MBE Advance Access published July 19, 2016

Lokiarchaeota Marks the Transition between the Archaeal and Eukaryotic Selenocysteine Encoding Systems

Marco Mariotti,[1,2,3] Alexei V. Lobanov,[1] Bruno Manta,[1] Didac Santesmasses,[2,3] Andreu Bofill,[2,3] Roderic Guigó,[2,3] Toni Gabaldón,[2,3,4] and Vadim N. Gladyshev*,[1]

[1]Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA

[2]Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain

[3]Universitat Pompeu Fabra (UPF); and Institut Hospital del Mar d'Investigacions Mèdiques (IMIM), Barcelona, Spain

[4]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

***Corresponding author:** E-mail: vgladyshev@rics.bwh.harvard.edu.

Associate Editor: Michael Purugganan

Abstract

Selenocysteine (Sec) is the 21st amino acid in the genetic code, inserted in response to UGA codons with the help of RNA structures, the SEC Insertion Sequence (SECIS) elements. The three domains of life feature distinct strategies for Sec insertion in proteins and its utilization. While bacteria and archaea possess similar sets of selenoproteins, Sec biosynthesis is more similar among archaea and eukaryotes. However, SECIS elements are completely different in the three domains of life. Here, we analyze the archaeon *Lokiarchaeota* that resolves the relationships among Sec insertion systems. This organism has selenoproteins representing five protein families, three of which have multiple Sec residues. Remarkably, these archaeal selenoprotein genes possess conserved RNA structures that strongly resemble the eukaryotic SECIS element, including key eukaryotic protein-binding sites. These structures also share similarity with the SECIS element in archaeal selenoprotein *VhuD*, suggesting a relation of direct descent. These results identify *Lokiarchaeota* as an intermediate form between the archaeal and eukaryotic Sec-encoding systems and clarify the evolution of the Sec insertion system.

Key words: **Selenocysteine, selenoprotein, SECIS, Lokiarchaeota, archaea, evolution.**

Introduction

Selenoproteins Across the Domains of Life

Selenoproteins are a rare class of proteins that contain a selenocysteine (Sec) residue, often referred to as the 21st amino acid (Labunskyy et al. 2014). Sec is inserted co-translationally, like standard amino acids, and possesses its own tRNA (*tRNAsec*). However, Sec does not have a fully dedicated codon in the genetic code. Instead, Sec residues are inserted in response to UGA codons that are recoded in the presence of specific Sec designation signals; UGAs are otherwise interpreted as stop signals in most organisms. In response to those signals, a Sec-specific elongation factor (*EFsec*) replaces the standard *EF-Tu* uniquely for the translation of Sec UGA codons. *EFsec* recruits the *Sec-tRNA*, promoting the specific insertion of Sec residues at these locations.

Most selenoproteins are enzymes with oxidoreductase function, with Sec being the catalytic redox active site. For the great majority of selenoproteins, standard homologues (orthologues and/or paralogues) that replace Sec with cysteine (Cys) exist, and are known to perform essentially the same molecular function, albeit less efficiently (Fomenko and Gladyshev 2012). Although the exchangeability of Sec and Cys is debated (Gromer et al. 2003; Castellano 2009; Hondal and Ruggles 2011; Hondal et al. 2013), it is generally accepted that Sec is used instead of Cys for reasons related to its higher reactivity, which leads to improved catalytic efficiency or resistance to inactivation in redox reactions. Many selenoproteins are enzymes involved in redox homeostasis, including some that are essential for human, mouse, and other vertebrates (Labunskyy et al. 2014).

Selenoproteins require a specific set of genes dedicated to Sec synthesis and insertion (here denoted as the "Sec machinery"), including *EFsec* and *tRNAsec*. However, selenoproteins, along with the capacity to code for Sec (the "Sec trait"), are not present in all organisms. Their distribution is scattered, but encompasses lineages of the three domains of life: bacteria, archaea, and eukaryotes (Mariotti et al. 2015). Selenoprotein families are quite diverse, with little overlap between prokaryotic and eukaryotic selenoproteomes (Driscoll and Chavatte 2004). Bacteria utilize selenoproteins to carry out functions such as redox homeostasis, electron transport/energy metabolism, compound detoxification, and oxidative protein folding. In contrast, the archaeal selenoproteins with known functions are involved in hydrogenotrophic methanogenesis, with the only exception of selenophosphate synthetase (*SPS*), involved in Sec biosynthesis (Stock and Rother 2009). Eukaryotic selenoproteins, on the other hand,

© The Author 2016. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Open Access

Downloaded from http://mbe.oxfordjournals.org/ at Biblioteca de la Universitat Pompeu Fabra on July 22, 2016

Article

carry out very diverse functions (Lobanov et al. 2009). In comparison to prokaryotes, eukaryotic selenoproteomes are largely expanded in families involved in redox regulation, antioxidant defense, protein repair, and oxidative protein folding, while they are decreased or depleted in families involved in compound detoxification, electron transport, and energy metabolism (Labunskyy et al. 2014).

The mechanisms of Sec synthesis and insertion also exhibit differences in the three domains of life (fig. 1). However, while archaea and bacteria share a larger number of selenoprotein families than with eukaryotes, Sec biosynthesis is more similar between archaea and eukaryotes. Organisms in these two lineages first catalyze the synthesis of phospho-Ser with the protein *PSTK*, and then convert it to Sec, whereas bacteria directly synthesize Sec from Ser.

SEC Insertion Sequence (SECIS) Elements

The main signals for Sec insertion are RNA structures found in the selenoprotein transcripts, designated SECIS elements (Berry et al. 1991). Strikingly, SECIS elements do not share any obvious resemblance in sequence or structure between the three domains of life (Krol 2002). In bacteria, the SECIS element (bSECIS) is a stem–loop structure located within the coding sequence, immediately downstream of the Sec UGA site it acts upon (fig. 1*B*) (Hüttenhofer et al. 1996). These structures show considerable variability across genes and species, but always feature at least one stem and a small loop, with one or more G in the first loop positions (Zhang and Gladyshev 2005). The recognition of bSECIS elements is carried out by a specialized domain of the bacterial *EFsec* (*SelB*) (Yoshizawa and Böck 2009).

The eukaryotic SECIS elements are located in the 3′UTR of selenoprotein transcripts. The distance between the SECIS and the Sec UGA codon varies substantially in vertebrates (∼0.2/5.2 kB—Mariotti et al. 2012). Eukaryotic SECISes contain an RNA motif called kink-turn (Latrèche et al. 2009), characterized by an unusual GA-GA antiparallel pair, and surrounded by two stems. An additional stem is found in the apical loop of certain SECIS elements, known as "type II" (Grundner-Culemann et al. 1999). Eukaryotic SECISes show a few other conserved features besides the invariant kink-turn core. The dinucleotide preceding the first GA is strongly conserved as AU (forming the AUGA tetranucleotide), or rarely substituted with GU. The first nucleotides in the loop at the top of stem 2 are strongly conserved as adenines, with few exceptions. Lastly, additional positions around the core show preference toward specific nucleotides (Chapple et al. 2009). Eukaryotic SECISes are bound by *SBP2*, a L7AE-domain containing protein that contacts the region around the core (Fletcher et al. 2001). Besides recognizing SECIS elements, *SBP2* interacts with *EFsec* (Tujebajeva et al. 2000).

In archaea, SECIS elements (aSECIS) are also located in the 3′UTR of selenoprotein genes, with a single documented exception in which it is found in the 5′UTR (Wilting et al. 1997). Archaeal SECISes (Krol 2002; Kryukov and Gladyshev 2004; Stock and Rother 2009) are characterized by two stems separated by an invariant asymmetric bulge, consisting of a GAA trinucleotide at the 5′ and a single adenine at the 3′. The first stem is GC rich and encompass 10 bp. The second stem is shorter, generally only 3 bp, and is entirely composed of GC pairs. The apical loop has variable length, and may contain additional pairings. To date, the archaeal SECIS element has no known interactor protein, and the question of how the SECIS and the Sec UGA site communicate remains open. The *SBP2* counterpart has never been observed in archaea, while the archaeal *EFsec* was shown not bind archaeal SECIS elements (Stock and Rother 2009).

Evolution of the Sec Trait

The distribution of selenoproteins in living organisms, particularly in prokaryotes and protists, testifies a dynamic process in which the Sec trait was lost in many lineages independently (Zhang et al. 2006; Lobanov et al. 2008; Mariotti et al. 2015). Although some isolated events of horizontal gene transfer have occurred, in general both the distribution of proteins in the Sec pathway and their reconstructed phylogeny are consistent with the known phylogenetic relationship between bacteria, archaea, and eukaryotes (Mariotti et al. 2015). In addition, considering the evident homology of the core Sec pathway (*tRNAsec*, *EFsec*, *SPS*) it becomes compelling that the Sec trait has evolved only once in the history of life, and its origin can be dated back to the last universal common ancestor. Yet, SECIS elements bear no obvious resemblance between the major domains of life. This is particularly surprising when we consider archaea and eukaryotes, in which SECIS elements are situated in the same location (the 3′UTR), and are thus expected to be homologous structures.

We must consider, however, that selenoproteins are a very rare feature among the archaea analyzed so far. The markers for Sec utilization were identified only in two phylogenetic orders, *Methanococcales* and *Methanopyrales* (*Methanopyrus*), while all other major archaeal groups seem to be devoid of selenoproteins (Stock and Rother 2009). Since the time of the discovery of archaeal selenoproteins (Wilting et al. 1997), many new archaeal sequences became available, but, until today, the Sec trait has remained limited to the same two genera.

The Origin of Eukaryotes and the Discovery of *Lokiarchaeota*

Eukaryotes possess several features that set them apart from prokaryotes, such as a larger cell size and a complex cell compartmentalization (including the presence of the nucleus, mitochondria, and cytoskeleton). Remarkably, these complex features are present in nearly all extant eukaryotes without any apparent intermediate grade, posing the problem of how the first eukaryotes came about. In recent years, the increasing availability of genome sequences and the development of sophisticated phylogenomics tools have finally shed light into the origin of this domain of life. It is now widely accepted that eukaryotes originated from an archaeal ancestor, and that a key event was the incorporation of an alpha-proteobacterial endosymbiont, giving rise to mitochondria (Koonin 2010). The phylogenetic relationships between extant archaea and the proto-eukaryotic archaeal ancestor,

Downloaded from http://mbe.oxfordjournals.org/ at Biblioteca de la Universitat Pompeu Fabra on July 22, 2016

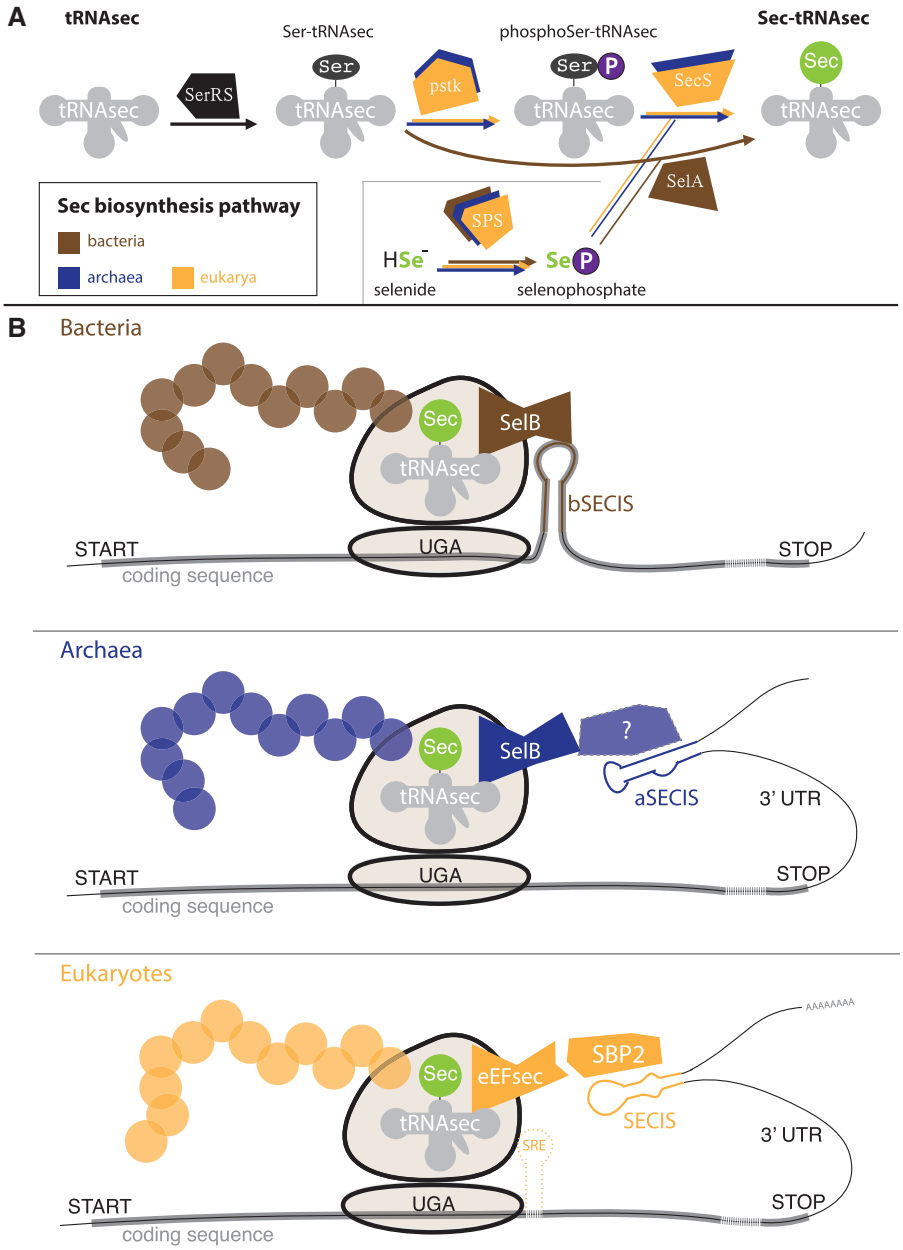

Fig. 1. Sec machinery in the three domains of life. The figure shows two modules of the Sec pathway: Sec biosynthesis (*A*), and Sec insertion during selenoprotein translation (*B*). The differences between bacteria, archaea, and eukaryotes (Stock and Rother 2009; Yoshizawa and Böck 2009; Labunskyy et al. 2014) are marked with different colors. *SerRS* serine-tRNA synthetase, *pstk* phosphoseryl-tRNA kinase, *SecS* archaeal/eukaryotic Sec synthase, *SelA* bacterial Sec synthase, *SPS* selenophosphate synthetase, *SelB* bacterial/archaeal Sec elongation factor, *eEFsec* eukaryotic Sec elongation factor, *SBP2* SECIS binding protein 2, bSECIS bacterial SECIS, *aSECIS* archaeal SECIS (*SECIS* eukaryotic SECIS), *SRE* Sec redefinition element (Howard et al. 2007).

however, remained uncertain: it was unclear whether eukaryotes should be considered a sister group of archaea, or rather branched from within archaea.

A recent study (Spang et al. 2015) has reported the discovery of a novel archaeal phylum, *Lokiarchaeota*, in marine sediments collected near a hydrothermal vent known as Loki's Castle. In their study, the authors describe how they obtained a metagenomic assembly of nonamplified DNA from this site (LCGC14, Genbank accession LAZR01000000.1), hereafter referred to as the *Laz* sequence set. From this assembly, they derived a contig subset by performing supervised binning driven by a carefully selected set of genes with a peculiar phylogenetic signal. The resulting set of contigs (Lokiarchaeum genome bins, Genbank accession JYIM00000000.1), hereafter designated as *Loki*, was subject of extensive phylogenetic analysis. The *Loki* sequences turned out to belong to a new archaeal lineage that appears as a sister group of eukaryotes in phylogenetic reconstructions. This strongly supports the hypothesis that eukaryotes evolved from within the archaeal domain, with *Lokiarchaeota* being

Downloaded from http://mbe.oxfordjournals.org/ at Biblioteca de la Universitat Pompeu Fabra on July 22, 2016

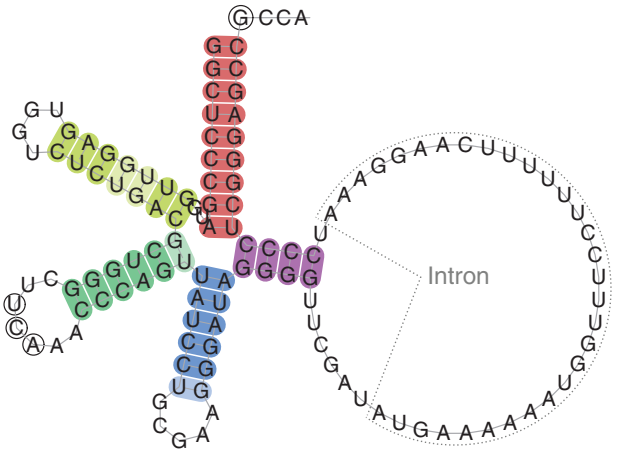

Fig. 2. tRNAsec in *Lokiarchaeota*. The color scheme highlights the tRNA stems: red for the acceptor arm, yellow for the D arm, green for the anticodon arm, blue for the variable arm, and purple for the T arm. This tRNA has all the characteristics expected for an archaeal *tRNAsec*: UCA anticodon, G discriminator base, 9/4 fold of the acceptor/T stems, a D arm of 7 bp, a D loop of 4 bp, and a long variable arm (Chiba et al. 2010). Two copies of *tRNAsec* with very similar sequence were identified in *Loki* (supplementary table S1, Supplementary Material online), but since their gene context is identical (supplementary material S1.A, Supplementary Material online), we conclude that they correspond to the same, sole copy of this gene in *Lokiarchaeota*.

the closest known relative to the archaeal ancestor of eukaryotes (Eme and Doolittle 2015; Koonin 2015; Finn et al. 2016). Consistent with this, several eukaryotic signature proteins were found in *Lokiarchaeota*.

In this study, we present the first analysis of selenoproteins and Sec pathway in *Lokiarchaeota*. A complete set of known archaeal genes to encode for Sec was found in *Loki*, together with selenoprotein genes belonging to five distinct families. Remarkably, the selenoprotein genes in *Lokiarchaeota* possess conserved RNA structures that resemble the eukaryotic SECIS elements. These results identify *Lokiarchaeota* as an intermediate form between the typical archaeal and eukaryotic Sec encoding systems, contributing to the understanding of the origin and evolution of the Sec insertion system.

Results

We analyzed selenoproteins and Sec biosynthesis genes in the *Loki* sequences and further expanded our searches to the *Loki* superset assembly, *Laz*. Both *Laz* and *Loki* are metagenomes, representing a mixture of sequences from a multitude of species (although *Loki* obviously displays a much smaller diversity, approximable to a single genome assembly). Both *Laz* and *Loki* are affected by the potential confounding factor of "apparent paralogues": genes in multiple copies in the metagenome, which actually correspond to orthologous genes of closely related strains or species. To avoid this confounding issue, we decided to analyze each gene in its genomic context (Methods, supplementary material S1, Supplementary Material online). This analysis reduced redundancy of the sets of genes of interest in *Loki*, which allowed us to analyze the sequences in greater detail.

A Complete Sec Machinery in Loki Sequences

We used a variety of computational approaches to search the *Loki* sequences for genes involved in the Sec pathway (Methods). We identified the full set of known genes required for archaeal Sec utilization, here designated as the Sec machinery: *tRNAsec*, *pstk*, *SecS*, *EFsec*, *SPS* (genomic locations are provided in supplementary table S1, Supplementary Material online). All Sec machinery genes were found to be present in single copy in *Loki* and on distinct contigs, with the exception of *EFsec* and *tRNAsec*. *EFsec* was present in the same contig with *tRNAsec*, approximately ~6 kb and eight genes upstream, on the opposite strand. This pair of genes was observed twice, on two distinct contigs in *Loki*. However, we could classify these as apparent paralogues, belonging to very closely related *Lokiarchaeota*. In fact, in both contigs the homology extended to the region between *tRNAsec* and *EFsec*, and beyond (supplementary material S1.A, Supplementary Material online). The only nonprotein coding gene in the Sec machinery, *tRNAsec*, contained a 31-nt intron interrupting the loop in the TΨC arm (fig. 2). Although this position is considered not canonical for tRNA introns, it was previously observed in other archaeal tRNAs (Yoshihisa 2014). The sequence and structure of the mature tRNA resembles unequivocally the archaeal *tRNAsec*. Similarly, all Sec machinery proteins exhibited an archaeal phylogenetic signal, being more similar to *Methanococcales* or *Methanopyrus* orthologues than to their bacterial or eukaryotic counterparts (supplementary material S2, Supplementary Material online).

The *SPS* gene (*SPS*) is singular in that it is both part of the Sec machinery and a selenoprotein itself in many organisms. *SPS* is an ancestral selenoprotein, and although it has replaced Sec with Cys in many organisms, most likely its original form contained Sec (Mariotti et al. 2015). In *Loki*, we identified only a fragment of the SPS gene. It was located on a short contig, which ended in the middle of the *SPS* coding sequence, resulting in an apparently truncated sequence that lacked the Sec-containing N-terminal region. However, we could deduce the complete *Loki SPS* protein sequence analyzing *Laz* sequences. We found several *SPS* genes, some of which were almost identical in sequence to the *Loki SPS* fragment. We thus collected all *SPS* genes in *Laz*, aligned them with a reference set of *SPS* proteins from the whole tree of life (Mariotti et al. 2015), and applied a phylogenetic reconstruction procedure (Methods). The resulting tree offered a "phylogenetic fingerprint" of all *SPS* containing species present in *Laz* (supplementary material S2.A, Supplementary Material online). While some of the *Laz SPS* genes clustered within diverse bacterial lineages, a group of nine *SPS* genes (including the *Loki* sequence fragment) branched within archaea, forming a cluster that resembled the *Methanococcales SPS* gene, yet exhibited considerable sequence divergence. We further refer to these *Laz* genes as the *Loki*-like *SPS*. Their coding sequence contained one in-frame UGA codon aligned to the homologous Sec position in other species, supporting the existence of *SPS* as

Downloaded from http://mbe.oxfordjournals.org/ at Biblioteca de la Universitat Pompeu Fabra on July 22, 2016

Fig. 3. Selenoprotein genes in *Lokiarchaeota*. The *VhuU.1* gene (first on top) serves as legend. Only a fragment of the SPS gene was identified in *Loki*, but the missing sequence could be deduced from similar gene occurrences in *Laz*. The lokiSECIS in SPS is colored differently to indicate its peculiar characteristics (figs. 4 and 5). Note that the coding sequence of genes *HdrA.3* and *VhuD.3* overlap in different frames by 28 nucleotides. The genes *HdrA.4*/*VhuD.4*, *PrxL.1* and *VhuU.1* are present in one extra copy in the *Loki* contigs (supplementary table S1, Supplementary Material online), but these are not shown here since their gene context is identical (supplementary material S1.A, Supplementary Material online).

selenoprotein in *Lokiarchaeota*. We also searched the *Loki* and *Laz* assemblies for the presence of the selenouridine synthase (*ybbB* gene), a marker of utilization of selenouridine in certain tRNAs (Romero et al. 2005). However, none of the possible candidates withstood our manual inspection. Similarly, the search for gene markers of Se utilization as cofactor to certain molybdenum-dependent hydroxylases (*YqeB* and *YqeC*) (Lin et al. 2015) did not return any suitable candidate. Thus, the use of Sec in selenoproteins is the only known Se utilization trait supported by *SPS* in *Lokiarchaeota*.

Selenoprotein Families in *Lokiarchaeota*

We applied multiple approaches to search for selenoprotein genes in *Lokiarchaeota* (Methods). These methods were aimed to identify genes belonging both to known selenoprotein families and to novel families. This search yielded *Loki* selenoprotein genes belonging to five protein families: *SPS*, *VhuU*, *VhuD*, *HdrA*, and *PrxL* (fig. 3). Of these, *VhuU*, *VhuD*, and *HdrA* were orthologous to the selenoprotein genes previously identified in other archaea genera, such as *Methanococcus* and *Methanopyrus* (Stock and Rother 2009), and they clustered together in phylogenetic reconstructions (supplementary material S3, Supplementary Material online).

VhuU and *VhuD* encoded two subunits of the same enzyme, the F_{420}-nonreducing hydrogenase (Vhu), which includes also nonselenoprotein subunits *VhuA* and *VhuG* (both found in the *Loki* genome, see supplementary table S1, Supplementary Material online). In *Loki* and in previously described Sec-containing archaea alike, *VhuU* is a small oxidase (~50 amino acids) with a single Sec residue, while *VhuD* is a ~150 amino acid-long hydrogenase with two distinct Sec positions (Wilting et al. 1997). *HdrA* represents the largest subunit (~660 amino acids) of heterodisulfide reductase (Hdr) complex, which participates, together with Vhu, in the reduction of Coenzyme B-Coenzyme M heterodisulfide (CoM-S-S-CoB) during the last step of methanogenesis. Hdr includes also subunits *HdrB* and *HdrC*. These do not contain Sec, and were also identified in the *Loki* genome (supplementary table S1, Supplementary Material online). In the Sec utilizing archaeal species previously described, *HdrA* has a single Sec residue in position ~200 ("canonical position"). In contrast, in *Loki* sequences we identified several multi-Sec *HdrA* genes, with a total of five possible Sec positions, always including the canonical one (fig. 3). All these positions aligned to conserved Cys residues in other archaeal *HdrA* proteins. Strikingly, all *Loki HdrA* and *VhuD* genes were located in tandem syntenic pairs, with *VhuD* always located downstream of *HdrA* on the same strand (fig. 3). This syntenic block was also conserved in all *Methanococcales* archaea as well as many other archaeal lineages. The intergenic space between these two genes in *Loki* is invariably very short. In one case (*HdrA.3*–*VhuD.3*), the two coding sequences actually overlap by 28 bp, in different frames.

The remaining Sec-containing proteins identified in *Loki* resemble the peroxiredoxin family (Prx), which was never

Downloaded from http://mbe.oxfordjournals.org/ at Biblioteca de la Universitat Pompeu Fabra on July 22, 2016

observed as selenoprotein in archaea. Interestingly, these *Loki* selenoproteins belong to a subclass that lacks typical Prx features. This diverse family, never experimentally characterized, was previously referred to as Prx-like (Fomenko and Gladyshev 2003; Cui et al. 2012). Whereas Prx proteins contain an active site with Thr/Ser–X–X–Cys (where X stands for any residue) (Poole and Nelson 2016), Prx-like proteins possess a Cys–X–X–Cys motif ("redox box") typical of thioredoxins, with the first Cys replaced by Sec in some bacteria (Zhang and Gladyshev 2008) and eukaryotes (Jiang et al. 2012; Mariotti et al. 2013). We identified two distinct Prx-like genes in *Loki*, hereafter referred to as *PrxL*. The *PrxL.2* gene contains a single Sec residue corresponding to the first Cys in the redox box, while *PrxL.1* contains two, thus replacing both Cys residues with Sec. Searching *Laz* sequences, we identified additional *PrxL* homologues, either with one or two Sec residues at the same positions (supplementary table S1, Supplementary Material online). Strikingly, none of the *PrxL* proteins contain any additional Cys residue outside the redox box.

The Loki SECIS Element Resembles the Eukaryotic SECIS

After identifying the selenoprotein genes in *Loki*, we searched their sequences for the occurrence of SECIS elements. To our surprise, none of the tools designed to identify archaeal, bacterial, or eukaryotic SECIS elements reported any significant hit in *Loki* selenoprotein genes. Hence, we outlined a *de novo* computational procedure aimed to identify any RNA motif conserved in these genes (Methods). With this, we detected a common RNA structure located downstream of the coding sequences in every *Loki* selenoprotein gene with the exception of *SPS* (explained later). Following our initial discovery of the structures in the selenoprotein genes located on the *Loki* contigs (supplementary material S4.A, Supplementary Material online), we expanded the search to the full set of *Loki*-like selenoproteins in *Laz* (see Methods), ending up with a final set of 25 different "lokiSECIS" elements (supplementary material S4.B, Supplementary Material online). All these structural motifs were located in close proximity to the end of the selenoprotein coding sequences (fig. 3). In all occurrences of the tandem genes *HdrA* and *VhuD*, only a single lokiSECIS was found downstream of the second gene (*VhuD*), but not anywhere else nearby, suggesting that these pairs of genes share their Sec insertion signal. All lokiSECIS elements in *PrxL*, *VhuU*, *HdrA–VhuD* folded in a very precise shape, featuring a 9-bp stem and an 11-nt apical loop. In a few *VhuD* SECIS elements, the region corresponding to the apical loop was predicted with additional pairings, forming an additional short stem (supplementary material S4.A, Supplementary Material online). From the alignment of all lokiSECIS elements, their conserved features were clearly discernible (fig. 4C).

From the analysis of lokiSECIS elements, we observed the following: (1) the nucleotides preceding the stem were always AUGA; (2) the first five nucleotides of the apical loop showed a strong preference for adenines; (3) downstream of the stem, the first two nucleotides were GA (except that they were AA in *VhuU*). Remarkably, all these characteristics of the lokiSECIS were shared with the eukaryotic SECIS elements. At glance, the lokiSECIS looks like a eukaryotic SECIS with a shorter stem. The lokiSECIS exhibited a few additional conserved features (fig. 4C), the most prominent being an ultraconserved guanine at the last position of the apical loop, most often preceded by an adenine. We detected a lokiSECIS element with these characteristics downstream of every selenoprotein gene in the *Loki* contigs, as well as downstream of those selenoprotein genes in *Laz* which we also attributed to *Lokiarchaeota*, with one exception (*SPS* genes).

Applying an alternative discovery approach (Methods), we found that the *Loki*-like *SPS* genes also possessed a conserved structure in the same homologous position. This structure (fig. 5, supplementary material S4.C, Supplementary Material online) also resembled the eukaryotic SECIS, and had some obvious similarities with the other lokiSECIS elements, but showed important differences as well. All lokiSECIS elements of *PrxL*, *VhuU*, *HdrA–VhuD* genes possessed the same number of nucleotides in the stem and apical loop. In contrast, the main stem was shorter in the *SPS* lokiSECIS, but the overall structure was longer, featuring in all cases an additional stem in the apical region. In this regard, the *SPS* lokiSECIS resembled a eukaryotic type II SECIS, except that it had a shorter stem.

Next, we set out to investigate where the lokiSECIS originated from, searching all publicly available archaeal genomes for similar structures using lokiSECISearch (Methods). This analysis resulted in no significant hits in the great majority of genomes (90%), and only one or two hits in each of the remaining genomes (with the sole exception of *Loki*, in which the program recovered all lokiSECIS elements previously described). However, we noticed that for *Methanococci* species *M. aeolicus* and *M. vannielii*, the unique lokiSECIS hits were located in orthologous positions. Strikingly, they mapped just downstream of the selenoprotein gene *VhuD* and corresponded to the archaeal SECIS of these genes. To further investigate this, we predicted the archaeal SECIS elements of all selenoprotein genes in *Methanococcales* and *Methanopyrus*. We then extended their sequence boundaries beyond the limits normally considered for archaeal SECIS (i.e., fig. 4A), and aligned them. Although this procedure did not uncover new features conserved across all SECIS elements of archaea, we detected motifs conserved in a gene-specific fashion (supplementary material S4.D, Supplementary Material online). The archaeal *VhuD* SECISes were particularly intriguing: they all possessed a short extra stem. Moreover, an invariant AUGA motif on one side, and a GAC/AAC trinucleotide on the other, was observed in between the extra stem and the main stem (fig. 4B). Thus, the *Methanococcales* and *Methanopyrus VhuD* SECIS elements exhibit both the typical features of archaeal SECIS elements and the distinctive motifs of the *Loki* and eukaryotic SECIS.

SBP2 Is Not Detected in *Lokiarchaeota*

At present, no SECIS binding protein is known in the Sec utilizing archaea *Methanococcales* and *Methanopyrus*. In contrast, it is well established that eukaryotic SECIS elements are recognized by *SBP2*, a master regulator in the Sec insertion process (Kossinova et al. 2014), although recent results

Downloaded from http://mbe.oxfordjournals.org/ at Biblioteca de la Universitat Pompeu Fabra on July 22, 2016

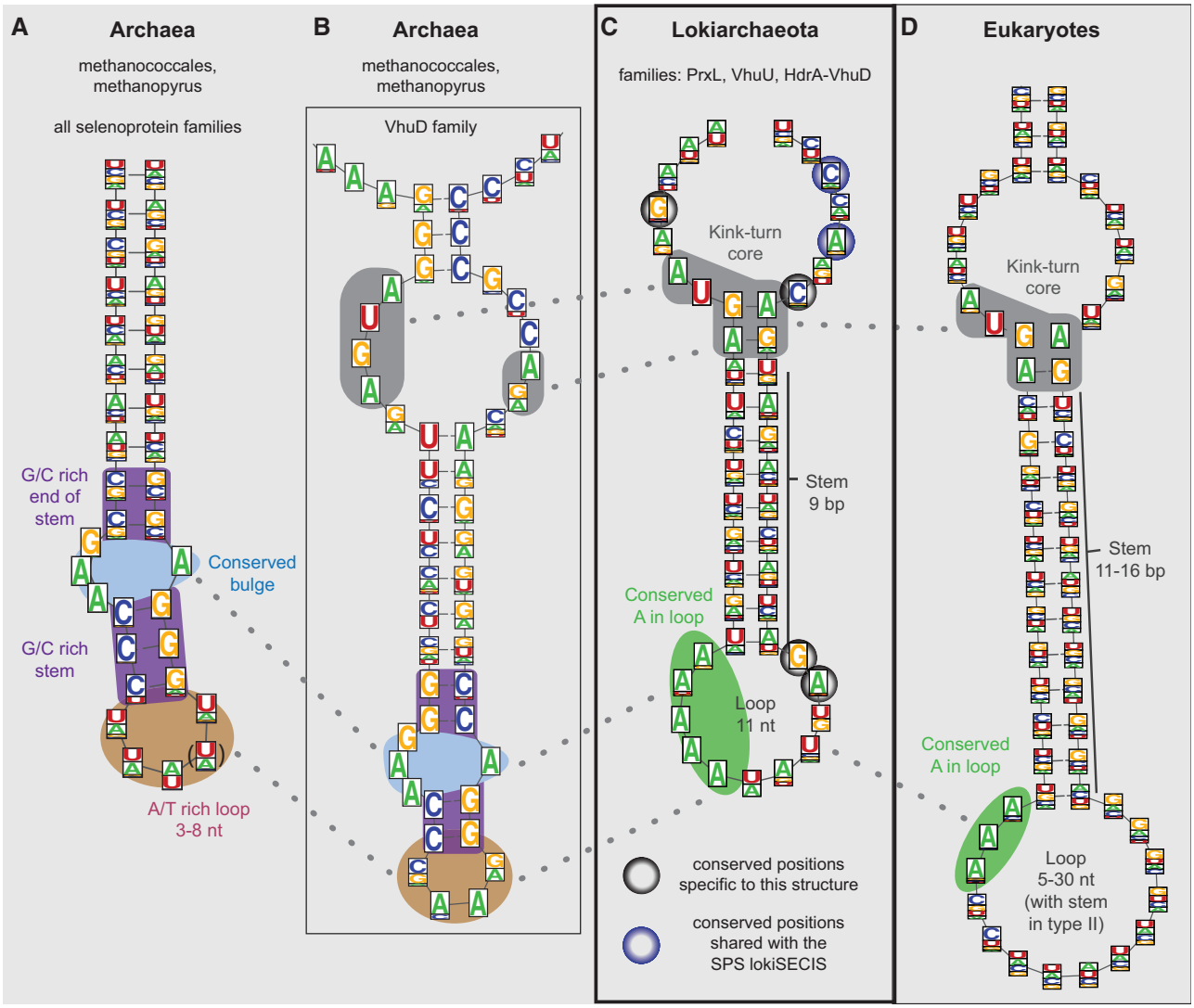

Fig. 4. SECIS elements in *Lokiarchaeota* compared with eukaryotes and other archaea. SECIS elements in known Sec utilizing archaea (*A*), archaeal *VhuD* genes (*B*), *Lokiarchaeota* (see also fig. 5) (*C*), and eukaryotes (*D*). These images were derived from curated SECIS alignments as explained in Methods. The structures are predicted by RNAalifold (Bernhart et al. 2008) and incorporate the frequency logos (Crooks et al. 2004) of the nucleotides found at each position.

questioned the absolute requirement of *SBP2* for selenoprotein expression (Seeher and Schweizer 2014). Although with remarkable variability in size, *SBP2* is found in every selenoprotein containing eukaryotic genome, and always includes an L7AE domain (Donovan and Copeland 2009). This domain is shared with other RNA binding proteins that also recognize kink-turn motifs (Huang and Lilley 2013). In consideration of the similarity of the lokiSECIS with the eukaryotic SECIS, and in particular the conservation of the kink-turn region, we devised a thorough phylogeny-based search for *SBP2* homologues in *Lokiarchaeota* (see Methods). We used an L7AE protein profile to identify all proteins containing this domain both in the *Laz* proteome, and in all eukaryotic and archaeal proteomes available at NCBI. After reducing the set to workable size by automatic selection of representative sequences, we ran our phylogenetic reconstruction pipeline. In the resulting phylogenetic tree (supplementary material S4.E, Supplementary Material online), all known *SBP2* and *SBP2-like* sequences clustered together, enabling us to classify as *SBP2* all the genes in this cluster. This procedure identified known *SBP2* genes throughout all eukaryotic lineages, including the very diverse protozoans and also nematodes, which possess an elusive, very short isoform (Otero et al. 2014). Despite our efforts, however, we could not detect any *SBP2* candidate in any *Laz* (or *Loki*) sequences, or in any other archaea.

Discussion

Selenoproteins are synthesized through a conserved UGA-recoding mechanism, considered an expansion of the genetic code (Böck et al. 1991). Living organisms use this intriguing strategy to insert the amino acid Sec in catalytic sites of certain oxidoreductases. Although Sec was likely present in the last universal common ancestor, and has since been preserved in organisms in the three domains of life, including humans, it is absent in a considerable fraction of extant organisms. Although approximately half of sequenced eukaryotes contain selenoproteins, these proteins are found only in

Downloaded from http://mbe.oxfordjournals.org/ at Biblioteca de la Universitat Pompeu Fabra on July 22, 2016

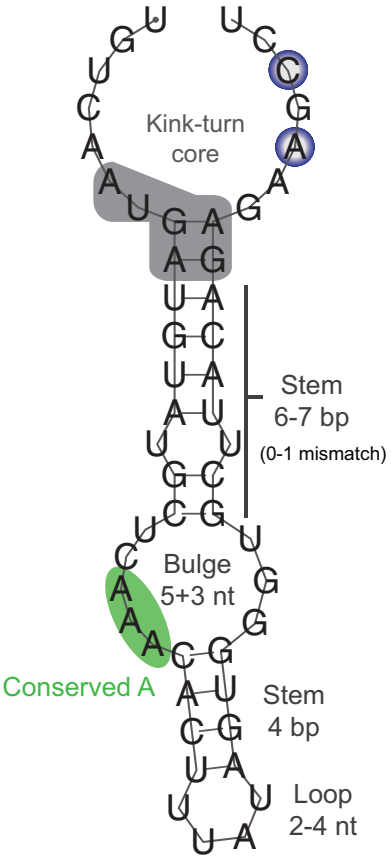

Fig. 5. A distinct SECIS element in Loki *SPS*. The figure shows the SECIS predicted downstream of the *Loki* SPS fragment. This structure, predicted by RNAfold (Lorenz et al. 2011), is analogous to that predicted for all other *Loki*-like *SPS* in *Laz*, and also to their consensus structure predicted by RNAalifold (supplementary material S4.C, Supplementary Material online). The figure highlights the features shared with the rest of *Loki* SECIS and the eukaryotic SECIS (fig. 4). Those nucleotide positions conserved in all *Lokiarchaeota* SECISes, but not in eukaryotes, are circled in blue.

~20% of sequenced bacterial genomes (Lin et al. 2015; Mariotti et al. 2015), and Sec utilization is even more scarce within archaea: selenoproteins were so far detected only in two archaeal orders, *Methanopyrus* and *Methanococcales*, comprising ~12% of currently available sequenced archaeal genomes. In this study, we describe the discovery of the Sec trait in a novel archaeal phylum, *Lokiarchaeota*, the closest archaeal relative to eukaryotes known to date (Spang et al. 2015). The identity and functions of the selenoprotein genes we identified is backed by the monophyly with their experimentally characterized orthologues. Our results show that *Lokiarchaeota* possess the full set of genes to support Sec biosynthesis and incorporation, as well as selenoprotein genes belonging to five different families (*SPS*, *VhuU*, *VhuD*, *HdrA*, and *PrxL*, fig. 3). All selenoprotein families identified in *Loki*, except one, were previously identified in other archaeal lineages (Stock and Rother 2009). The only exception, *PrxL*, belongs to the thioredoxin-like superfamily, like many other selenoproteins, and it was reported in other computational studies both in bacteria (Zhang and Gladyshev 2008) and eukaryotes (Jiang et al. 2012; Mariotti et al. 2013).

We noticed that some of the *Loki* selenoproteins carry a rather unusual number of Sec residues. Whereas the known Prx and Prx-like selenoproteins have a single Sec site, we observed two in *Loki PrxL.1*, as well as in some of its homologues in the *Laz* assembly (supplementary table S1, Supplementary Material online). The Sec residues are found in a redox box, suggesting that the *PrxL.1* protein forms a diselenide bond, a feature that has only one additional known case (Shchedrina et al. 2007). *HdrA* has also never been observed with more than one Sec residue, but in *Loki* we found such genes containing up to four Sec sites, located in five possible positions (fig. 3). *HdrA* is FAD-dependent protein with four [4Fe–4S] clusters coordinated by strictly conserved Cys residues. All Sec sites align to Cys positions that are conserved across the entire *HdrA* family, but that are not the ones involved in cluster coordination, with the only exception of the most C-terminal Sec in *HdrA.2* (fig. 3). This Sec residue is aligned with one of the four Cys in the last [4Fe–4S] binding motif; however, Sec is here followed by an additional Cys right next to Sec, so that the motif contains four Cys and one Sec, instead of the standard four Cys residues. *HdrA* functions as an electron-transfer relay in the Hdr complex, providing reducing power to the catalytic subunit *HdrB* (Hedderich et al. 2005). The presence of several Sec residues in *Loki HdrA* would indicate that *HdrA* possesses multiple catalytic Cys residues. This suggests an unusual redox chemistry that has not been explored for this protein, traditionally considered as an intermediate in electron transfer.

The most intriguing discovery in our analysis was that the *Lokiarchaeota* selenoprotein genes possess a eukaryotic-like SECIS element. In fact, the lokiSECIS exhibits a conserved core which is essentially identical to the eukaryotic SECIS, and also shows the same conserved stretch of adenosines in the apical loop (fig. 4). Since the core of the eukaryotic SECIS folds in a kink-turn motif (Latrèche et al. 2009), we expect the same structure from the lokiSECIS. This implies that this motif, and the general fold of the eukaryotic SECIS, had been already established in *Lokiarchaeota* and was later maintained in eukaryotes. Furthermore, the lokiSECIS also distantly resembles one particular SECIS element of the selenoprotein gene *VhuD* in several archaea. Remarkably, all *Methanococcales* and *Methanopyrus VhuD* SECIS elements conserve the core motif AUGA preceding the main stem, which is not found in the rest of SECIS elements in the same genomes. This suggests that the archaeal *VhuD* SECIS is the prototypical SECIS element, and that every lokiSECIS, and by extension every eukaryotic SECIS element, descended from it.

Further support to this idea comes from the analysis of *VhuD* and its genomic neighbor *HdrA*. In an unprecedented configuration, in *Lokiarchaeota* these two genes share a single lokiSECIS located downstream of *VhuD*, and thus are most likely translated from the same polycistronic mRNA. In addition, these genes always co-occur with the same genomic organization (fig. 3) in *Lokiarchaeota* (supplementary material S1, Supplementary Material online). In every single case, *HdrA* is upstream of *VhuD*, with the same orientation and

Downloaded from http://mbe.oxfordjournals.org/ at Biblioteca de la Universitat Pompeu Fabra on July 22, 2016

extremely short intergenic distance. Remarkably, it seems that the single lokiSECIS in *HdrA–VhuD* can support Sec incorporation for up to six UGA codons, which span a distance of ~2 kb. This would be unparalleled in the tree of life. Outside *Lokiarchaeota*, very few selenoprotein genes with more than one Sec residue are known: *SelP* (vertebrates) (Hill et al. 1993; Lobanov et al. 2008), *SelL* (some metazoans) (Shchedrina et al. 2007), *MsrB* of *Metridium senile* (cnidarian) (Lee et al. 2011), and the *VhuD* homologue in other Sec utilizing archaea. *SelP* and *M. senile MsrB* have several Sec residues, and possess two SECIS elements each. *SelL* possess a single SECIS, which acts on two UGA codons in very close proximity (two codons apart). The *Methanococcales* and *Methanopyrus VhuD* has been the record holder for most distant Sec residues inserted by a single (archaeal) SECIS (50 codons apart). In *Lokiarchaeota*, the *VhuD* gene extended this pattern to its neighboring gene, *HdrA*. Plausibly, the peculiar features of the archaeal *VhuD* SECIS were advantageous in *Lokiarchaeota*, so that it took over the role of the neighboring *HdrA* SECIS, which degenerated. Ultimately, the *VhuD* SECIS 'dictated' its characteristics to the rest of SECIS elements, becoming effectively the new prototype structure for SECISs in *Lokiarchaeota*, and then eukaryotes. The AUGA–GA motif in particular, which is conserved in the SECIS of archaeal *VhuD*, but not in other genes, spread to all selenoprotein genes and became what is known today as the SECIS core.

By examining the identity of the stop codons at the end of the *Lokiarchaeota* selenoprotein coding sequences, we noticed that many of them are UGA codons (supplementary table S1, Supplementary Material online). Since LokiSECISes are located in close proximity to stop codons, structural constraints would preclude SECIS function for these positions. Thus, it appears that the lokiSECIS is able to support Sec incorporation in distant UGA codons, but has a minimal distance requirement, analogously to the eukaryotic SECIS (Martin et al. 1996; Labunskyy et al. 2014). This model finds support in the observation that, while the various *VhuD* genes bear diverse stop codons including UGA, *HdrA* always carries UAA or UAG. We reasoned that, if this stop codon was a UGA, it would be also translated as Sec, and this mutation is selected against. It appears that the lokiSECIS can overlap the coding sequence, at least partially (fig. 3). In the *VhuU.1* gene, the overlap is particularly extensive, so that its AUGA motif actually works also as the stop codon for this gene (UGA).

Despite our efforts, we could not detect *SBP2* in *Loki*. Its apparent absence suggests that the transformation of the SECIS predates the origin of this protein. A recent study (Seeher and Schweizer 2014) showed that, surprisingly, the deletion of *SBP2* in mouse neurons did not fully abolish selenoprotein expression. It is tempting to make a connection between this evidence of *SBP2*-independent Sec insertion in mammals and the situation in *Loki*, where eukaryotic-like SECIS elements are present but no SBP2 could be detected. However, the *SBP2*-independent selenoprotein expression in mouse may also be explained by the presence of *SBP2L*, an L7AE-domain containing paralog of *SBP2* that emerged at the root of vertebrates, that was previously shown to weakly bind SECIS elements *in vitro* (Donovan and Copeland 2009). Unfortunately, the question of archaeal SECIS recognition remains, to date, unanswered. Like for the rest of Sec utilizing archaea, we suppose that the *SBP2* function in *Lokiarchaeota* is either carried out by an unknown dedicated protein, or performed by a constitutive ribosomal component (e.g., *L30*) as an accessory function. Given the different SECIS structures in *Lokiarchaeota* and other Sec encoding archaea, it is plausible that the protein responsible for their binding is distinct in the two cases.

In conclusion, this study sheds light on the transition of the Sec recoding pathway from archaea to eukaryotes, testifying the value of the *Lokiarchaeota* genome as a "living fossil" between prokaryotes and eukaryotes. The selenoprotein genes in *Lokiarchaeota* are mostly typical of the archaeal world, but they possess conserved RNA structures with unmistakable similarity to eukaryotic SECIS elements.

Materials and Methods

Identification of Selenoprotein and Sec Machinery Genes

We searched the *Loki* sequences for selenoprotein genes using a combination of computational methods. First, we applied Selenoprofiles, a homology-based pipeline able to correctly predict known selenoprotein families in genomes even in complete automation (Mariotti and Guigó 2010). Second, we ran blast searches (Altschul et al. 1997) (tblastn program) using a comprehensive set of selenoproteins annotated in other species, and further manually inspected the results. Third, we used a modified version of Seblastian (Mariotti et al. 2013). This program identifies potential SECIS elements as a first step, and then searches for known or novel selenoproteins in the sequence upstream of each SECIS, using blastx against a database of known selenoproteins or potential Cys homologues. For *Loki*, due to the modest total sequence size, we could apply a modified Seblastian that bypassed SECIS finding and searched the full metagenome instead. All candidates from the three approaches were merged in a single set, and were subjected to extensive manual analysis and gene structure refinement. This resulted in a set of 13 UGA-containing *Loki* selenoprotein genes, belonging to the families *HdrA*, *VhuD*, *VhuU*, and *PrxL* (supplementary table S1, Supplementary Material online). Analyzing their genomic neighborhoods (see Synteny Analysis), we reduced this to a nonredundant set of ten UGA-containing selenoprotein genes. This set allowed the first discovery of the lokiSECIS motif, which was then progressively enriched with homologous sequences in *Laz* to form a lokiSECIS model (see Search for *Lokiarchaeota* SECIS). Performing searches with this model, we obtained a (potentially redundant) set of 33 selenoproteins in *Laz* (a superset of the *Loki* selenoprotein set), which all belong to aforementioned protein families. In addition to these genes, we identified ten *Loki*-like *SPS* selenoproteins in *Laz*. Furthermore, the *Laz* sequence set contained many other selenoprotein sequences attributed to a variety of bacterial species, which were filtered out. All selenoprotein sets were subjected to phylogenetic analysis (supplementary material S1, Supplementary Material online). Sec machinery

Downloaded from http://mbe.oxfordjournals.org/ at Biblioteca de la Universitat Pompeu Fabra on July 22, 2016

protein coding genes (*SecS*, *EFsec*, *pstk*) were identified using Selenoprofiles and manual tblastn searches, followed by phylogenetic analysis to ensure the correct assignment of gene family (supplementary material S3, Supplementary Material online). The search for *SBP2* was carried out with a particularly exhaustive procedure (see Search for SBP2). *tRNAsec* was identified using a newly developed method based on covariance models, which was built specifically to find this gene in nucleotide sequences (Santesmasses D, in preparation).

SECIS Models for Archaea and Eukaryotes

For archaea, a first structural alignment of a few SECIS elements was manually assembled based on known sequences of archaeal selenoprotein genes (Kryukov and Gladyshev 2004). This first "seed" model was subsequently enriched with additional archaeal SECISes, in the following way. From all archaeal genomes available at NCBI, we selected those that, based on the presence of Sec machinery (analogously to Mariotti et al. 2015), were predicted to code for Sec. All such genomes belonged to the *Methanococcales* and *Methanopyrus* orders. We used Selenoprofiles to identify their selenoprotein genes, and finally scanned their downstream sequences with the seed model, using the program infernal (Nawrocki and Eddy 2013). The final set of potential archaeal SECISes was inspected and filtered to obtain a *bona fide* set (71 sequences). The criteria chosen for filtering were the distance with the coding sequence and the fit with the seed consensus structure. This archaeal SECIS model was used for two purposes. First, it was used to search the *Loki* selenoproteins for archaeal SECISes, which returned no hits. Second, it was used to obtain a graphical representation of the archaeal SECIS consensus (fig. 4). To this aim, we removed all columns with >85% gaps, and the resulting alignment was run with weblogo (frequency plot) (Crooks et al. 2004) and RNAalifold (Bernhart et al. 2008). The images obtained in this way were assembled and colored to produce the scheme shown in figure 4.

For eukaryotic SECISes, searches were performed using the program SECISearch3 (Mariotti et al. 2013), set at best sensitivity. To obtain a graphical representation of the eukaryotic SECIS consensus, we used the structural alignment underlying SECISearch3. This large set (1,121 sequences) was reduced by selecting the ~250 most representative sequences using trimal (Capella-Gutiérrez et al. 2009), and further trimmed by removing all columns with >25% gaps. The resulting alignment was processed with weblogo and with RNAalifold to produce the two graphical components (frequency logos and structure) then assembled in figure 4. For aesthetic reasons, the noncanonical pairing of the two GA dinucleotides at the kink-turn core was imposed as constraint when running RNAalifold.

Search for the *Lokiarchaeota* SECIS and RNA Structure Prediction

Initially, we searched the *Loki* selenoprotein gene sequences using SECIS covariance models (Nawrocki and Eddy 2013) from archaea, eukaryotes (see above), and bacteria (Santesmasses D, unpublished), but this gave no significant matches. We then set out to search for motifs *de novo*. Our sequence dataset for motif search consisted in the nonredundant *Loki* selenoprotein set with the addition of the *Loki* SPS fragment (supplementary table S1, Supplementary Material online). We considered three sequence classes: the coding sequences, the regions just upstream, and those just downstream. We exploited the NCBI annotation of *Loki* sequences to delimit the upstream and downstream regions, by cutting just before the start or end of the next annotated gene. Regions that were shorter than 20 nucleotides were excluded from all subsequent analyses, removing in the process the small intergenic regions between the *HdrA-VhuD* pairs. All remaining upstream and downstream regions were further extended by 30 nucleotides towards their gene, enabling the motifs to have a partial overlap with coding sequences. We then ran the motif search program Glam2, able to predicted motifs potentially containing gaps (Frith et al. 2008), on the three sequence sets separately (coding sequences, upstream, downstream). To assess significance of the predicted motifs, we searched for them again in the full genome and examined the locations and scores of their occurrences along the *Loki* selenoprotein genes, as well as their predicted structure and free energy. The best scoring motif in the downstream sequences appeared to be a suitable candidate: in contrast to all other motifs, it folded consistently in very similar shapes and all localized in homologous positions, in close proximity to the putative translation termination site. The occurrences of this motif were used to train a first "lokiSECIS" infernal model (Nawrocki and Eddy 2013), using RNAalifold (Bernhart et al. 2008) to predict a consensus RNA structure and emacs ralee (Griffiths-Jones 2005) to inspect it. A pattern expression for the program scan_for_matches (http://blog.theseed.org/servers/2010/07/scan-for-matches.html, last accessed June 22, 2016) was also manually designed to fit the occurrences of this motif in *Loki* selenoproteins. We then built a "lokiSECISearch" program. Inspired by the original SECISearch algorithm (Kryukov et al. 1999), this program uses both the motif pattern and the infernal model to search nucleotide sequences. It then predicts the structure of the hits using RNAfold (Lorenz et al. 2011), and filters out those with free energy greater than −5.0 kcal/mol. Finally, the program reports any UGA-containing ORFs located upstream of SECIS candidates. We ran lokiSECISearch on the *Loki* and *Laz* metagenomes. While no further selenoprotein candidate was predicted in *Loki*, after filtering and manual curation we obtained a set of 33 selenoprotein genes in *Laz*, all belonging to the families *HdrA*, *VhuD*, *VhuU*, *PrxL*. We used the downstream motif occurrences in these genes to enrich our lokiSECIS model, which finally consisted of 25 sequences. To produce a graphical representation like for the archaeal and eukaryotic models, the lokiSECIS alignment was trimmed removing columns with >85% gaps, processed with weblogo and RNAalifold (imposing the pairing of the GA dinucleotides at the core), and finally colored and assembled (fig. 4).

With our enriched lokiSECIS model, we could now detect a match downstream of each single *Loki* selenoprotein gene (for *HdrA–VhuD* pairs, the lokiSECIS was located only

Downloaded from http://mbe.oxfordjournals.org/ at Biblioteca de la Universitat Pompeu Fabra on July 22, 2016

downstream of the second gene). Notable exceptions were the *Loki* SPS fragment, as well as the *Loki*-like SPS in *Laz*, which did not exhibit this motif. At this stage, however, due to searches in *Laz*, we had at hand an adequate number of *Loki*-like selenoprotein sequences to attempt a different *de novo* approach to find conserved structures. For each selenoprotein family separately (*HdrA*, *VhuD*, *VhuU*, *PrxL*, and *SPS*), we extracted the genes' coding sequences extended by 200 nucleotides at each side and aligned them using clustal omega (Sievers et al. 2011). We then used the program RNAz (Gruber et al. 2010) to detect any family-specific conserved structure supported by compensatory mutations. RNAz reported significant hits corresponding to every lokiSECIS element predicted so far (downstream of *PrxL*, *VhuD*, and *VhuU* genes). Furthermore, it predicted a stable structure downstream of the *Loki*-like *SPS* genes. This structure, although exhibiting obvious similarities to the lokiSECIS of the other *Loki* selenoprotein genes, has also several important differences (see Results and fig. 5), which make it evade the detection power of both the enriched lokiSECIS infernal model and patscan pattern. RNAz did not report any significant hits other than those just described. We built an additional infernal model for the SPS lokiSECIS structure, and incorporated it in the program lokiSECISearch.

Synteny Analysis

We set out to find homologous relationships in the genomic context of our genes of interest (the selenoproteins and Sec machinery in *Loki* and *Laz*). In order to do that, we downloaded the NCBI protein annotations of *Loki* and *Laz*, and we integrated them with our curated selenoprotein gene sets (since selenoproteins were not correctly annotated in NCBI). The resulting protein set was run with BLASTp against itself, and partitioned in protein families by single link clustering based on the BLASTp matches with an *e*-value lower than 10^{-12}. We then used the syntheny_view program (available at https://github.com/marco-mariotti/tree_classes, last accessed June 22, 2016), a ETE2-based (Huerta-Cepas et al. 2010) script that produces a graphical representation of gene surrounds, highlighting with the same color the genes that belong to the same family. With this visualization, it became evident that several genes were present in multiple copies with a homologous gene context (supplementary material S1, Supplementary Material online). We considered these to be different "versions" of the same gene (i.e., orthologues in closely related species, or just variants in a population), and based on this we reduced our initial *Loki* gene set to a nonredundant set (supplementary table S1, Supplementary Material online).

Search for SBP2

Our initial searches for *SBP2* in *Loki* were performed using Selenoprofiles (Mariotti and Guigó 2010) and tBLASTn of known eukaryotic homologues, but gave no suitable candidates. We thus performed a more thorough search, with the following procedure. We downloaded all archaeal and eukaryotic NCBI genome assemblies and extracted their protein coding annotation using the Entrez module of Biopython (Cock et al. 2009) and GBParsy (Lee et al. 2008). We then scanned all the proteomes obtained in this way using hmmer version 3.1b1 (Eddy 2009), with the pfam profile (Finn et al. 2016) of the L7AE domain (PF01248.22), which is conserved in all *SBP2* proteins but also shared with other ribonucleoproteins. We performed the same hmmer search on all proteins annotated in the *Laz* sequences. We extracted the protein sequences of all domains matching this profile with an *e*-value lower than 0.01 (4,527 sequences), and we aligned them using the program mafft v7.215 (Katoh and Standley 2013). We then used trimal (Capella-Gutiérrez et al. 2009) to select the best representative sequences with a maximum sequence identity of 92%, obtaining a reduced set of 1,962 proteins with an L7AE domain. Finally, we ran our phylogenetic reconstruction pipeline (see below) on this alignment, and we inspected the resulting tree (supplementary material S4.E, Supplementary Material online). The sequences consistently clustered by their annotated family, with known *SBP2* proteins (including here the vertebrate *SBP2-like* subfamily—Donovan and Copeland 2009) falling in a single cluster. We concluded that all domains in this cluster, and only these, belonged to *bona fide SBP2*.

Phylogenetic Analysis

Phylogenetic trees were computed by maximum likelihood with the evolutionary model resulting from an automated selection procedure, as explained in Mariotti et al. (2012) after Huerta-Cepas et al. (2011). The input were protein sequence alignments generated by mafft v7.215 (Katoh and Standley 2013) or clustal omega v1.2.1 (Sievers et al. 2011). For each protein family of interest (selenoproteins and Sec machinery), we included in the alignment their most similar proteins in NCBI NR, identified using BLASTp. The *e*-value threshold was manually adjusted for each family, making sure that searches were permissive enough to include also similar, but nonorthologous proteins (e.g., *EF-TU* sequences were included in the *EFsec* alignment). Trimal (Capella-Gutiérrez et al. 2009) was run to remove very similar sequences (>90% identity) and also to trim noninformative columns using a method optimized for maximum likelihood (*−automated1* option, see trimal manual). For *SPS* alone, instead of collecting homologous sequences with BLASTp, we used the manually curated set in Mariotti et al. (2015).

Supplementary Material

Supplementary materials S1–S4 and table S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

Authors thank anonymous reviewers for useful suggestions. This study was supported by NIH GM061603, GM065205 and CA080946. B.M. is partly supported by The Pew Charitable Trust postdoctoral fellow program.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation

Downloaded from http://mbe.oxfordjournals.org/ at Biblioteca de la Universitat Pompeu Fabra on July 22, 2016

of protein database search programs. *Nucleic Acids Res*. 25:3389–3402.
Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF. 2008. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics* 9:474.
Berry MJ, Banu L, Chen YY, Mandel SJ, Kieffer JD, Harney JW, Larsen PR. 1991. Recognition of UGA as a selenocysteine codon in type I deiodinase requires sequences in the 3' untranslated region. *Nature* 353:273–276.
Böck A, Forchhammer K, Heider J, Baron C. 1991. Selenoprotein synthesis: an expansion of the genetic code. *Trends Biochem Sci*. 16:463–467.
Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–1973.
Castellano S. 2009. On the unique function of selenocysteine—insights from the evolution of selenoproteins. *Biochim Biophys Acta*. 1790:1463–1470.
Chapple CE, Guigó R, Krol A. 2009. SECISaln, a web-based tool for the creation of structure-based alignments of eukaryotic SECIS elements. *Bioinformatics* 25:674–675.
Chiba S, Itoh Y, Sekine S, Yokoyama S. 2010. Structural basis for the major role of O-phosphoseryl-tRNA kinase in the UGA-specific encoding of selenocysteine. *Mol Cell*. 39:410–420.
Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25:1422–1423.
Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res*. 14:1188–1190.
Cui H, Wang Y, Wang Y, Qin S. 2012. Genome-wide analysis of putative peroxiredoxin in unicellular and filamentous cyanobacteria. *BMC Evol Biol*. 12:220.
Donovan J, Copeland PR. 2009. Evolutionary history of selenocysteine incorporation from the perspective of SECIS binding proteins. *BMC Evol Biol*. 9:229.
Driscoll DM, Chavatte L. 2004. Finding needles in a haystack. In silico identification of eukaryotic selenoprotein genes. *EMBO Rep*. 5:140–141.
Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. *Genome Inf*. 23:205–211.
Eme L, Doolittle WF. 2015. Archaea. *Curr Biol*. 25:R851–R855.
Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al. 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res*. 44:D279–D285.
Fletcher JE, Copeland PR, Driscoll DM, Krol A. 2001. The selenocysteine incorporation machinery: interactions between the SECIS RNA and the SECIS-binding protein SBP2. *RNA* 7:1442–1453.
Fomenko DE, Gladyshev VN. 2003. Identity and functions of CxxC-derived motifs. *Biochemistry* 42:11214–11225.
Fomenko DE, Gladyshev VN. 2012. Comparative genomics of thiol oxidoreductases reveals widespread and essential functions of thiol-based redox control of cellular processes. *Antioxid Redox Signal*. 16:193–201.
Frith MC, Saunders NF, Kobe B, Bailey TL. 2008. Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput Biol*. 4:e1000071.
Griffiths-Jones S. 2005. RALEE—RNA ALignment editor in Emacs. *Bioinformatics* 21:257–259.
Gromer S, Johansson L, Bauer H, Arscott LD, Rauch S, Ballou DP, Williams CH Jr, Schirmer RH, Arnér ES. 2003. Active sites of thioredoxin reductases: why selenoproteins? *Proc Natl Acad Sci U S A*. 100:12618–12623.
Gruber AR, Findeiß S, Washietl S, Hofacker IL, Stadler PF. 2010. RNAz 2.0: improved noncoding RNA detection. *Pac Symp Biocomput*. 15:69–79.
Grundner-Culemann E, Martin GW 3rd, Harney JW, Berry MJ. 1999. Two distinct SECIS structures capable of directing selenocysteine incorporation in eukaryotes. *RNA* 5:625–635.
Hedderich R, Hamann N, Bennati M. 2005. Heterodisulfide reductase from methanogenic archaea: a new catalytic role for an iron-sulfur cluster. *Biol Chem*. 386:961–970.
Hill KE, Lloyd RS, Burk RF. 1993. Conserved nucleotide sequences in the open reading frame and 3' untranslated region of selenoprotein P mRNA. *Proc Natl Acad Sci USA*. 90:537–541.
Hondal RJ, Marino SM, Gladyshev VN. 2013. Selenocysteine in thiol/disulfide-like exchange reactions. *Antioxid Redox Signal*. 18:1675–1689.
Hondal RJ, Ruggles EL. 2011. Differing views of the role of selenium in thioredoxin reductase. *Amino Acids* 41:73–89.
Howard MT, Moyle MW, Aggarwal G, Carlson BA, Anderson CB. 2007. A recoding element that stimulates decoding of UGA codons by Sec tRNA[Ser]Sec. *RNA* 13:912–920.
Huang L, Lilley DM. 2013. The molecular recognition of kink-turn structure by the L7Ae class of proteins. *RNA* 19:1703–1710.
Huerta-Cepas J, Capella-Gutierrez S, Pryszcz LP, Denisov I, Kormes D, Marcet-Houben M, Gabaldón T. 2011. PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny. *Nucleic Acids Res*. 39(Database issue):D556–D560.
Huerta-Cepas J, Dopazo J, Gabaldón T. 2010. ETE: a python Environment for Tree Exploration. *BMC Bioinformatics* 11:24.
Hüttenhofer A, Westhof E, Böck A. 1996. Solution structure of mRNA hairpins promoting selenocysteine incorporation in *Escherichia coli* and their base-specific interaction with special elongation factor SELB. *RNA* 2:354–366.
Jiang L, Ni J, Liu Q. 2012. Evolution of selenoproteins in the metazoan. *BMC Genomics* 13:446.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 30:772–780.
Koonin EV. 2010. The origin and early evolution of eukaryotes in the light of phylogenomics. *Genome Biol*. 11:209.
Koonin EV. 2015. Archaeal ancestors of eukaryotes: not so elusive any more. *BMC Biol*. 13:84.
Kossinova O, Malygin A, Krol A, Karpova G. 2014. The SBP2 protein central to selenoprotein synthesis contacts the human ribosome at expansion segment 7L of the 28S rRNA. *RNA* 20:1046–1056.
Krol A. 2002. Evolutionarily different RNA motifs and RNA-protein complexes to achieve selenoprotein synthesis. *Biochimie* 84:765–774.
Kryukov GV, Gladyshev VN. 2004. The prokaryotic selenoproteome. *EMBO Rep*. 5:538–543.
Kryukov GV, Kryukov VM, Gladyshev VN. 1999. New mammalian selenocysteine-containing proteins identified with an algorithm that searches for selenocysteine insertion sequence elements. *J Biol Chem*. 274:33888–33897.
Labunskyy VM, Hatfield DL, Gladyshev VN. 2014. Selenoproteins: molecular pathways and physiological roles. *Physiol Rev*. 94:739–777.
Latrèche L, Jean-Jean O, Driscoll DM, Chavatte L. 2009. Novel structural determinants in human SECIS elements modulate the translational recoding of UGA as selenocysteine. *Nucleic Acids Res*. 37:5868–5880.
Lee BC, Lobanov AV, Marino SM, Kaya A, Seravalli J, Hatfield DL, Gladyshev VN. 2011. A 4-selenocysteine, 2-selenocysteine insertion sequence (SECIS) element methionine sulfoxide reductase from Metridium senile reveals a non-catalytic function of selenocysteines. *J Biol Chem*. 286:18747–18755.
Lee TH, Kim YK, Nahm BH. 2008. GBParsy: a GenBank flatfile parser library with high speed. *BMC Bioinformatics* 9:321.
Lin J, Peng T, Jiang L, Ni JZ, Liu Q, Chen L, Zhang Y. 2015. Comparative genomics reveals new candidate genes involved in selenium metabolism in prokaryotes. *Genome Biol Evol*. 7:664–676.
Lobanov AV, Hatfield DL, Gladyshev VN. 2008. Reduced reliance on the trace element selenium during evolution of mammals. *Genome Biol*. 9:R62.
Lobanov AV, Hatfield DL, Gladyshev VN. 2009. Eukaryotic selenoproteins and selenoproteomes. *Biochim Biophys Acta*. 1790:1424–1428.
Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithm Mol Biol*. 6:26.

Downloaded from http://mbe.oxfordjournals.org/ at Biblioteca de la Universitat Pompeu Fabra on July 22, 2016

Mariotti M, Guigó R. 2010. Selenoprofiles: profile-based scanning of eukaryotic genome sequences for selenoprotein genes. *Bioinformatics* 26:2656–2663.

Mariotti M, Lobanov AV, Guigo R, Gladyshev VN. 2013. SECISearch3 and Seblastian: new tools for prediction of SECIS elements and selenoproteins. *Nucleic Acids Res*. 41:e149.

Mariotti M, Ridge PG, Zhang Y, Lobanov AV, Pringle TH, Guigo R, Hatfield DL, Gladyshev VN. 2012. Composition and evolution of the vertebrate and mammalian selenoproteomes. *PLoS One* 7:e33066.

Mariotti M, Santesmasses D, Capella-Gutierrez S, Mateo A, Arnan C, Johnson R, D'Aniello S, Yim SH, Gladyshev VN, Serras F, et al. 2015. Evolution of selenophosphate synthetases: emergence and relocation of function through independent duplications and recurrent subfunctionalization. *Genome Res*. 25:1256–1267.

Martin GW 3rd, Harney JW, Berry MJ. 1996. Selenocysteine incorporation in eukaryotes: insights into mechanism and efficiency from sequence, structure, and spacing proximity studies of the type I deiodinase SECIS element. *RNA* 2:171–182.

Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29:2933–2935.

Otero L, Romanelli-Cedrez L, Turanov AA, Gladyshev VN, Miranda-Vizuete A, Salinas G. 2014. Adjustments, extinction, and remains of selenocysteine incorporation machinery in the nematode lineage. *RNA* 20:1023–1034.

Poole LB, Nelson KJ. 2016. Distribution and features of the six classes of peroxiredoxins. *Mol Cells* 39:53–59.

Romero H, Zhang Y, Gladyshev VN, Salinas G. 2005. Evolution of selenium utilization traits. *Genome Biol*. 6:R66.

Seeher S, Schweizer U. 2014. Targeted deletion of Secisbp2 reduces, but does not abrogate, selenoprotein expression and leads to striatal interneuron loss. *Free Radic Biol Med*. 75(Suppl 1):S9.

Shchedrina VA, Novoselov SV, Malinouski MY, Gladyshev VN. 2007. Identification and characterization of a selenoprotein family containing a diselenide bond in a redox motif. *Proc Natl Acad Sci U S A*. 104:13919–13924.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*. 7:539.

Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, van Eijk R, Schleper C, Guy L, Ettema TJ. 2015. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* 521:173–179.

Stock T, Rother M. 2009. Selenoproteins in Archaea and Gram-positive bacteria. *Biochim Biophys Acta*. 1790:1520–1532.

Tujebajeva RM, Copeland PR, Xu XM, Carlson BA, Harney JW, Driscoll DM, Hatfield DL, Berry MJ. 2000. Decoding apparatus for eukaryotic selenocysteine insertion. *EMBO Rep*. 1:158–163.

Wilting R, Schorling S, Persson BC, Böck A. 1997. Selenoprotein synthesis in archaea: identification of an mRNA element of *Methanococcus jannaschii* probably directing selenocysteine insertion. *J Mol Biol*. 266:637–641.

Yoshihisa T. 2014. Handling tRNA introns, archaeal way and eukaryotic way. *Front Genet*. 5:213.

Yoshizawa S, Böck A. 2009. The many levels of control on bacterial selenoprotein synthesis. *Biochim Biophys Acta*. 1790:1404–1414.

Zhang Y, Gladyshev VN. 2005. An algorithm for identification of bacterial selenocysteine insertion sequence elements and selenoprotein genes. *Bioinformatics* 21:2580–2589.

Zhang Y, Gladyshev VN. 2008. Trends in selenium utilization in marine microbial world revealed through the analysis of the global ocean sampling (GOS) project. *PLoS Genet*. 4:e1000095.

Zhang Y, Romero H, Salinas G, Gladyshev VN. 2006. Dynamic evolution of selenocysteine utilization in bacteria: a balance between selenoprotein loss and evolution of selenocysteine from redox active cysteine residues. *Genome Biol*. 7:R94.

Downloaded from http://mbe.oxfordjournals.org/ at Biblioteca de la Universitat Pompeu Fabra on July 22, 2016