which is highly supported. Notably, the phylogenetic placement of these protein clades relative to each other and within the phylogenetic backbone is not resolved due to lack of statistical support in most deeper nodes of the tree. For example, the placement of the two aEF-2p clades differs depending on whether or not a bacterial outgroup is included in the analysis (fig. 2 and supplementary fig. S7, Supplementary Material online). In part, 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 of members of the aEF-2p clades as well as in the node leading to aEF-2 (fig. 2 and supplementary files S2 and S3, Supplementary Material online).

Surprisingly, while bathyarchaeran 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 bathyarchaeran EF-2s as well as potential paralogs (i.e., RBG_13_46_16b and GB8-32-3). 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 are 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 Heimdallarchaeota LC3 clade (Heimdallarchaeota LC3 and Heimdallarchaeota B3) (fig. 2, support: 100/100).

Close inspection of the EF-2 sequence alignment revealed that eukaryotic and Heimdallarchaeota LC3 clade EF-2 homologs share common indels to the exclusion of all other archaeal EF-2 family protein sequences (supplementary figs. S8 and S9, Supplementary Material online). Notably, these highly conserved indels were found to be encoded by the genomic bins of two distantly related members of the Heimdallarchaeota LC3 lineage, which were independently assembled and binned from geographically distinct metagenomes (Spang et al. 2015; Zaremba-Niedzwiedzka et al. 2017). This refutes recently raised claims stating that these indels in Heimdallarchaeota LC3 may be the results of contamination from eukaryotes (Da Cunha et al. 2017) while supporting the sister-relationship of eukaryotes and Asgard archaea (Spang et al. 2015, 2018; Eme et al. 2017; Zaremba-Niedzwiedzka et al. 2017). The observed phylogenetic topology and the presence of the full complement of $dph$ biosynthesis genes in Heimdallarchaeota LC3 genomes (figs. 1 and 2), support an evolutionary scenario in which Heimdallarchaeota LC3 and eukaryotes share a common ancestry with EF-2 being vertically inherited from this archaeal ancestor.

**Discussion**

The use of metagenomic approaches has led to an expansion of genomic data from a large diversity of previously unknown archaeal and bacterial lineages and has changed our perception of the tree of life, microbial metabolic diversity and evolution, as well as the origin of eukaryotes (Brown et al. 2015; Castelle et al. 2015; Spang et al. 2015; Hug et al. 2016; Parks et al. 2017; Zaremba-Niedzwiedzka et al. 2017). Since most of what is known about archaeal informational processing machineries is based on a few model organisms, we aimed to use the expansion of genomic data to investigate key elements of the translational machinery—EF-2 and diphthamidylation—across the tree of life.

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

The evolution of archaeal diphthamide biosynthesis and EF-2 is especially intriguing in the context of eukaryogenesis. Recent findings based on comparative genomics...
indicate that eukaryotes evolved from a symbiosis between an alphaproteobacterium with an archaeal host
that shares a most recent common ancestor with extant members of the Asgard archaea, possibly a
Heimdallarchaeota-related lineage (Spang et al. 2015; Zaremba-Niedzwiedzka et al. 2017). Our study adds
additional data to support this scenario by revealing close sequence and predicted structural similarity of canonical
EF-2 proteins of the Heimdallarchaeote LC3 lineage and euarchaeal EF-2 proteins, including shared indels.
Furthermore, phylogenetic analyses of EF-2 family proteins reveals that EF-2 of the Heimdallarchaeote LC3 lin-
eage forms a monophyletic group with EF-2 family proteins of eukaryotes, and therefore suggests that the
archaeal ancestor of eukaryotes was equipped with an EF-2 protein similar to the homologs found in this lineage.
The subsequent evolution of the euarchaeal EF-2 family appears to have included at least two ancient duplication
events leading to Ria1 and Snu114. Importantly, the presence of characteristic eukaryotic indels in EF-2 of all mem-
bers 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. 2018).

In addition, the Heimdallarchaeote 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, euarchaeal
Dph5 appears to be most closely related to homologs of Woesearchaeota (supplementary fig. S10 and
file S3, Supplementary Material online), an archaeal lineage belonging to the proposed DPANN superphylum (Rinke et al.
2013; Castelle et al. 2015; Williams et al. 2017), comprising various additional lineages with putative symbiotic and/or parasitic members (reviewed in Spang et al. 2017). Notably, a previous study has also revealed an affiliation of some euarchaeal
rRNA 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 equi-
tans 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 (Iwabe et al. 1989; Baldauf
et al. 1996; Hashimoto and Hasegawa 1996; Elkins et al. 2008), and multiple-gene alignments of universal single
copy genes (Williams et al. 2012; Guy et al. 2014; Raymann et al. 2015; and others) to assess the relationships between
Archea, 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 data sets. 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 (Wu and Scott 2012;
Parks et al. 2015). 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 diph-
theramide was previously thought to be a conserved feature across Archaea and eukaryotes (Schafrath et al. 2014), and
critical for ensuring translational fidelity (Ortiz et al. 2006). Parabasalid parasites are known to infect the mucosal environ-
ments (e.g., the urogenital, digestive, and respiratory tracts) of different animals (for review see Maritz et al. 2014). Since
some of these environments in present-day animals contain diphtheria-producing bacteria (e.g., Corynebacterium
diphtheria; Human Microbiome Project 2012; Krishna et al. 2016) or toxin-encoding phages and prophages (Al-Jarbou
2012), it is possible that the ancestor of parabasalids was exposed to diphtheria toxin. Thus, loss of the diphthamideylation
machinery could have been selectively advantageous in the ancestral parabasalid to overcome the deleterious effects of
the diphtheria toxin. While we currently cannot rule out the possibility that dph-lacking archaea and parabasalids perform
the multistep process of diphthamideylation using a set of yet-unknown enzymes, future proteomics studies will be needed
to conclusively rule out the presence of diphthamide in these taxa. It is also possible that the EF2 of dph-lacking archaea and
parabasalids are subject to an alternate postranslational mod-
ification in domain IV, as has been shown for bacterial EF-P (Rajkovic et al. 2015). 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 diphthamide-deficient aEF-2 proteins are likely
under strong selective pressure to maintain translocase func-
tion. In contrast, analyses of the aEF-2 suggest that, while this paralog is a member of the translational GTPase super-
family, aEF-2 is unlikely to function in the same manner as canonical aEF-2. In fact, the complete lack of sequence con-
vention in aEF-2 paralogs indicates that

these paralogs are not likely to act as translases (fig. 3 and supplementary fig. S2a, Supplementary Material online) (Rodnina et al. 1997; Ortiz et al. 2006) and instead perform alternative roles. A similar lack of sequence conservation for key residues at the tips of the domain IV loops is seen in the bacterial EF-G paralog Tet. In that case, rather than participating in translocation, the paralog functions to dislocate tetracycline from the ribosome in a GDP-dependent manner (Donhofer et al. 2012).

It also seems possible that aEF-2p may compensate for the absence of diphthamide in at least some dph-lacking lineages. Eukaryotic EF-2 has recently been shown to function as a back-translocase (Susorov et al. 2018). Interestingly, this process was inhibited by ADP-ribosylation of eEF-2 diphthamide, and diphthamide may play a key role in back-translocation. While this remains to be explored further, these results support the hypothesis that aEF-2p could promote back-translocation in dph- archaea. Alternatively, given proposed regulation of translation via ADP-ribosylation of diphthamide (Schaffrath et al. 2014) and a role of diphthamide in responding to oxidative stress (Argüelles et al. 2013, 2014), aEF-2p could perform another, yet unknown role in translation regulation. No matter the true function or functions of aEF-2p, sequence homology suggests aEF-2p hydrolyzes GTP. This is in contrast to another duplication seen in an ancient paralog of aEF-2: based on sequence analysis, the translational GTPase superfamily member aSelBL appears to have lost the ability to hydrolyze GTP, and has a currently unknown function (Atkinson et al. 2011). Overall, sequence homology, predicted fold, and numerous evolutionarily analogous duplications within the larger superfamily all suggest aEF-2p is likely to interact with the ribosome in a GTP-dependent manner, but additional study is needed to determine its precise function.

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

**Supplementary Material**

Supplementary data are available at Genome Biology and Evolution online.

**Acknowledgments**

We thank Jordan Angle, Kay Stefanik, Rebecca Daly, and Kelly Wrighton for assistance with sampling of OWC sediments, and Felix Homa for computational support. Sequencing of OWC metagenomes was conducted in part by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility that is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. We thank Minh Bui Quang and Stephen Crotty for useful discussions. Sequencing of Aarhus bay metagenomes was performed by the National Genomics Infrastructure sequencing platforms at the Science for Life Laboratory at Uppsala University, a national infrastructure supported by the Swedish Research Council (VR-RFI) and the Knut and Alice Wallenberg Foundation. We thank the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) at Uppsala University and the Swedish National Infrastructure for Computing (SNIC) at the PDC Center for High-Performance Computing for providing computational resources. This work was supported by grants of the European Research Council (ERC Starting grant 310039-PUZZLE_CELL), the Swedish Foundation for Strategic Research (SSF-FFL5) and the Swedish Research Council (VR grant 2015-04959) to T.J.G.E. C.W.S. is supported by a European Molecular Biology Organization long-term fellowship (ALTF-997-2015) and the Natural Sciences and Engineering Research Council of Canada postdoctoral research fellowship (PDF-487174-2016).

**Literature Cited**


EF-2 and Diphthamide in Archaea and Eukaryotes


