Evolution of Elongation Factor G and the Origins of Mitochondrial and Chloroplast Forms

Gemma C. Atkinson*†,1 and Sandra L. Baldauf1

1Department of Systematic Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden
†Present address: University of Tartu, Centre of Biomedical Technology, Institute of Technology, Tartu, Estonia

*Corresponding author: E-mail: atkinson@ut.ee.
Associate editor: Martin Embley

Abstract

Protein synthesis elongation factor G (EF-G) is an essential protein with central roles in both the elongation and ribosome recycling phases of protein synthesis. Although EF-G evolution is predicted to be conservative, recent reports suggest otherwise. We have characterized EF-G in terms of its molecular phylogeny, genomic context, and patterns of amino acid substitution. We find that most bacteria carry a single “canonical” EF-G, which is phylogenetically conservative and encoded in an str operon. However, we also find a number of EF-G paralogs. These include a pair of EF-Gs that are mostly found together and in an eclectic subset of bacteria, specifically δ-proteobacteria, spirochaetes, and plionctomycetes (the “spd” bacteria). These spdEFGs have also given rise to the mitochondrial factors mtEFG1 and mtEFG2, which probably arrived in eukaryotes before the eukaryotic last common ancestor. Meanwhile, chloroplasts apparently use an α-proteobacterial–derived EF-G rather than the expected cyanobacterial form. The long-term comaintenance of the spd/mtEFGs may be related to their subfunctionalization for translocation and ribosome recycling. Consistent with this, patterns of sequence conservation and site-specific evolutionary rate shifts suggest that the faster evolving spd/mtEFG2 has lost translocation function, but surprisingly, the protein also shows little conservation of sites related to recycling activity. On the other hand, spd/mtEFG1, although more slowly evolving, shows signs of substantial remodeling. This is particularly extensive in the GTPase domain, including a highly conserved three amino acid insertion in switch I. We suggest that subfunctionalization of the spd/mtEFGs is not a simple case of specialization for subsets of original activities. Rather, the duplication allows the release of one paralog from the selective constraints imposed by dual functionality, thus allowing it to become more highly specialized. Thus, the potential for fine tuning afforded by subfunctionalization may explain the maintenance of EF-G paralogs.

Key words: EF-G, elongation factor G, organelle, xenology, paralogy, ribosome, translation.

Introduction

Elongation factor G (EF-G) is an ancient translational GTPase (trGTPase) and the bacterial homolog of eukaryotic eEF2 and archaeal aEF2. EF-G/eEF2 is a universal protein, one of at least four trGTPases that were present in the last universal common ancestor along with EF1, IF2, and possibly, SelB (Leipe et al. 2002; Margus et al. 2007). The EF2 family altogether consists of at least nine protein subfamilies of which five are bacterial (EF-G, LepA, RF3, Tet, TypA, and EFG-L), one is archaeal (aEF2), and three are eukaryotic (eEF2, Ria1p, and Snu114p) (Leipe et al. 2002; Atkinson and Baldauf 2009).

In all three domains of life, EF2 functions in the elongation stage of protein synthesis, promoting the translocation of peptidyl-transfer RNA (tRNA) from the A to the P site of the ribosome (Rodnina et al. 1997). In addition, bacterial EF-G has recently been shown to be an essential component of the ribosome recycling complex, which causes the two ribosomal subunits to dissociate. This final stage of protein synthesis requires EF-G, the ribosome recycling factor (RRF), and initiation factor 3 (IF3) (Hirokawa et al. 2005; Zavialov et al. 2005).

EF-G comprises five domains, the G (GTPase) domain and four additional domains referred to by number (II–V). The G domain is responsible for GTP binding and hydrolysis and is common to all GTPases, whereas domain II is common to all trGTPases, and domains III–V are EF2 family specific. Together, the latter three domains are proposed to mimic the structure of aminoacyl-tRNA (Nyborg et al. 1996) and to bind to the same region of the ribosome, the A site. EF-G domains III and IV also interact directly with RRF during ribosome recycling (Ito et al. 2002; Gao et al. 2007). EF-G promotes an interdomain rotation of RRF, which in turn forces the two ribosomal subunits apart (Gao et al. 2007).

EF-G is encoded by the gene fusA, which tends to occur in the highly conserved str operon. The operon classically consists of four protein-coding genes in the order 5′-rpsL-rpsG-fusA-tufA-3′, an organization found widely in bacteria and also in some archaea (Lathe et al. 2000). In addition, str is often followed immediately downstream by all or part of the ribosomal 16S operon. Together, these two operons are hypothesized to have formed an ancient “str–rpsL–rpsG–fusA–tufA–rpsJ–splC–3′” (Lathe et al. 2000).
Eukaryotes conduct cytoplasmic protein synthesis with eEF2, but most eukaryotic nuclear genomes also encode organelle-targeted EF-Gs that function in mitochondrial and/or plastid protein synthesis (mtEFG and cpEFG, respectively). Organelle targeting of these proteins has been demonstrated by their purification from mitochondrial and chloroplast cellular fractions, respectively (Grandi and Kuntzel 1970; Ciferri and Tiboni 1973), and by whole organelle proteome analyses of both mitochondria and chloroplasts (Sickmann et al. 2003; Heazlewood et al. 2007; Smith et al. 2007). Nonetheless, all organelar EF-Gs are nuclear encoded. This is presumably the result of endosymbiotic gene transfer early in eukaryotic evolution, as has occurred more recently for its 3′ str neighbor tufA in both plastids (Baldauf and Palmer 1990; Martin 2003; Timmis et al. 2004) and mitochondria (Lang et al. 1997).

Due to their antiquity, universality, and high levels of sequence conservation, various members of the EF-G protein family have been used to address important evolutionary questions. These include whether or not archaea are monophyletic (Cammarano et al. 1999) and the position ary questions. These include whether or not archaea are monophyletic (Cammarano et al. 1999) and the position of the root of the universal tree (Iwabe et al. 1989; Baldauf 1997). These gene families have been used to address important evolutionary questions: their antiquity, universality, and high levels of sequence conservation. Various members of the EF-G protein family have been used to address important evolutionary questions. These include whether or not archaea are monophyletic (Cammarano et al. 1999) and the position of the root of the universal tree (Iwabe et al. 1989; Baldauf 1997).

Our results confirm that most bacteria carry a single slowly evolving EF-G encoded in a canonical str operon. However, operon-free EF-G duplicates are also found, and these are scattered across the tree. All but two of these duplicates have very limited taxonomic distribution, and these two also have very similar taxonomic distributions, including most mitochondriate eukaryotes. The frequent cooccurrence of these two EF-Gs suggests that they are codpendent, most likely due to subfunctionalization, as shown experimentally for Borrelia and human EF-Gs (Tsuboi et al. 2009; Suematsu et al. 2010). We investigate the evolution of these two proteins in light of the known functional roles of EF-G.

**Methods**

**Data Set Assembly**

EF-G sequences were retrieved from NCBI using BlastP (Altschul et al. 1997) with an E-value cutoff of e−40. Bacterial sequences were identified by searching all completely sequenced genomes in the NCBI database. Due to the large number of these genome sequences, target genomes were limited to one representative for each genus in the order they are listed on the NCBI genomes table. Preliminary phylogenetic analyses identified two paralogous EF-G clades in spirochetes, planctomycetes, and δ-proteobacteria (referred to here as “spd” bacteria). Therefore, additional searches were conducted against all spd genomes in the NCBI genome and nr database. Bacterial small subunit ribosomal RNA (SSU rDNA) sequences were downloaded from the ribosomal database project (Cole et al. 2009).

Eukaryotic sequences were identified through searches against a taxonomically wide sampling of complete genome sequences (supplementary table S1, Supplementary Material online). Additional eukaryotic EF-G sequences were retrieved from BlastP searches against genomes from the following species: *Eimeria tenella* (GeneDB, Sanger Institute—http://www.genedb.org/); *Aureococcus anophagefferens*, *Chlamydomonas reinhardtii*, *Micromonas pusilla*, *Naeegleria gruberi*, *Phaeodactylum tricornutum*, *Thalessiosira pseudonana*, *Volvox carterii*, and *Emiliana huxleyi* (Department of Energy Joint Genome Institute—http://www.jgi.doe.gov/); *Cyanidioschyzon merolae* (University of Tokyo—http://merolae.biol.u-tokyo.ac.jp/); *Toxoplasma gondii* (ToxoDB—http://ToxoDB.org/); and *Tetrahymena thermophila* (NCBI nr database via trGTPbase: www.trGTPbase.org; Atkinson GC and Baldauf SL, unpublished data). *Phaeodactylum tricornutum* mtEFG1 was assembled from two overlapping gene models: e_gw1.5.154.1_Phtar2:11200 and e_gw1.5.161.1_Phtar2:11121 (http://www.jgi.doe.gov/). The 5′ ends of predicted mtEFG2 sequences from three species of Leishmania were completed by BlastN searches of the genome sequences to retrieve the full-length gene sequences, which were then translated with Transseq (http://www.ebi.ac.uk/Tools/).

The EF-G identity of all sequences was confirmed by scanning against a set of EF2 family profile hidden markov models (HMMs) generated from subsets of a curated trGTPase superfamily alignment (trGTPbase; Atkinson GC and Baldauf SL, unpublished). The resulting EF-G data set is referred to here as EFG_dset1. EFG_dset1 was aligned using MAFFT v6.234b with strategy L-INS-I (Katoh et al. 2005). Consensus sequences were then computed from the full alignment using the Python script ConsensusFinder (Atkinson GC and Baldauf SL, unpublished data), which indicates sites with conservation of a single amino acid in uppercase characters and conservation of a substitution group in lowercase. A 70% consensus sequence was constructed and used to identify well-conserved regions. These unambiguously aligned regions were confirmed and extracted using the Gblocks server v.0.91b (http://molevol.cnmima.csic.es/castresana/Gblocks_server.html; Talavera and...
Consensus versus Canonical EF-G for Various Classes of EF-Gs.

<table>
<thead>
<tr>
<th>Change in Numbers of Consensus Sites vs. Canonical EF-G</th>
<th>Gain</th>
<th>Change</th>
<th>Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>EF-G Domain</td>
<td>G II</td>
<td>III IV V</td>
<td>G II</td>
</tr>
<tr>
<td>spd/mtEFGs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>spdEFG1</td>
<td>36</td>
<td>14 10 9 6 4 0 1 24 16 18 7 8 12 2 6 7</td>
<td></td>
</tr>
<tr>
<td>mtEFG1</td>
<td>31</td>
<td>14 10 10 10 10 2 2 23 19 21 10 14</td>
<td></td>
</tr>
<tr>
<td>spd+mtEFG1</td>
<td>22</td>
<td>11 5 11 10 4 2 1 13 14 17 7 4 7</td>
<td></td>
</tr>
<tr>
<td>spdEFG2</td>
<td>11</td>
<td>1 1 1 1 0 1 0 0 1 0 1 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>mtEFG2</td>
<td>11</td>
<td>4 2 1 1 0 0 0 1 0 2 2 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>spd+mtEFG2</td>
<td>4</td>
<td>1 2 0 0 1 0 0 0 0 0 0 0 0 0 0 0</td>
<td></td>
</tr>
</tbody>
</table>

Note.—Consensus sequences were calculated at the 80% level, after excluding highly divergent sequences and are summarized by EF-G structural domains (Avarsion et al. 1994).

The Total Number of Gains, Losses, and Changes in Consensus versus Canonical EF-G for Various Classes of EF-Gs.

The dimensions of the final data sets are as follows: ConsensusFinder sequences calculated using the Python script ConsensusFinder (Atkinson GC and Baldauf SL, unpublished data, table 1). The data were divided into five classes based on the EFG_dset1 phylogeny (supplementary fig. S3, Supplementary Material online) excluding highly divergent sequences: “canonical EF-G,” “spdEFG1,” “mtEFG1,” “spdEFG2,” and “mtEFG2.” Gains are sites that meet the 80% conservation threshold within a class but not in canonical EF-G. Changes are sites that are conserved both in the class and canonical EF-G, but the identity of the conserved residue has changed. Losses are sites that are unconserved within a class but are conserved in canonical EF-G.

Site-Specific Sequence Conservation and Functional Diversification Analyses

Two methods were used for analyzing patterns of sequence evolution: 1) site-specific sequence conservation and 2) ML identification of sites that have experienced shifts in evolutionary rate as implemented in DIVERGE 2.0 (Gu 2001, 2006; Gu and Vander Velden 2002). To analyze site-specific conservation across subfamilies, the total number of gains, losses, and changes in consensus versus canonical EF-G were calculated for each major EF-G type using 80% strict consensus sequences calculated using the Python script ConsensusFinder (Atkinson GC and Baldauf SL, unpublished data, table 1).

Evolution of Chloroplast and Mitochondrial EF-G

During the topology search, evaluating the final tree under a $I'$ distribution of rates. Optimal tree searches consisted of 100 replicates, each starting from a maximum parsimony tree. Clade robustness was assessed with 100 bootstrap replicates, with the bootstrap percentage (mlBP) shown for each branch. Relative evolutionary rates for different EF-G types were computed with the relative rate test as implemented in RRTree (Robinson-Rechavi and Huchon 2000).

Genomic Context

The genomic context of EF-G–encoding genes (fus) was determined by retrieving the identity of up- and downstream flanking genes using the Python script ContextFinder.
entirely of sequences from an eclectic but similar subset of EF-G or simply EF-G. The other two subtrees consist almost completely sequenced genome. Eukaryotic sampling includes mitochondria-carrying eukaryotes. In addition, all representatives of every major bacterial taxon except Spirochetes (supplementary table S1 and fig. S3, Supplementary Material online). As examined except Spirochetes (supplementary table S1 and fig. S3, Supplementary Material online). Within the canonical EF-G subtree, the bacteria are also roughly grouped into recognized higher order taxa (Ciccarelli et al. 2006; Cole et al. 2009), albeit with varying levels of support (supplementary fig. S3, Supplementary Material online). As expected for a single gene or protein tree of bacteria, there is no confident resolution of any deeper branches (Delsuc et al. 2005). Canonical EF-G also has a slower evolutionary rate (0.37 substitutions/site) compared with both spd/mtEFG1 and spd/mtEFG2 (0.42 and 0.58 substitutions/site, respectively). DIVERGE 2.0 (Gu 2001, 2006; Gu and Vander Velden 2002) was used to calculate the coefficients of Type I functional divergence ($\theta_1$) among EF-G subtypes (supplementary fig. S5, Supplementary Material online). Type I divergent sites are those with significantly different amino acid substitution rates among subtypes. $\theta_1$ values of the genes was assigned by searching for keywords in gene name, title, and COG identifiers in the annotations of each record. Where identity could not be automatically assigned using annotations, the Entrez Gene record was checked by eye.

Transit Peptides

All putative mitochondrial and plastid sequences were analyzed using TargetP, which scans the N terminus for the predicted ability of sequences to form functional chloroplast, mitochondrial, and signal peptides (Emanuelsson et al. 2000). Apicomplexan nonmitochondrial-type sequences were also analyzed with PATS (predict apicoplast-targeted sequences) (Zuegge et al. 2001) to predict possible apicoplast transit peptides.

Results and Discussion

Molecular Phylogeny of EF-G

We have reconstructed the phylogeny of bacterial and organellar EF-G, which together form a monophyletic subgroup within the EF2 family (Leipe et al. 2002; Atkinson GC and Baldauf SL, in preparation). Altogether, 218 bacterial and 52 eukaryotic genomes were sampled including at least one representative from each genus of bacteria with a completely sequenced genome. Eukaryotic sampling includes representatives of five of the six recognized eukaryotic supergroups (Keeling et al. 2005) as little molecular data are available for Rhizaria. We find that EF-G is universal in bacteria and mitochondria-carrying eukaryotes. In addition, it has given rise to at least 12 paralogs and xenologs derived by gene duplication or HGT.

The molecular phylogeny of EF-G shows three main sections or subtrees (fig. 1 and supplementary fig. S3, Supplementary Material online). The largest subtree includes sequences from nearly every examined major group of bacteria. These are also usually the only EF-G in their host cells (supplementary table S1, Supplementary Material online), and they are almost exclusively found in str operons (fig. 1 and supplementary fig. S3, Supplementary Material online). Therefore, these are referred to here as canonical EF-G or simply EF-G. The other two subtrees consist almost entirely of sequences from an ecletic but similar subset of bacteria. These taxa include all 9 examined species of Spirochetes, 4 of the 5 examined species of Planctomycetes, and 14 of the 21 examined species of $\delta$-Proteobacteria. This diverse set of bacteria is herein referred to as the “spd” bacteria and their EF-G sequences as spdEFG1 and spdEFG2 (fig. 1).

Each of the spdEFG1 and spdEFG2 subtrees also includes one of the two mitochondrially targeted EF-Gs, mtEFG1 and mtEFG2, respectively (fig. 1 and supplementary fig. S3, Supplementary Material online). Together these account for all the putative or proven mitochondrial EF-Gs (supplementary table S1, Supplementary Material online). The monophyly of the spd/mtEFG1 clade is strongly supported by both ML and BI phylogenetic methods (1.0 bIPP, 100% bootstrap percentages [mlBP]; fig. 1). The clade is also supported by the presence of a small well-conserved insertion with the consensus sequence GVG, which is found exclusively and universally in all spd/mtEFG1 sequences (positions 47–49 in fig. 2, see below). The grouping of all spdEFG2 sequences and mtEFG2 is less well or consistently supported (1.0 bIPP, 65% mlBP; fig. 1), but these sequences also have a very similar taxonomic distribution to spd/mtEFG1 and are therefore referred to here as spd/mtEFG2. Nearly all species with spdEFG1 possess spdEFG2 and vice versa, and these are usually the only EF-Gs encoded in the genome (supplementary table S1, Supplementary Material online). Only two closely related bacteria were found with all three main EF-G types (Anae-romyxbacter sp. and A. halogenans, supplementary table S1 and fig. S3, Supplementary Material online).

Canonical EF-G is by far the most widespread of the three EF-G types. It is found in at least some and, in most cases, all representatives of every major bacterial taxon examined except Spirochetes (supplementary table S1 and fig. S3, Supplementary Material online). Within the canonical EF-G subtree, the bacteria are also roughly grouped into recognized higher order taxa (Ciccarelli et al. 2006; Cole et al. 2009), albeit with varying levels of support (supplementary fig. S3, Supplementary Material online). As expected for a single gene or protein tree of bacteria, there is no confident resolution of any deeper branches (Delsuc et al. 2005). Canonical EF-G also has a slower evolutionary rate (0.37 substitutions/site) compared with both spd/mtEFG1 and spd/mtEFG2 (0.42 and 0.58 substitutions/site, respectively). DIVERGE 2.0 (Gu 2001, 2006; Gu and Vander Velden 2002) was used to calculate the coefficients of Type I functional divergence ($\theta_1$) among EF-G subtypes (supplementary fig. S5, Supplementary Material online). Type I divergent sites are those with significantly different amino acid substitution rates among subtypes. $\theta_1$ values

<table>
<thead>
<tr>
<th>Pairwise Combinations</th>
<th>$\theta_1$</th>
<th>$\theta_1$ SE</th>
<th>Total sites where PP &gt; 0.90</th>
<th>G</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>spdEFG1 vs. spdEFG2</td>
<td>0.32</td>
<td>±0.03</td>
<td>43</td>
<td>9</td>
<td>7</td>
<td>4</td>
<td>18</td>
<td>5</td>
</tr>
<tr>
<td>spdEFG1 vs. canEFG</td>
<td>0.30</td>
<td>±0.03</td>
<td>47</td>
<td>14</td>
<td>8</td>
<td>10</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>spdEFG2 vs. canEFG</td>
<td>0.26</td>
<td>±0.03</td>
<td>38</td>
<td>10</td>
<td>5</td>
<td>3</td>
<td>18</td>
<td>2</td>
</tr>
</tbody>
</table>

Note.—$\theta_1$ is the coefficient of functional divergence, and SE is standard error. PP is the Posterior probability of a site being functionally divergent, with sites >0.90 indicating a strong probability of divergence. The number of these sites in total and within each domain are shown.
were significantly greater than 0 and in the same range for all pairwise comparisons (0.26–0.32 with a standard error of 0.03; table 2), indicating significant differences in substitution rates among all the three groups.

In contrast to canonical EF-G, the two spd/mtEFG clades include sequences from only a small subset of bacteria (fig. 1). The eclectic taxonomic distribution of the spdEFGs suggests that the encoding genes have been horizontally transferred.
possibly multiple times. HGT of spdEFG1, between spirochaetes and \( \delta \)-proteobacteria, is also supported by the shared presence of an adjacent downstream gene in *Treponema denticola* and several \( \delta \)-proteobacteria (fig. 1 and supplementary fig. S1, Supplementary Material online). The two spd/mtEFG clades do not cluster together as a monophyletic group in the tree (fig. 1), suggesting that they were independently transferred. However, intervening branches between the two clades are few and lack strong support (fig. 1). Additionally, the fact that spdEFG1 and spdEFG2 cooccur in three very different groups of bacteria—spirochaetes, \( \delta \)-proteobacteria, and planctomycetes—is most parsimoniously explained by cotransfer. The simplest explanation for this would be that the spdEFG-encoding genes first arose by tandem duplication and that they remained adjacent in the genome long enough to be cotransferred several times. However, these genes are not found at adjacent locations in any examined bacterium. Thus, the co-HGT of these genes, if it did occur, would probably have been largely limited to a narrow window of time when the duplicates were proximal in the genome.

**Origin of Organelle EF-Gs**

No \( \alpha \)-proteobacterium was found to encode an spdEFG, and there is no apparent affinity of mtEFG for \( \alpha \)-proteobacteria in the EF-G tree (fig. 1). This in itself is not too unusual as only ~20% of nuclear-encoded mitochondrial proteins show an \( \alpha \)-proteobacterial affinity (Kurland and Andersson 2000; Brindefalk et al. 2007). Instead, the mtEFGs appear to have arisen from spd bacterial EF-Gs. The fact that both mtEFGs are present in all five of the examined lineages of mitochondriate eukaryotes (Opisthokonts, Amoebozoa, Archaeplastida, Chromalveolates, and Discobids) indicates that these EF-Gs probably arrived in eukaryotes before the eukaryotic last common ancestor. In fact, the simplest scenario is that they were present in the progenitor of the mitochondrion, and thus, the dual nature of mtEFG is an inherited rather than an eukaryote-derived feature. The link between mtEFG and mitochondrial protein synthesis is further strengthened by the fact that the only eukaryotes lacking mtEFGs are taxa that lack respiring mitochondria, none of which encode either protein. This includes *Giardia intestinalis*, *Entamoeba histolytica*, *Trichomonas vaginalis*, and the mitochondrial DNA–lacking *Cryptosporidium parvum* and *C. hominis* (Embley and Martin 2006). Because these species represent three widely separated eukaryotic supergroups (Excavates, Opisthokonts, and Alveolates), this also represents at least three parallel independent losses of mitochondrial respiration and mtEFGs (Embley and Martin 2006).
With the single exception of the obligate fungal parasite, *Cryptococcus neoformans*, the only aerobic eukaryotes with only one mtEF-G are the plastid/apicoplast-carrying eukaryotes: Archaeoplastida, stramenopile algae, and Apicomplexa (supplementary table S1, Supplementary Material online). Instead of a second mtEF-G, these species carry a plastid/apicoplast-targeted EF-G, termed cpEFG or apiEFG (fig. 1 and supplementary fig. S3, Supplementary Material online). Unlike the mtEFGs, cpEFG is a canonical-type EF-G, whereas the apicoplast-targeted apiEFGs are too divergent to classify (fig. 1 and supplementary fig. S3 and table S2). The latter is a common feature in apicoplast proteins (Blanchard and Hicks 1999).

All cpEFGs form a single tight clade in the canonical section of the EF-G tree with full support. However, unusual for cpEFGs, this clade shows a moderate to strong affinity for α-proteobacterial EF-G (76% mlBP, 0.99 biPP; fig. 1), whereas most nuclear-encoded proteins of bacterial ancestry that are targeted to the chloroplast show cyanobacterial ancestry (Reyes-Prieto et al. 2010). This makes it tempting to speculate that cpEFG could have originated with the α-proteobacterial progenitor of mitochondria. This would require either that the mitochondrial progenitor had all three EF-G types, an extremely rare situation among the bacteria examined here (supplementary table S1, Supplementary Material online), or that the original canonical mtEF-G was replaced later by an spdEFG pair acquired from another bacterial source or sources. However, both mtEFGs appear to have been present in the last common ancestor of eukaryotes, whereas plastids appear to have arisen later, early in the evolution of Archaeplastida (Archibald 2009). Thus, if cpEFG has been resident in eukaryotes as long as mitochondria, it must have been maintained under strong selective pressure until the advent of eukaryotic photosynthesis at which point it suddenly became dispensable in all nonphotosynthetic eukaryotes: Archaeplastida, stramenopile algae, and Apicomplexa (supplementary table S1, Supplementary Material online), or that the original canonical mtEF-G was replaced later by an spdEFG pair acquired from another bacterial source or sources. However, both mtEFGs appear to have been present in the last common ancestor of eukaryotes, whereas plastids appear to have arisen later, early in the evolution of Archaeplastida (Archibald 2009). Thus, if cpEFG has been resident in eukaryotes as long as mitochondria, it must have been maintained under strong selective pressure until the advent of eukaryotic photosynthesis at which point it suddenly became dispensable in all nonphotosynthetic lines. Alternatively, in a simpler explanation, the progenitor of the chloroplast could have had an α-proteobacterial EF-G acquired by HGT and xenologous replacement.

If the two spd/mtEFGs have complementary functions, as the data suggest (see below), then the fact that plastid-carrying eukaryotes have only spd/mtEFG1 suggests that cpEFG is filling in for the function of spd/mtEFG2. If this is the case, then cpEFG must be dually targeted to the chloroplast and the mitochondrion. Most of the cpEFGs appear to have amino-terminal extensions that can be confidently predicted to encode chloroplast-targeting (transit peptide) sequences (supplementary table S2, Supplementary Material online), and the chloroplast targeting of cpEFG has been demonstrated by plastid proteome analyses in *Arabidopsis thaliana* (Heazlewood et al. 2007). However, dual targeting of plastid and apicoplast proteins to the mitochondrion is known to occur, including other components of the translation machinery, aminoacyl-tRNA synthetases (Duchene et al. 2005; Pino et al. 2007).

**Additional EF-G Paralogs and Xenologs**

In addition to spdEFG, there are a number of other paired EF-Gs in the canonical EF-G subtree. In all these cases, the shorter branched EF-G is str operon encoded, and its phylogeny (supplementary fig. S3, Supplementary Material online) roughly follows the species phylogeny as predicted by small SSU rRNA (supplementary fig. S4, Supplementary Material online). Meanwhile, in all cases, the longer branched EF-G form is not str encoded and is often found elsewhere in the tree, suggesting that these are xenologs, that is, they arose by HGT. However, there are two examples where the two EF-Gs group together, which may indicate in-paralogs or very recent HGT from a close relative (supplementary fig. S3, Supplementary Material online).

**Alpha-Proteobacterial Subtree**

A mixed grouping of non-str-encoded EF-Gs from subsets of γ-proteobacteria and cyanobacteria is embedded in the α-proteobacterial subtree (gcEFG; fig. 1 and supplementary fig. S3, Supplementary Material online) forming a strongly supported clade with a subset of α-proteobacteria. This suggests at least two horizontal transfers of gcEFG, one at its origin and a second transfer some time later. The γ-proteobacteria in the gcEFG group are from at least three different divisions of γ-proteobacteria (supplementary figs. S3 and S4, Supplementary Material online), suggesting additional possible interphylum transfers or early HGT followed by multiple independent losses. In addition, the EF-Gs in the α-proteobacterial sister group to gcEFG are non-str encoded and group with *Hahella chejuensis* EF-G2 (supplementary fig. S3, Supplementary Material online), which suggests that the α-proteobacterial sequences may themselves have been derived by HGT.

**Beta-Proteobacterial Subtree**

Three unusual groupings are found embedded in the β-proteobacterial subtree (fig. 1 and supplementary fig. S3, Supplementary Material online). The bgEFG1 clade consists of non-str-encoded EF-Gs from two γ-proteobacteria. This suggests simple HGT, although these two γ-proteobacteria are not sister taxa in their home clade in either the EF-G or rDNA trees (supplementary figs. S3 and S4, Supplementary Material online, respectively), suggesting more than one HGT event. The bgEFG1 clade consists of non-str-encoded EF-Gs from four related β-proteobacteria, which means that these could be in-paralogs. However, the branching order among the bgEFG2 sequences does not entirely match that of the EF-G or rDNA species trees (supplementary figs. S3 and S4, Supplementary Material online, respectively), suggesting some additional HGT. Finally, the bgEFG2 clade consists of two γ-proteobacteria EF-Gs, indicating horizontal transfer, as previously reported (Brochier et al. 2002). Interestingly, these sequences are also str-encoded EF-Gs, which would indicate HGT followed by xenologous replacement as with ribosomal protein L29 (Omelchenko et al. 2003).

**Other Groups**

The γ-proteobacterium *Pseudomonas aeruginosa* has both an str-encoded EF-G and a longer branched non-str-encoded EF-G that group together. This looks like in-paralogs, although there is substantial distance between the two
sequences, so it could be HGT between closely related genomes that have not yet been sequenced (100% BP; fig. 1 and supplementary fig. S3, Supplementary Material online). Likewise, *Syntrophomonas wolfei* (home clade Synergistes) also carries both an str- and non-str-encoded EF-G. These both have very long branches and only group together weakly (25% BP; supplementary fig. S3, Supplementary Material online), although again the non-str-EF-G is more divergent. Finally, two actinobacteria, *Streptomyces coelicolor* and *St. avermitilis*, have str-encoded EF-Gs in their home clade (fig. 2) plus non-str-encoded EF-Gs that form a separate strongly-supported clade (strepeFG, 100% BP; supplementary fig. S3, Supplementary Material online) lying close to spdEFG2 (51% BP, supplementary fig. S3, Supplementary Material online). These extremely divergent sequences may be in-paralogs that are misplaced in the tree due to long-branch attraction to some of the extremely long spdEFG branches. No second copy EF-Gs were found in any of the other major bacterial groups represented here, but there is also very limited taxonomic sampling from most of these except Firmicutes.

**Other Possible HGTs**

Other than the above listed cases, there is general agreement between the tRNA tree (supplementary fig. S4, Supplementary Material online) and canonical EF-G subtree (supplementary fig. S3, Supplementary Material online). That is, all major groups and some subgroups are reconstructed similarly in both trees. However, many branches lack strong statistical support (BP > 70%; Hillis and Bull 1993), which makes inconsistencies difficult to detect with confidence. Nonetheless, there are a number of nontrivial differences between these trees. For example, *Maricaulis maris* and *Caulobacter crescentus* (α-proteobacteria) occupy quite different positions in the EF-G and SSU trees (supplementary figs. S3 and S4, Supplementary Material online, respectively), including a number of strongly supported intervening branches. However, all the solo EF-Gs are str encoded (fig. 1), meaning that HGT would need to have been followed by xenologous replacements. This would lead to strong inconsistencies between the EF-G and EF-Tu phylogenies of the same species, which so far we do not see (Atkinson GC and Baldauf SL, unpublished observations).

**Operon Dynamics**

Nearly all canonical EF-Gs are encoded in an str operon (5′-rpsL-rpsG-fusA-tufA-3′; fig. 2 and supplementary fig. S3, Supplementary Material online) often including a downstream tufS and sometimes even rpIC, thus forming a complete str über-operon (Lathe et al. 2000). Meanwhile, most spdEFG-encoding genes (referred to here as fusS1 and fusS2) show no signs of flanking str operon structure. There are two exceptions: The fusS1 of *Treponema denticola* and the fusS2 of Planctomycetes are both found in nearly canonical str operons missing only the downstream tufA, in the case of *Blastopirellula marina*, including even the farther downstream tufS (fig. 1). This presence of str operons deeply embedded in operon-free clades suggests three possible scenarios. First, the fusS copies originated from a whole operon duplication that was retained through much of spd evolution and then lost independently in multiple lineages. This would require a minimum of four operon losses for spdEFG1 and two for spdEFG2 (fig. 1). Second, “de novo” operon assembly may have occurred in one or both spdEFG lineages and been maintained through selection. Although this seems unlikely based on pure chance, “de novo” assembly would be simplified by the fact that rpsL and rpsG are already linked in most spd species (data not shown). Third, the sudden appearance of operon structure in both spd lineages may be due to multiple independent HGTs of nearly complete str operons into a close ancestor of the Spirochaetes and a common ancestor of the Planctomycetes. This would have been followed by xenologous replacement whereby the incoming str-encoded fusA was replaced by a resident fusS gene. This should lead to anomalies in the rpsL and rpsG phylogenies, but unfortunately, these genes are too small for reliable phylogenetic analysis. Xenologous replacement has been previously demonstrated for at least seven ribosomal proteins (Brochier et al. 2000; Makarova et al. 2001; Omelchenko et al. 2003). At present, there are not enough data to support one of these scenarios over the others; however, this may be resolved in the future with further genome sequencing of spd-type organisms.

**Functional Evolution of EF-G Types**

Examination of the distribution of EF-G types among bacteria shows that the two spdEFGs never occur alone. In most cases, they cooccur and are the only EF-Gs in the cell (supplementary table S1, Supplementary Material online). Four bacteria were found with only spdEFG1 and six with only spdEFG2 (supplementary table S1, Supplementary Material online). However, all these carry a second canonical-type EF-G, except for *Leptospira spp*. Where the second EF-G, while still clearly an EF-G based on HMM screening, is so divergent that its phylogenetic position could not be resolved (data not shown). Likewise, all mitochondriate eukaryotes have either both mtEFGs or one mtEFG and a cpEFG and an apIFG. The single exception to this is the fungal parasite, *Cryptococcus neoformans* (supplementary table S1, Supplementary Material online).

The long-term and widespread comaintenance of spd/mtEFG suggests that these are performing complementary essential functions. The fact that either can be substituted for by canonical EF-G suggests that these complementary functions are subsets of canonical EF-G function. Recent evidence indicates that EF-G has distinct roles in both translocation and ribosome recycling. These activities occur at different times in the translation cycle, involve different interacting partners, and reside at least partly in different domains of the protein. Ribosomal recycling has especially been shown by Cryo-EM mapping to mainly involve interaction of EF-G domains III and IV with RRF (*Itto et al. 2002; Gao et al. 2007*). This is also supported by domain swapping experiments on human mtEFGs (*Tsuboi...*
In fact, human mtEFG1 and 2 have been shown to be specific for translocation and recycling, respectively (Tsuboi et al. 2009), as have spdEFG1 and 2, respectively, of *Borrelia burgdorferi* (Suematsu et al. 2010).

If all spd/mtEFGs have indeed become specialized for these two functions, then we would expect to see different patterns of sequence conservation in the two EF-G types consistent with these two roles. Specifically, we expect to see stronger conservation of ribosome recycling-specific sites in spd/mtEFG2 and stronger conservation of translocation-specific sites in spd/mtEFG1. We therefore analyzed patterns of sequence conservation in the two spd/mtEFGs relative to canonical EF-G by comparing consensus sequences and site-specific evolutionary rates (Gu 2001, mtEFGs relative to canonical EF-G by comparing consensus patterns of sequence conservation in the two spd/mtEFG2 and stronger conservation of patterns of sequence conservation in the two EF-G (Suematsu et al. 2010).

Evolution of Chloroplast and Mitochondrial EF-G

In terms of overall sequence conservation, the different EF-G forms also show quite different patterns (table 1). Compared with canonical EF-G, spd/mtEFG1 has lost a total of 87–88 consensus sites but gained 79–88 new ones (table 1; yellow highlighting in fig. 2). In contrast, spd/mtEFG2 has lost 147–162 consensus sites relative to canonical EF-G but added only seven new ones (table 1; blue highlighting in fig. 2). In addition, spd/mtEFG1 has also “remodeled” 16 of the consensus sites shared with canonical EF-G, that is, these sites are still strongly conserved in spd/mtEFG1 but with a new consensus residue, often with very different chemical characteristics (table 1; green highlighting in fig. 2). Meanwhile, spd/mtEFG2 has only remodeled one canonical EF-G consensus site. Thus, it appears that spd/mtEFG1 has been substantially remodeled relative to canonical EF-G, possibly becoming more specialized or modifying its preexisting role in translocation. Meanwhile, spd/mtEFG2 has mostly just deteriorated. Although the latter is consistent with subfunctionalization, the actual pattern of conservation (fig. 2) does not appear to be consistent with retention of RRF interaction (Gao et al. 2009). Sites that DIVERGE 2.0 (Gu 2001, 2006; Gu and Vander Velden 2002) predicts at the 0.90 PP level to have experienced a shift in substitution rates (Type I divergence) between spd/mtEFG1 and spd/mtEFG2 are indicated on figure 2. Similar to the consensus comparison results, the sites with a shift in substitution rate show no correlation with predicted RRF interaction sites. Additionally, DIVERGE failed to identify any substitutions that both shift physiochemical properties and are completely fixed among all members of the subgroups (Type II functionally divergent [θII] sites; θII ML <0).

The lack of signal of subfunctionalization at RRF-interacting sites seems inconsistent with the fact that spdEFG1 and 2 have been shown to be subfunctionalized for translocation and recycling, respectively, in *Borrelia* (Tsuboi et al. 2009) as have mtEFG1 and 2 in human mitochondria (Tsuboi et al. 2009; Suematsu et al. 2010). There are various possible explanations for the lack of RRF interacting site conservation in spd/mtEFG2. 1) It is possible that only some spd/mtEFG2s are subfunctionalized for ribosome recycling. This would require another explanation for the long-term coexistence of both spd/mtEFGs. 2) There could be balancing mutations in RRF that compensate for spdEFG2 mutations. However, inspection of an RRF alignment shows the same pattern of sequence conservation at EF-G-interacting sites in canonical and spdEFG2-carrying bacteria (supplementary fig. S4, Supplementary Material online). 3) EF-G:RRF interactions may be highly species specific and therefore rapidly evolving. This is supported by the observation that mutational defects in one RRF:EF-G intermolecular interface can be compensated for by gain-of-function mutations at another interface (Ito et al. 2002). The latter also suggests that ribosome recycling may not require highly specific interactions between EF-G and RRF but rather simply the maintenance of an EF-G-like structure with ribosome-binding capabilities and appropriate surface charges. This could be sufficient to cause the necessary repulsion of the RRF head domain (Gao et al. 2007). Indeed, many of the sites that are strongly conserved
in spd/mtEFG2 are internal to the structure and responsible for forming networks of interactions for structural integrity (fig. 2).

On the other hand, there is strong evidence of loss of translocation function by spd/mtEFG2. The greatest concentration of sites with substitution rate shifts between spd/mtEFG1 and spd/mtEFG2 is around the tRNA-interacting loops (positions 500–520 and 579–590; fig. 2) located at the very tip of the domain IV structure (Gao et al. 2009). These sites are highly conserved in spd/mtEFG1, but degraded in spd/mtEFG2, consistent with loss of translocation function by the latter. However, the region of EF-G (positions 459–466; fig. 2) that interacts with the sarcin–ricin loop (SRL) of tRNA is 100% conserved in all three EF-G subtypes. The SRL has been proposed to be critical for EF-G and EF-Tu discrimination and EF-G*GTP binding to the ribosome A site (Sergiev et al. 2005; Garcia-Ortega et al. 2010). This suggests that specific interactions on the ribosome may differ among subtypes, but important structural features required for ribosome binding are preserved.

Thus, the key to the subfunctionalization of EF-G may lie with spd/mtEFG1. Unlike spd/mtEFG2, the latter shows substantial remodeling relative to canonical EF-G (fig. 2; table 1). Much of this remodeling appears to be in the GTPase domain, where 13 canonical consensus sites are lost, 22 new ones are gained, and 7 are conserved but with a new residue (vs. 4–17, 4–11, and 0–4 for the other four domains, respectively, table 1). This domain is also the location of the three amino acid insertion unique to spd/mtEFG1 (fig. 2). Together, these G domain modifications suggest that spd/mtEFG1 evolution has involved changes in how it interacts with and/or is regulated by nucleotides. The three amino acid insertion is particularly interesting as it occurs in the “switch I” region of the G domain (fig. 2). Switch I is a flexible structural element between the G1 and G2 GTP-binding loops, responsible for accommodating of the nucleotide in the G domain. Its length varies among trGTPase families, reflecting family-specific differences in nucleotide interactions (Hauryliuk et al. 2009). In EF-G’s GDP-bound and nucleotide-free states, switch I is disordered in structure, but on binding GTP, it becomes structured by locking onto the gamma phosphate of GTP (Connell et al. 2007; Hauryliuk et al. 2008; Gao et al. 2009; Ticu et al. 2009). After GTP hydrolysis and peptidyl-tRNA translocation, the switch changes conformation again, flipping out from the ribosome and promoting release of GDP and EF-G from the ribosome (Connell et al. 2007; Hauryliuk et al. 2008; Gao et al. 2009; Ticu et al. 2009).

In this respect, it is interesting to note that human mtEFG1 (Bhargava et al. 2004) and B. burgdorferi spdEFG1 (Suematsu et al. 2010) are resistant to fusidic acid (FA), an antibiotic that inhibits release of EF-G from the ribosome, thus blocking translation. FA binding appears to require an open disordered conformation of the switch I loop (Gao et al. 2009). Thus, the GVG insertion could change the structure of the open conformation of switch I and make spd/mtEFG1 inaccessible to FA. Interestingly, spd/mtEFG2 is also resistant to FA, which also inhibits ribosome recycling by E. coli EF-G (Savelsbergh et al. 2009; Suematsu et al. 2010). The cause here might also be switch I, which has also experienced multiple substitutions in spd/mtEFG2 relative to canonical EF-G (fig. 2). However, unlike spdEFG1, this is more of a loss of conservation rather than a strongly differently conserved motif. Therefore, FA resistance of spdEFG2 might be a lineage-specific phenomenon.

Our analyses of site-specific conservation and substitution rates clearly show retention and loss of translocation functions in spd/mtEFG1 and spd/mtEFG2, respectively, consistent with experimental data on subfunctionalization (Tsuboi et al. 2009; Suematsu et al. 2010). However, this subfunctionalization appears asymmetric as we find surprisingly little signal in the primary sequence of spd/mtEFG2 for specialization for ribosome recycling. As spdEFG2 is in fact most strongly conserved in structurally important regions, the advantage of maintaining spdEFG2 may have been as a structural mimic of EF-G, capable of binding the ribosome and promoting recycling primarily through surface charge repulsion. This would have released spdEFG1 from its recycling responsibilities, allowing it to become more specialized for its role in translocation. Substantial remodeling of the G domain suggests that this specialization involved modifying or refining its interactions with nucleotides. The spdEFG pair subsequently entered eukaryotes already subfunctionalized in a frozen accident that survived both mtEFG genes being transferred from the mitochondrial genome to the nucleus.

Supplementary Material

Supplementary tables 1–2 and figs. 1–5 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We wish to thank S. Andersson, V. Hauryliuk, T. Margus, T. Tenson, O. Fiz-Palacio, and T. Ettema for helpful discussions and critical reading of the manuscript. We also wish to thank the Department of Energy Joint Genome Institute for the use of unpublished sequence data. This work was primarily supported by a Westin Foundation Grant (S.L.B. and G.C.A.) and VR grant 70495101 (S.L.B.), with additional funds from Estonian Science Foundation Mobilitas grant MJ99 (G.C.A.), and the European Regional Development Fund through the Center of Excellence in Chemical Biology (G.C.A.).

References

evolution of bacterial ribosomal proteins. *Genome Biol.* 2:RESEARCH 0033


