January 2004

Reconsidering the evolution of eukaryotic selenoproteins: a novel nonmammalian family with scattered phylogenetic distribution

Sergi Castellano
Grup de Recerca en Informàtica Biomèdica, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra

Sergey V. Novoselov
University of Nebraska-Lincoln

Gregory V. Kryukov
University of Nebraska-Lincoln

Alain Lescure
UPR 9002 du CNRS, Institut de Biologie Moléculaire et Cellulaire, 15 Rue René Descartes, 67084 Strasbourg Cedex, France

Enrique Blanco
Grup de Recerca en Informàtica Biomèdica, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, Dr. Aiguader 80, 08003 Barcelona, Catalonia, Spain

See next page for additional authors

Published in *EMBO Reports* 5:1 (January 2004), pp. 71–77. doi: 10.1038/sj.embor.7400036. http://www.nature.com/embor/index.html

Submitted August 28, 2003; revised October 15, 2003; accepted October 15, 2003; published online December 19, 2003.

Scientific Report

Reconsidering the evolution of eukaryotic selenoproteins: a novel nonmammalian family with scattered phylogenetic distribution

Sergi Castellano,[1] Sergey V Novoselov,[2] Gregory V Kryukov,[2] Alain Lescure,[3] Enrique Blanco,[1] Alain Krol,[3] Vadim N Gladyshev,[2] and Roderic Guigó[1,4]*

[1]Grup de Recerca en Informàtica Biomèdica, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, Dr. Aiguader 80, 08003 Barcelona, Catalonia, Spain

[2]Department of Biochemistry, University of Nebraska–Lincoln, Lincoln, Nebraska 65588, USA

[3]UPR 9002 du CNRS, Institut de Biologie Moléculaire et Cellulaire, 15 Rue René Descartes, 67084 Strasbourg Cedex, France

[4]Programa de Bioinformàtica i Genòmica, Centre de Regulació Genòmica, Barcelona, Catalonia, Spain

*Corresponding author. E-mail: rguigo@imim.es

Abstract: While the genome sequence and gene content are available for an increasing number of organisms, eukaryotic selenoproteins remain poorly characterized. The dual role of the UGA codon confounds the identification of novel selenoprotein genes. Here, we describe a comparative genomics approach that relies on the genome-wide prediction of genes with in-frame TGA codons, and the subsequent comparison of predictions from different genomes, wherein conservation in regions flanking the TGA codon suggests selenocysteine coding function. Application of this method to human and fugu genomes identified a novel selenoprotein family, named SelU, in the puffer fish. The selenocysteine-containing form also occurred in other fish, chicken, sea urchin, green algae and diatoms. In contrast, mammals, worms and land plants contained cysteine homologues. We demonstrated selenium incorporation into chicken SelU and characterized the SelU expression pattern in zebrafish embryos. Our data indicate a scattered evolutionary distribution of selenoproteins in eukaryotes, and suggest that, contrary to the picture emerging from data available so far, other taxa-specific selenoproteins probably exist.

Introduction

Selenium is a micronutrient found in proteins in the eubacterial, archaeal and eukaryotic domains of life. It is present in selenoproteins in the form of selenocysteine (Sec), the 21st amino acid. Sec is inserted co-translationally in response to UGA codons, a stop signal in the canonical genetic code. The alternative decoding of UGA depends on several *cis*- and *trans*-acting factors. In eukaryotes, the main *cis*-factor is an mRNA element, the selenocysteine insertion sequence (SECIS), located in the 3'UTR of selenoprotein genes (Walczak *et al*, 1998; Grundner-Culemann *et al*, 1999). About 25 Sec-containing proteins have been identified in eukaryotes (Kryukov *et al*, 2003), but distribution among taxa varies greatly. For instance, no selenoproteins have been found in yeast and land plants, only one in worms and three in flies. The majority of selenoproteins have homologues in which Sec is replaced by cysteine (Cys), even in genomes lacking the Sec-containing gene.

Because of the dual role of the UGA codon, identification of novel selenoproteins in eukaryotes is very difficult. The more direct approach is to search for occurrences of the SECIS structural pattern. Although this approach has been successfully applied in expressed sequence tag (EST) and other cDNA sequences (Kryukov *et al*, 1999; Lescure *et al*, 1999), the low specificity of SECIS searches produces a large number of predictions when applied to eukaryotic genomes. Thus, for the analysis of *Drosophila melanogaster* (Castellano *et al*, 2001, Martin-Romero *et al*, 2001), we devised a strategy that coordinated SECIS identification with prediction of genes with in-frame TGA codons. Again, while this strategy efficiently identified novel selenoproteins in the fly, it resulted in a large number of potential selenoprotein candidates when applied to larger and more complex vertebrate genomes.

Here, we describe a comparative genomics strategy to target bona fide selenoproteins in such complex genomes. Underlying comparative genome methods is the assumption that conservation of function is often reflected in sequence conservation. Indeed, we have already used the fact that SECIS sequences are characteristically conserved between orthologous

genes in our recent characterization of human and mouse selenoproteomes (Kryukov *et al*, 2003). Here, we compare computational predictions of genes with in-frame TGA codons in two different vertebrate genomes, and then search for sequence alignments with conservation around Sec–Sec or Cys–Sec aligned pairs, as suggestive of selenoprotein function. The underlying assumption is that sequence conservation in regions flanking a UGA codon strongly argues for protein coding function across the codon.

We have applied this strategy to human (*Homo sapiens*) and puffer fish (*Takifugu rubripes*) genomes. Our method led to the discovery of a novel selenoprotein family (SelU) in puffer fish, whereas its human counterpart contained Cys. In addition, Sec-containing homologues exist in other fish, chicken, sea urchin, green algae and diatoms. The results presented argue for a scattered phylogenetic distribution of selenoprotein genes, suggesting a quite dynamic Sec/Cys evolutionary exchange.

Results

Comparative gene prediction of novel selenoproteins

We used the geneid program (Guigó *et al*, 1992; Parra *et al*, 2000) to predict standard and TGA-containing genes. geneid predicted 42,357 and 41,127 standard genes in the human and fugu genomes respectively, and 27,605 and 28,603 TGA-containing genes (see Methods and supplementary information online). In all, 20 out of the 23 human selenoprotein genes and 18 out of the 22 fugu selenoprotein genes that were mapped on these genomes were among the predicted TGA-containing genes.

Inter- and intragenomic comparisons in search of Sec–Sec- and Sec–Cys-containing conserved alignments reduced the set of TGA-containing predictions to 133 selenoprotein candidates: 49 orthologous human–fugu selenoprotein predictions, including the 17 known selenoproteins that mapped to both genomes; 58 human selenoproteins with standard fugu orthologues; and 26 fugu selenoproteins with standard human orthologues. Here, we rely on the assumption that coding sequence conservation across a UGA codon between two DNA sequences from different species is strongly suggestive of Sec coding function.

To validate the resulting human–fugu pairs, we undertook an exhaustive search against a number of databases of known coding (proteins and ESTs) and genomic sequences (see supplementary information attached or online). These searches narrowed the number of predicted selenoproteins to 19. This set included two novel human–fugu pairs. Both pairs contained a human standard gene and a fugu selenoprotein gene orthologue, and belonged to the same family. A similar secondary structure pattern around the Sec or Cys residue common to the majority of selenoproteins was found (Castellano *et al*, 2001).

We tested whether newly discovered selenoproteins had SECIS elements in their 3′UTRs. SECIS element prediction was performed in the genomic regions of the two predicted fugu selenoproteins using SECISearch 2.0 (Kryukov *et al*, 2003) with a loose pattern (see Methods). A type 1 SECIS was found for each gene that fitted the established free-energy criteria.

Further homology searches in the fugu and human genomes expanded the fugu selenoprotein family with a third member having also Sec in fugu and Cys in human. This third SelU fugu gene bears a form 2 SECIS and it was not predicted because it lies in a partial contig, missing the 5′ end of the gene.

SelU in *Takifugu rubripes*

The Fugu SelU family (Fig 1) is composed of four members: SelUa and SelUb both have five coding exons with the in-frame TGA located in the second exon; SelUc has four coding exons (although the prediction is incomplete because of the lack of upstream genomic sequence) and the in-frame TGA lies in the first exon; and SelUd has Cys and its gene structure is not known.

SelU in *Homo sapiens*

The human SelU family (Fig 2) is composed of three Cys-containing members. They are uncharacterized predictions by the Ensembl system: ENSG00000122378 is a five-exon gene on chromosome 10, ENSG00000158122 is a six-exon gene on chromosome 9, and ENSG00000157870 has seven exons and maps to chromosome 1. Sequence homology does not apparently suffice to establish the unambiguous orthologous genealogy of the fugu and human SelU proteins (human SelUs named 1–3 in Fig 3).

SelU distribution in eukaryotes

The SelU family is widely distributed across the eukaryotic domain with either Cys- or Sec-containing proteins (Fig 3). Available sequences show that mammals, land plants, arthropods, worms, amphibians, tunicates and slime molds have Cys-containing SelUs, whereas fish, birds, echinoderms, green algae and diatoms carry Sec-containing proteins, although fish and possibly other genomes also have Cys paralogues. Apparently, yeast and flies (among arthropods) lack proteins of this family. Sec is located in SelU proteins close to a conserved Cys such that the two residues form a motif that resembles the CxxC motif that is present in various thiol-dependent redox proteins. Similar motifs are present in a number of eukaryotic selenoproteins, including SelP, SelW, SelV, SelT, SelM and SelH. Conversely, no SelU homologue is present in prokaryotes (see supplementary information online).

Metabolic labelling of SelU with ^{75}Se

To determine whether the SelU family indeed contains Sec (Fig 4), we developed a construct containing the green fluorescent protein (GFP), fused to the carboxy (C)-terminal region of chicken SelU, and the entire 3′UTR (including the predicted SECIS element). The fusion protein was designed such that its size would be different from those of endogenous mammalian selenoproteins. Monkey CV-1 cells transfected with the construct were metabolically labelled with ^{75}Se, and ^{75}Se-containing selenoproteins were analysed by SDS–polyacrylamide gel electophoresis (SDS–PAGE) and a PhosphorImager analysis. This experiment revealed the presence of a ^{75}Se-labelled band corresponding in size to the GFP–SelU fusion protein, if TGA encoded Sec. Thus, SelU is a true selenoprotein.

Expression of SelU during zebrafish embryogenesis

Tissue and temporal expression of the SelU gene during embryogenesis was addressed in the zebrafish model. A probe complementary to the zebrafish SelU cDNA (EST fz58h06.y2,

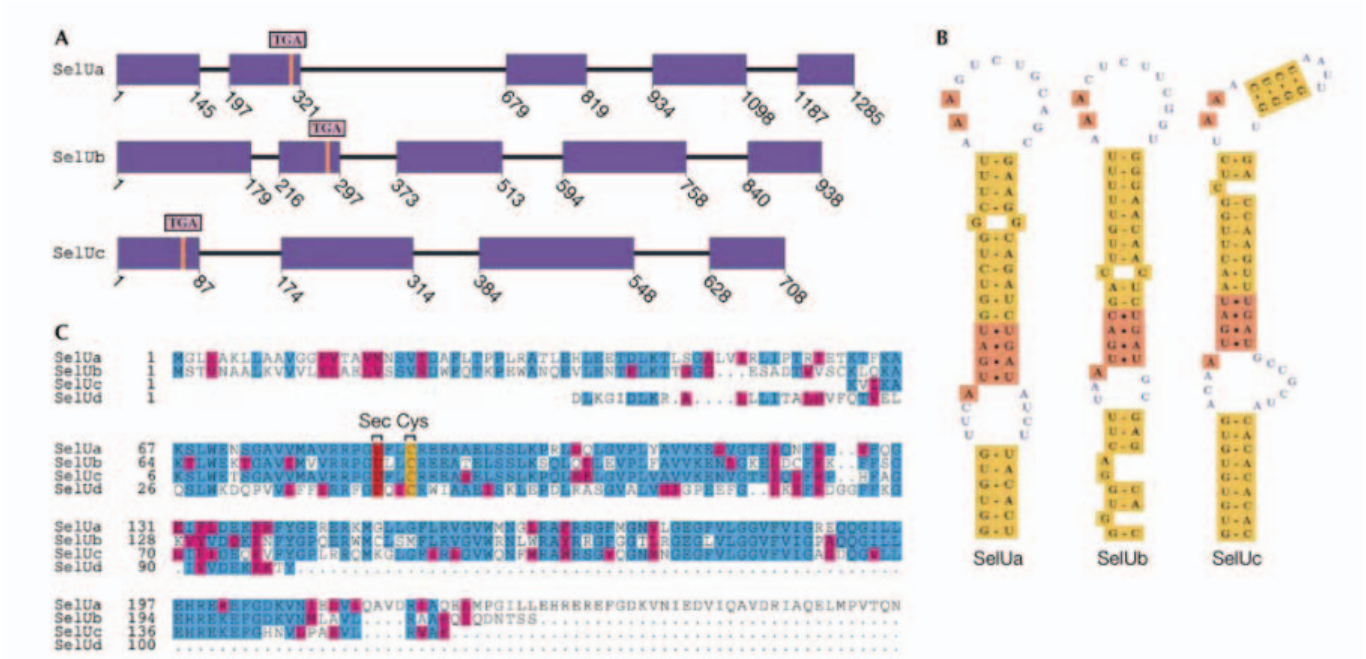

Figure 1. Fugu SelU family. (A) Gene structure (coding exons in purple) plotted using gff2ps (Abril & Guigo´, 2000). Red lines mark the TGA triplet. SelUc is a partial gene lacking the upstream region. (B) SECIS structures. SelUa and SelUb bear a type 1 SECIS and SelUc a type 2 SECIS. (C) Alignment of SelU paralogues using CLUSTAL_W (Thompson et al, 1994). U is Sec.

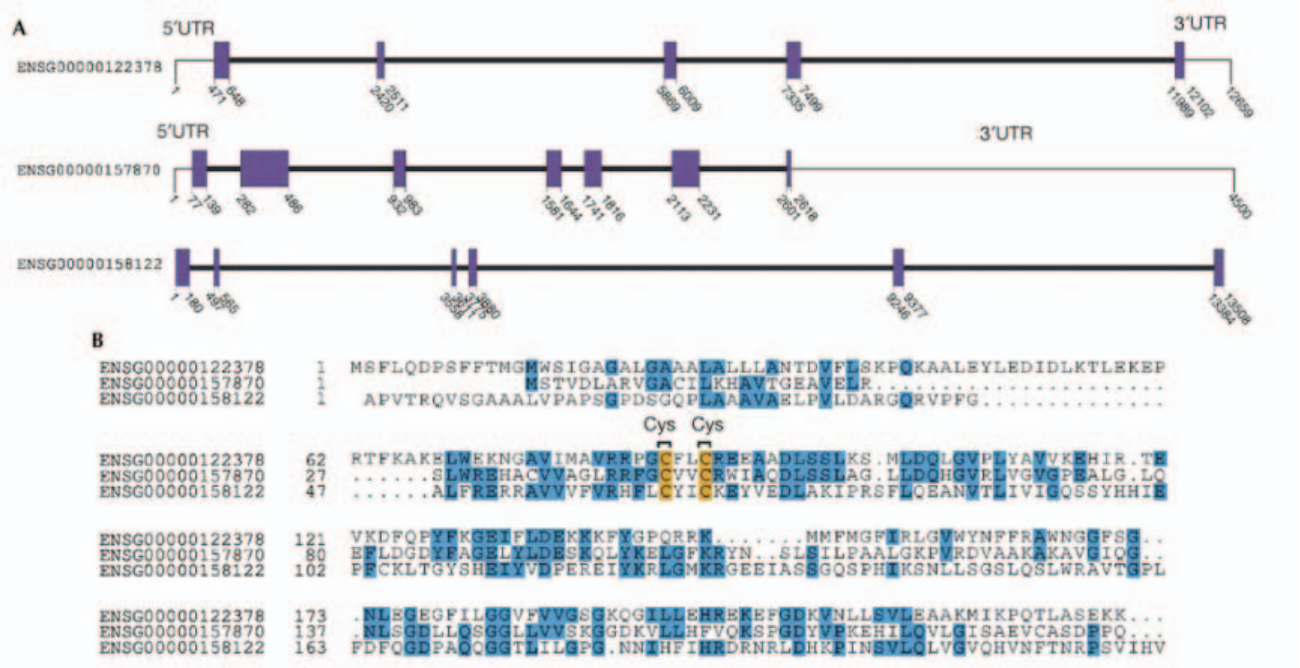

Figure 2. Ensembl human SelU family. (**A**) Gene structure (coding exons) for ENSG00000122378, ENSG00000157870 and ENSG00000158122 genes. (**B**) Alignment of SelU paralogues.

homologue to fugu SelUa) was designed, and *in situ* hybridization was performed on whole zebrafish embryos from different developmental stages. The hybridization sites were revealed by a chromogenic reaction and the expression patterns were analysed. The SelU gene was widely expressed in all embryonic tissues from all stages (Fig 5). Expression was already detectable at the early stages from gastrula and somitogenesis (Fig 5A–C), but within the embryonic tissues only; there was no expression within the nutrient cells of the yolk syncytial layer. Later in development, expression remained high and nonrestricted (Fig 5D–F), demonstrating ubiquitous expression of the SelU gene.

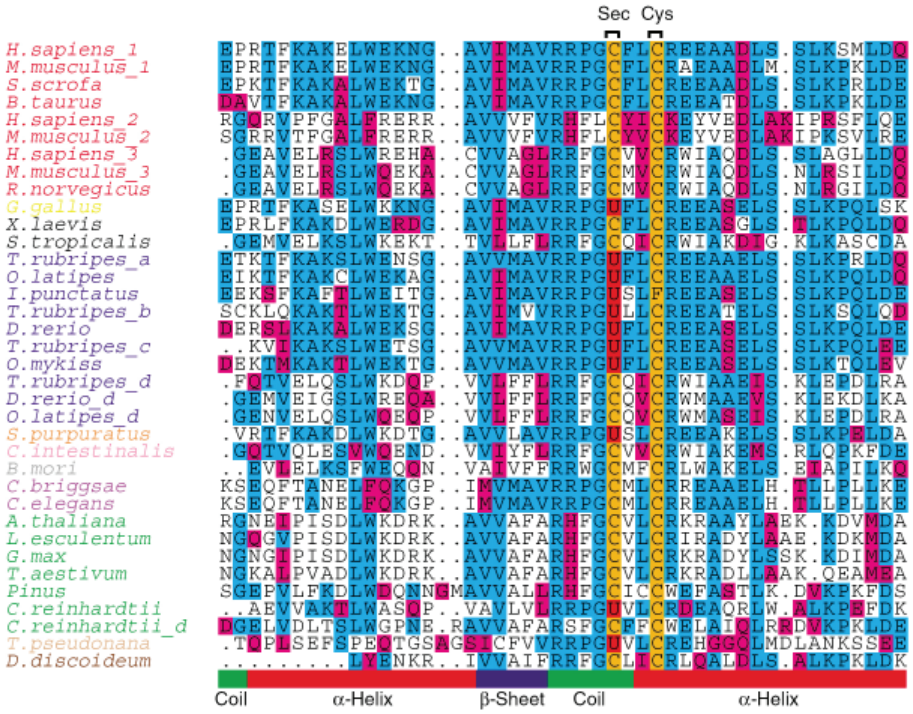

Figure 3. Multiple alignment of SelU proteins across the eukaryotic lineage (the sequence around the Sec (U) amino acid in red and Cys (C) in orange is shown). The sequences are clustered phylogenetically and by sequence similarity. The predicted protein secondary structure is shown at the bottom (also see supplementary information online). Species colours: mammals, red; birds, yellow; amphibians, black; fish, blue; echinoderms, orange; tunicates, pink; arthropods, grey; worms, violet; plants, green; diatoms, light orange; slime molds, brown.

Discussion

A growing body of evidence relates selenium to cancer prevention, immune system function, male fertility, cardiovascular and muscle disorders and prevention and control of the ageing process (Hatfield, 2001). Selenoproteins are thought to be responsible for a majority of these biomedical effects of selenium. To understand the role of selenium in health, the identification and characterization of eukaryotic selenoproteins is thus essential. Despite the increasing availability of eukaryotic genome sequences, the dual role of the UGA codon limits our ability to identify novel selenoproteins. The discovery here of the SelU family shows that comparative genomics could play an important role in overcoming this limitation.

While our comparative method aims at the exhaustive characterization of selenoproteomes, it is certainly unclear how complete is our set of fugu selenoproteins. However, recognition of the majority of known selenoproteins in this organism by this method argues for the identification of all or almost all fugu selenoproteins. In addition, because it assumes no restriction in the SECIS structure, our approach can identify genes with noncanonical SECIS. Although no such elements were found here, they may exist in more divergent lower eukaryotic genomes.

At present, neither sequence database searches nor more specialized motif searches identify similar proteins of known function (data not shown). However, *in situ* hybridization shows ubiquitous expression of SelU in fish embryos (Fig 5), and EST searches also suggest a widespread expression of SelU in human adult tissues (data not shown) pointing to a basic function in the cell.

The SelU family is widely distributed across the eukaryotic lineage, either as Sec- or Cys-containing proteins (Fig 3), but lacks the counterpart in prokaryotes. The scattered and taxa-specific distribution of Sec and Cys forms of a SelU, although common in prokaryotic selenoprotein families, is unexpected in eukaryotes. Besides SelU, other eukaryotic families show an unbalanced distribution, but are constantly present in mammals as true selenoproteins. Therefore, it has been implicitly assumed that mammalian selenoproteins recapitulate the eukaryotic selenoproteome. Our finding challenges this statement and suggests a more discrete distribution of Sec-containing proteins. This hypothesis is reinforced by the recent discovery that methionine-*S*-sulphoxide reductase (MsrA) occurs as a selenoprotein in *Chlamydomonas reinhardtii*, a green algae, but has Cys in vertebrates (including mammals) and other invertebrates (Fu *et al*, 2002; Novoselov *et al*, 2002). Furthermore, a glutathione peroxidase homologue (GPX6) was recently reported to have Sec in humans and pigs, but Cys in rodents (Kryukov *et al*, 2003).

The fact that selenoproteins are distributed discretely at very different taxonomic levels raises the question of whether Sec loss or Sec gain is favoured by evolution. Arguments exist in favour of both possibilities. Replacement of Sec by Cys is plausible because it yields a protein with diminished, but still functional, catalytic activity (Axley *et al*, 1991; Berry *et al*, 1992), and allows an organism to be independent of the supply of the trace element selenium. The fact that a 'fossil' SECIS has been identified in the Cys-containing GPX6 in rodents (Kryukov *et al*, 2003) and in human GPX5 (data not shown) suggests that this event has indeed occurred during evolutionary

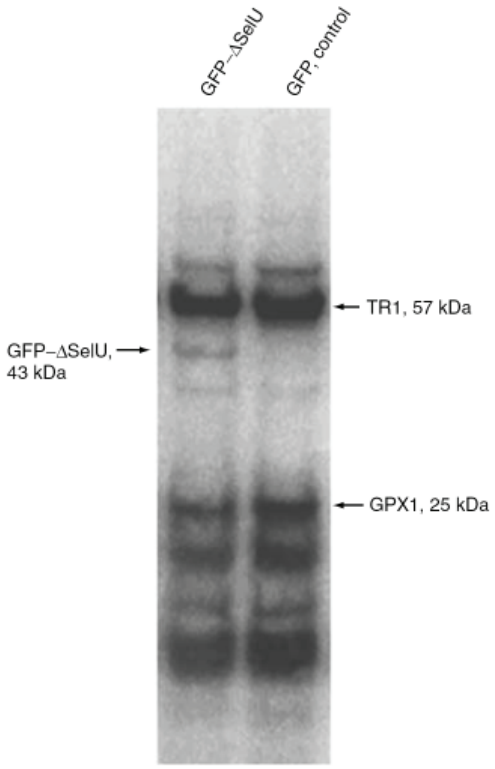

Figure 4. Detection of 75Se-labelled SelU. CV-1 cells were transfected with either GFP–ΔSelU fusion construct (left line) or GFP vector as a control (right line), and grown in the presence of 75Se[selenite] for 24 h. Cell extracts containing 75Se-labelled selenoproteins were resolved by SDS–polyacrylamide gel electrophoresis and visualized with a PhosphorImager System. Locations of major endogenous selenoproteins TR1 (57 kDa) and GPX1 (25 kDa) are shown on the right, and the GFP–ΔSelU fusion protein on the left.

time. In this regard, we searched for vestigial SECIS in human, rodent, amphibian and fish (Cys paralogues) SelU UTRs (see supplementary information, attached or online) with inconclusive results. The conversion in the other direction, a Cys to Sec mutation, is apparently more difficult, since the introduction of an in-frame stop codon must be compensated by the simultaneous emergence of a functional SECIS element in the 3′UTR of the gene. However, gene duplications, the pre-existence of SECIS-like signals, mobile genomic elements, horizontal transfer and the superior catalytic efficiency of Sec could make this process feasible. In any case, it remains to be settled why some organisms prefer Sec, while others prefer Cys-containing forms of orthologous proteins. The presence of SelU Sec and Cys paralogues in fish genomes, however, is suggestive of a particular history for each family and taxa, mediated by an ongoing evolutionary process of Sec/Cys interconversion, in which contingent events could play a role as important as functional constraints.

In any case, if the results obtained here through the analysis of the fugu genome are representative of more divergent eukaryotic genomes, the certain conclusion is that we comprehend today only a fraction of the selenium-dependent world.

Methods

Prediction of selenoproteins in nucleotide sequences. A general scheme is shown in Fig 6. Briefly, for each genome, we predict independently standard and selenoprotein genes, using the standard geneid and a modification that allows the prediction of genes interrupted by in-frame TGA (Castellano *et al*, 2001) (see supplementary information, attached or online).

Protein sequence comparisons: identification of Sec–Sec and Sec–Cys conserved pairs. Proteins predicted in fugu and human are compared using blastp (Altschul *et al*, 1997). Conserved protein sequence alignments with conservation in regions flanking Sec–Sec or Sec–Cys aligned pairs are selected as potential selenoproteins (see supplementary information, attached or online).

Prediction of SECIS in nucleotide sequences. SECIS elements are predicted in selected selenoprotein genes with the SECISearch program (Kryukov *et al*, 2003) (see supplementary information, attached or online).

Metabolic labelling of SelU with ^{75}Se. A 760 bp fragment of chicken SelU cDNA coding for a 16 kDa C-terminal portion and 3′UTR (including the SECIS element) was amplified with AGTGCTCGAGGTGATCATGGCTGTGCGAAGAC and TTATGGATCCGGTTTTGCTCCCCTGGGTAGAC primers and cloned into the *Xho*I/

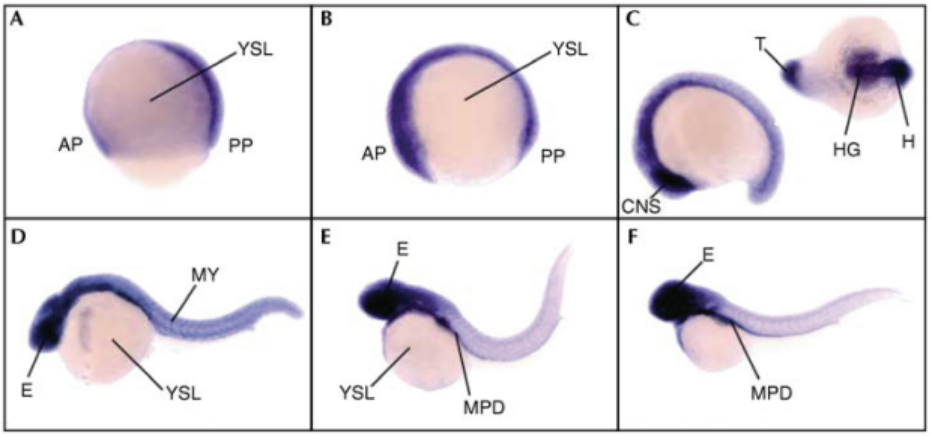

Figure 5. Expression pattern of the SelU gene during development in zebrafish embryos. Developmental stages are (A) gastrula, (B) early somitogenesis, (C) late somitogenesis, (D) 24 h postfertilization, (E) 36 h postfertilization and (F) 48 h postfertilization. All views are lateral except the one in the upper right corner in (C) which is dorsoventral. AP, anterior pole; CNS, central nervous system; E, eye; H, head; HG, hatching gland; MPD, medial part of the pronephric duct; MY, myotomes; PP, posterior pole; T, tail; YSL, yolk syncytial layer.

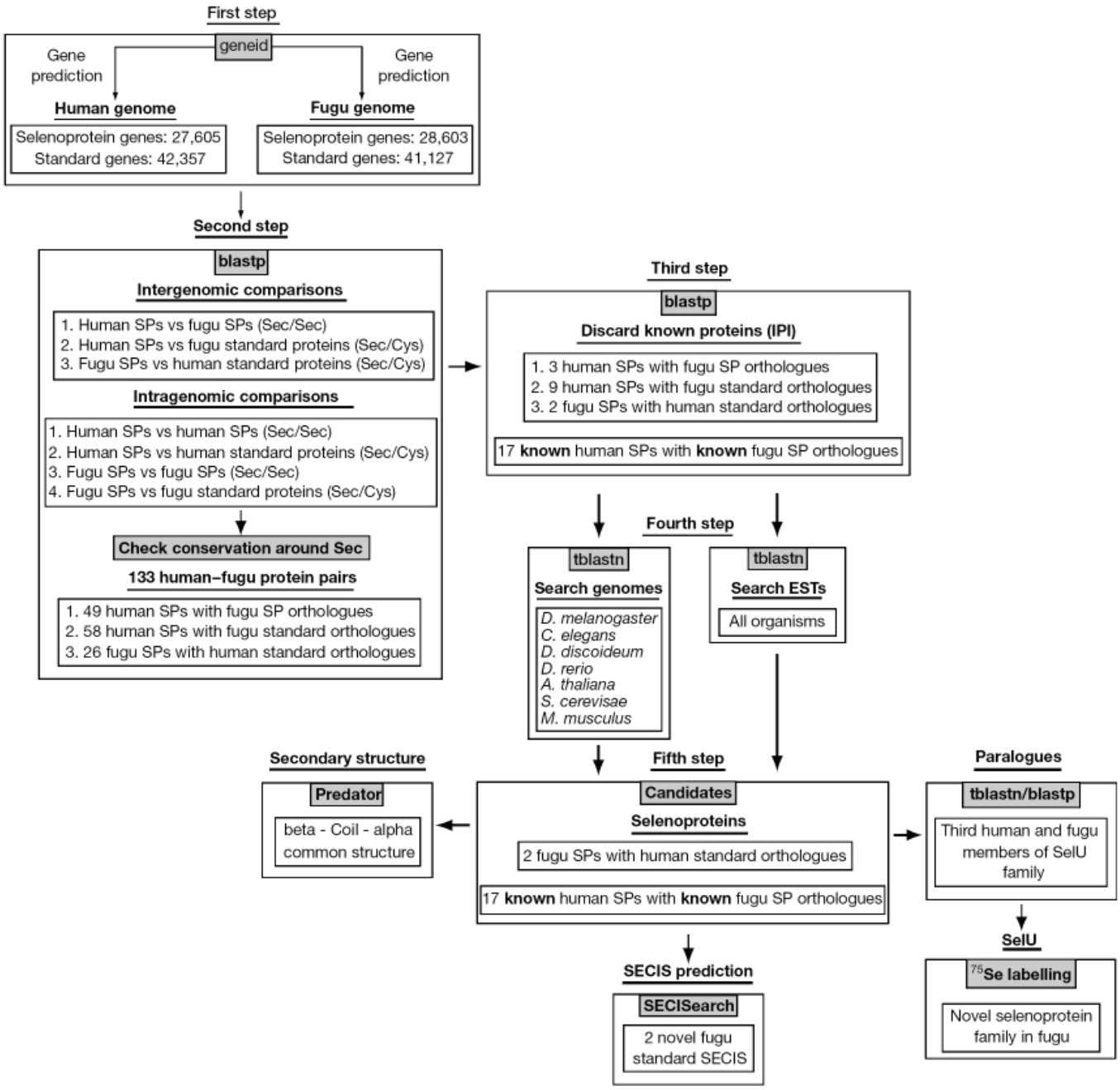

Figure 6. General schema for selenoprotein identification.

*Bam*HI sites of pEGFP-C3 vector (Clontech). CV-1 cells were transfected with either the resulting construct or corresponding vector as a control. In all, 5 µg of DNA and 20 µl of lipofectamine (Invitrogen) were used for transfection of each 60-mm-diameter plate, followed by incubation of cells with 25 µCi ^{75}Se[selenite] (University of Missouri Research Reactor). Samples were analysed on sodium dodecyl sulphate (SDS)–10% NuPAGE gels (Invitrogen). ^{75}Se-labelled proteins were visualized with a Storm PhosphorImager system (Molecular Dynamics). Transfection efficiency was followed by a parallel transfection of cells with a GFP construct. In addition, CV-1 cells were separately transfected with a human SelM construct and labelled with ^{75}Se, which provided a positive control.

***In situ* hybridization.** Eight different zebrafish ESTs, encoding a protein homologous to the fugu SelU protein, were compiled. These EST sequences generated a 1,292 bp contiguous nucleotide sequence encompassing the entire open reading frame and the 3'UTR containing the SECIS motif. A DNA probe complementary to the entire zebrafish SelU cDNA was PCR amplified from an oligo-dT cDNA library (a gift from C. Thisse and B. Thisse) and cloned with compatible restriction sites into pSK(–). Antisense probe synthesis and whole-mount *in situ* hybridization were performed according to Thisse *et al* (1993). The fully detailed protocol is accessible at http://zfin.org/zf_info/zfbook/chapt9/9.82.html. Specificity was assessed using antisense and other irrelevant probes (data not shown).

Data and software availability. Sequence data and software can be found at http://genome.imim.es/databases/spfugu2004

Supplementary information is attached. It is also available at *EMBO reports* online (http://www.nature.com/embor/journal/vaop/ncurrent/extref/7400036s1.pdf).

Acknowledgments

We thank the referees for helpful suggestions and J.F. Abril for technical assistance. S. Obrecht-Pflumio, C. Thisse and B. Thisse are gratefully thanked for technical expertise with *in situ* hybridization. S.C. is the recipient of a predoctoral fellowship from Generalitat de Catalunya. This work was supported by grant BIO2000-1358-C02-02 from Ministerio de Educación y Ciencia (Spain) to R.G. and by NIH grant GM061603 to V.N.G.

References

Abril JF, Guigó R (2000) gff2ps: visualizing genomic annotations. Bioinformatics 16: 743–744.

Altschul SF, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.

Axley MJ, Böck A, Stadman TC (1991) Catalytic properties of an *Escherichia coli* formate dehydrogenase mutant in which sulfur replaces selenium. Proc Natl Acad Sci USA 88: 8450–8454.

Berry MJ, Mai AL, Kieffer J, Harney JW, Larsen P (1992) Substitution of cysteine for selenocysteine in type i iodothyronine deiodinase reduces the catalytic efficiency of the protein but enhances its translation. Endocrinology 131: 1848–1852.

Castellano S, Morozova N, Morey M, Berry MJ, Serras F, Corominas M, Guigó R (2001) *In silico* identification of novel selenoproteins in the *Drosophila melanogaster* genome. EMBO Rep 2: 697–702.

Fu LH, Wang XF, Eyal Y, She YM, Donald LJ, Standing KG, Ben-Hayyim G (2002) A selenoprotein in the plant kingdom. Mass spectrometry confirms that an opal codon (UGA) encodes selenocysteine in *Chlamydomonas reinhardtii* glutathione peroxidase. J Biol Chem 277: 25983–25991.

Grundner-Culemann E, Martin GW III, Harney JW, Berry MJ (1999) Two distinct SECIS structures capable of directing selenocysteine incorporation in eukaryotes. RNA 5: 625–635.

Guigó R, Knudsen S, Drake N, Smith TF (1992) Prediction of gene structure. J Mol Biol 226: 141–157.

Hatfield DL (ed.) (2001) Selenium: Its Molecular Biology and Role in Human Health. Dordrecht: Kluwer Academic Publishers.

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.

Kryukov GV, Castellano S, Novoselov SV, Lobanov AV, Zehtab O, Guigó R, Gladyshev VN (2003) Characterization of mammalian selenoproteomes. Science 300: 1439–1443.

Lescure A, Gautheret D, Carbon P, Krol A (1999) Novel selenoproteins identified *in silico* and *in vivo* by using a conserved RNA structural motif. J Biol Chem 274: 38147–38154.

Martin-Romero FJ, Kryukov GV, Lobanov AV, Carlson BA, Lee BJ, Gladyshev VN, Hatfield DL (2001) Selenium metabolism in *Drosophila*: selenoproteins, selenoprotein mRNA expression, fertility, and mortality. J Biol Chem 276: 29798–29804.

Novoselov SV, Rao M, Onoshko NV, Zhi H, Kryukov GV, Xiang Y, Weeks DP, Hatfield DL, Gladyshev VN (2002) Selenoproteins and selenocysteine insertion system in the model plant cell system, *Chlamydomonas reinhardtii*. EMBO J 21: 3681–3693.

Parra G, Blanco E, Guigó R (2000) Geneid in *Drosophila*. Genome Res 10: 511–515.

Thisse C, Thisse B, Schilling TF, Postlethwait JH (1993) Structure of the zebrafish snail1 gene and its expression in wild-type, spadetail and no tail mutant embryos. Development 119: 1203–1215.

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL_W: improving the sensitivity of progressive sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680.

Walczak R, Carbon P, Krol A (1998) An essential non-Watson–Crick base pair motif in 3'UTR to mediate selenoprotein translation. RNA 4: 74–84.

Supplementary data

Sergi Castellano[1], Sergey V. Novoselov[2], Gregory V. Kryukov[2],
Alain Lescure[3], Enrique Blanco[1], Alain Krol[3], Vadim N. Gladyshev[2] and Roderic Guigó[1,4+]

[1] Grup de Recerca en Informàtica Biomèdica. Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, Dr. Aiguader 80, 08003 Barcelona, Catalonia, Spain
[2] Department of Biochemistry, University of Nebraska, Lincoln NE 65588, USA
[3] Institut de Biologie Moléculaire et Cellulaire, 15 Rue René Descartes, 67084 Strasbourg Cedex, France
[4] Programa de Bioinformàtica i Genòmica, Centre de Regulació Genòmica, Barcelona, Catalonia, Spain

[+] Corresponding author. Tel: +34 93 224 0877; Fax: +34 93 224 0875; E-mail: rguigo@imim.es

RESULTS

SECIS sequences are divided into standard structural units.

Gallus gallus (chicken) SECIS (TIGR ID: TC4619)

CCUUUUGUGUCUG ACUGUAUUAA UGAA AGGCUGGGUC UAAAAUCU GACGUACCCUGGAU GUUUU CAGUCAGAGACAGUCGG

Danio rerio (zebrafish) SECIS (TIGR ID: TC76454)

GUGUUUAAUGGUGUGU GUAUUAA UGAU AGUCUGACUC CAAACUCAGUGUAGAAAG AGCAGAUUUGAU GUCA ACACAUGCUUAUAUAC

METHODS

Gene prediction

`geneid` is a program to predict protein coding genes in anonymous eukaryotic sequences designed with a hierarchical structure (see Parra *et al*. 2000, and the `geneid` documentation at http://genome.imim.es/geneid for details).

Basically, gene prediction involves three main steps:

1) **prediction of sites.** That is, start (ATG), stop (TAA,TAG and TGA) and splice signals (GT and AG) that define potential exon boundaries. When predicting selenoproteins the TGA site is allowed two contrasting meanings, stop and selenocysteine codon (Castellano *et al*.,2001). Position Specific Scoring Matrices are used to predict splice sites and start codons. Thus, predicted sites are scored as the log-likelihood ratio of the site sequence under the site model and under the random model.

2) **prediction of coding exons.** `geneid` builds all possible exons compatible with the predicted sites and scores them according to the scores of the exon defining sites and to a coding potential function. The coding function reflects the species-specific bias in the usage of codons in protein coding regions. In `geneid`, a Markov Model of order five trained in known species-specific coding exons is used. These models have been typically applied to discriminate coding from non coding regions (Borodovsky and McIninch, 1993; Guigó, 1999).

   We had previously shown that the region comprised between the in-frame TGA codon and the stop codon in selenoproteins bears the codon bias characteristic of protein coding regions, whereas the region comprised between the stop codon TGA, and the next stop codon in-frame in non-selenoproteins do not castellano: 2001a, as otherwise expected. Therefore, coding potential is in general much higher in selenoproteins than in no selenoproteins in this region, and this value can be used to distinguish between actual selenoproteins and false positive predictions.

3) assembly of genes. From the set of predicted exons, `geneid` assembles the gene structure that maximizes the sum of the scores of the assembled exons. When assembling gene structures, `geneid` can take into account additional information about gene elements along the sequence. This information is provided externally, and may include previous knowledge about coding regions, or predictions obtained by other programs. It is in this way, that predicted SECIS elements can be introduced into gene predictions (Castellano *et al*., 2001)

On the other hand, to be assembled into a gene structure, predicted exons and other genomic elements provided to `geneid` must conform to a number of user-defined biological constraints, such as frame compatibility, minimum and maximum distance between consecutive elements, and the order in which different genomic elements can be chained. All this rules are stated in the gene model, which is specified externally. When predicting selenoproteins the model may specify that predicted genes with TGA in-frame interrupted exons are only allowed when a suitable SECIS element has been predicted within a given range of nucleotides of the predicted gene stop codon (Castellano *et al*., 2001).

Prediction of standard genes in the human and fugu genomes

Gene structure prediction using `geneid` was done in the human and fugu genomes to predict standard genes.

Human genome

`geneid` was ran on the August 6, 2001 Golden Path assembly (release hg8) of the *Homo sapiens* genome (http://genome.cse.ucsc.edu/). 42357 genes were predicted.

Fugu genome

`geneid` was ran on the October 25, 2001 Joint Genome Institute (JGI, release 1.0) assembly of the *Takifugu rubripes* genome (http://www.jgi.doe.gov/). This initial assembly provides short contigs, but the gene compactness of the fugu genome makes gene prediction feasible. 41127 genes were predicted.

Prediction of selenoprotein genes in the human and fugu genomes

As indicated above, we have modified slightly `geneid` in order to include the possibility of predicting selenoproteins. Essentially, the codon TGA can be understood both as stop and selenocysteine codon when building exons. Therefore, `geneid` is able to predict, at the same time, both standard genes and selenoprotein genes.

In contrast to the method presented in (Castellano *et al*., 2001), where candidate selenoprotein genes were predicted only when a suitable SECIS prediction was present at the appropiate downstream distance, here we introduce a SECIS independent gene prediction approach. Potential selenoprotein gene candidates are predicted regardless of the presence of a downstream SECIS structure. Gene predictions interrupted by in-frame TGA codons, are likely to occur only when the strong coding bias characteristic of coding regions is present across the in-frame TGA codon. However, SECIS independent selenoprotein prediction results in an overwhelming number of selenoprotein candidates, due to the additional number of exons predicted (those that contain a TGA in-frame), which decrease accuracy of final gene structures. Consequently, in the approach presented here, a different biological contraint is used. A comparative protocol is followed, in such a way, that homology assessments at the protein level (see below) take place of SECIS restriction.

Known selenoproteins: human and fugu genomes

Known selenoprotein genes were mapped in both, human and fugu genomes through BLAT (http://genome.cse.ucsc.edu/) and BLAST (Altschul *et al*., 1997) searches.

23 known human selenoprotein genes belonging to 15 different families (known at that time) were mapped onto the human genome. The modified `geneid` version was used to predict them and sensitivity of the program was assessed. 20 out of 23 selenoprotein genes were properly predicted. Only SelK, SelT and SelS genes were not predicted as selenoproteins.

22 known fugu selenoprotein genes belonging to 14 different families were mapped onto the fugu genome (SelW gene was not found in this genome). The modified `geneid` version was used to predict them and

sensitivity of the program was assessed. 18 out of 22 selenoprotein genes were properly predicted. Only SelK, SelH, SelS and SelM genes were not predicted as selenoproteins.

In conclusion, 1) both genomes, as shown by the mapping of all but one fugu selenoprotein gene, are complete enough to run a gene prediction program on them; and 2) the modified `geneid` program is able to predict most selenoprotein genes without the SECIS constraint. Sensitivity (that is, predicting only as non-selenoprotein genes non-selenoprotein genes. Sn >80% in both genomes) is sufficient to make reasonable the prediction of novel selenoprotein genes in the human and fugu genomes.

In addition, the same seventeen (out of 22 common selenoprotein genes mapped on both genomes. Sn >75%) are properly predicted in the two genomes. This fact, makes also reasonable the asumption of, by means of a comparative approach between genomes, true selenoprotein genes can be pinpoint from false positive predictions.

Potential selenoproteins: human genome

The modified version of `geneid` able to predict TGA in-frame genes was run on the August 25, 2001 Golden Path assembly of the *H. sapiens* genome. 27605 selenoprotein genes and 21603 standard genes were predicted. The modified version of `geneid` yields, in a single gene prediction, standard genes and potential TGA in-frame genes. This set of standard genes was discarded because gene structures are more reliably retrieved from standard `geneid` (see Prediction of standard genes in the human and fugu genomes) and selenoprotein gene prediction is intended only to provide genes bearing a TGA in-frame.

On the other hand, the set of potential selenoprotein genes is, in number, more than half of the total standard genes predicted by the standard `geneid` program. In other words, specificity (that is, predicting as selenoproteins only real selenoproteins) of the modified version of `geneid` able to predict TGA in-frame genes is extremely low at the level of sensitivity demanded (see above). Reasons for this are 1) coding potential, despite higher and positive in coding open reading frames (ORFs), can not discriminate as well when admitting a stop codon (TGA) in-frame. Many genes add short ORFs after a real stop codon (TGA), having that untranslated regio a low, but positive, coding potential; and 2) `geneid` parameters of the modified version, are slightly bias to include TGA in-frame exons. In this way, and because our aim

is finding novel selenoprotein families, we minimize the chance of missing yet unknown selenoproteins by overpredicting them. False positive predictions are removed at later stages (see below).

Potential selenoproteins: fugu genome

A modified version of `geneid` able to predict TGA in-frame genes was run on the October 25, 2001 JGI assembly of the *T. rubripes* genome (http: //www.jgi.doe.gov/). 28603 selenoprotein genes and 4523 standard genes were predicted. Same considerations, as for gene prediction in the human genome, apply to gene prediction in the fugu genome (see above).

Comparison of human and fugu standard protein and selenoprotein sets

Selenoprotein families can have cysteine-homologs in the same or different genomes, but the Sec/Cys pattern for novel selenoproteins is unknown. Distribution of homologs can help to pinpoint selenoproteins and, in consequence, we introduced a protocol to predict and compare both types of genes.

Given human and fugu selenoprotein and standard gene complements we do the following set of intra and inter-genomic comparisons, at the protein level with `blastp` (query sequences were not filtered for low compositional complexity and a expectation value of 1e-10 was used. Stop codons in BLOSUM62 matrix were treated as cysteines), to reproduce possible Sec/Cys distribution patterns:

1. Inter-genomic comparisons
   (a) Predicted human selenoproteins against predicted fugu selenoproteins (Sec/Sec)
   (b) Predicted human selenoproteins against predicted fugu standard genes (Sec/Cys)
   (c) Predicted fugu selenoproteins against predicted human standard genes (Sec/Cys)

2. Intra-genomic comparisons
   (a) Predicted human selenoproteins against predicted human selenoproteins (Sec/Sec)
   (b) Predicted human selenoproteins against predicted human standard genes (Sec/Cys)
   (c) Predicted fugu selenoproteins against predicted fugu selenoproteins (Sec/Sec)
   (d) Predicted fugu selenoproteins against predicted fugu standard genes (Sec/Cys)

However, these two types of comparisons (inter and intra-genomic), are not processed in the same way. First and separately for each predicted human and fugu selenoprotein (27605 human and 28603 fugu proteins), all possible inter-genomic comparisons are computed to define potential selenoprotein pairs having selenocysteine in either human, fugu or, alternatively, in both genomes. The result is a collection (subset of initial human and fugu predicted selenoproteins) of individual human and fugu potential selenoproteins with orthology support. Some cases having only Sec-Sec support, some others having only Sec-Cys and the rest both of them. Second, and once putative ortholog pairs have already been selected, paralogy data, if exist, is included for each of them (previously calculated from intra-genomic comparisons). In this way, and because paralogy is not as informative as orthology (see below), potential selenoprotein orthologs between human and fugu define pairs of putative selenoprotein families, and paralogs add additional support to them.

The rational behind this approach is that intra-genomic comparisons are false positive prone. Because of genome organization, where genes duplicate and may conserve sequence and gene structure, a false positive prediction in a genome (that is a gene with an incorrect TGA in-frame) may appear several times. Posterior comparisons would regard this gene as a potential selenoprotein family. However, this contingency is much more unlikely between genomes. The TGA (which is a false codon for Sec) may not be conserved and, at the same time, coding potential may be different (which can make that exon not to be included into predicted gene structure).

This procedure is consistent with the fact that human and fugu have all known selenoprotein families in Sec or Cys form. Therefore, we expect to predict a potential selenoprotein or cysteine homolog gene in both genomes and, at the same time, we use paralog information (too noisy by itself). Finally, human and fugu uniq selenoproteins, that have been treated independently up to now, are collapsed when define the same human-fugu or fugu-human pair (that is, query and subject are the same but inverted).

Results were the following, 1) 368 human selenoprotein - fugu selenoprotein pairs (including 17 known human-fugu selenoprotein pairs); 2) 296 human selenoprotein - fugu cysteine homolog

pairs; and 3) 216 fugu selenoprotein - human cysteine homolog pairs. Note that Sec-Sec pairs may also have Sec-Cys homologs, though are included only in the Sec-Sec division.

3. Conservation around the selenocysteine amino acid

   Selected ortholog pairs were further analyzed to assess protein sequence conservation around the selenocysteine amino acid. A block of 20 amino acids (10 at each side of the Sec residue aligned to either Sec or Cys) was checked for having at least 4 similar residues (according to BLOSUM62 matrix) on both parts. In order to gain sensitivity, when there were less than 10 residues on one, or both, sides the conservation assessment was skiped on that side(s). When applied, all known human and fugu selenoprotein pairs were recovered.

The results of this filtering step were the following, 1) 49 human selenoprotein - fugu selenoprotein pairs (including 17 known human-fugu selenoprotein pairs); 2) 58 human selenoprotein - fugu cysteine homolog pairs; and 3) 26 fugu selenoprotein - human cysteine homolog pairs.

Search for homologs

In order to further validate the resulting human-fugu pairs, we undertook an exhaustive search against a number of databases of known coding sequences (proteins and ESTs) and several partial and full-length genomes. This approach should elicit real selenoprotein genes along with their Sec/Cys eukaryotic distribution. Each human and fugu selenoprotein member of potential pairs was studied.

International Protein Index

The International Protein Index (IPI, human version 2.0) (http://www.ebi.ac.uk/IPI/) is a protein database that provides a minimally redundant yet maximally complete set of human genes and proteins. IPI is assembled from human protein sequence information taken from the following 5 data sources: 1) SWISS-PROT; 2) TrEMBL; 3) Ensembl (http://www.ensembl.org); 4) RefSeq NPs; and 5) RefSeq XPs. This database was used to discard sequences highly similar to known proteins with functions apparently unrelated to those of selenoproteins.

In this way, blast searches against the IPI database narrowed the number of potential pairs, that is containing unknown proteins, to 1) 21 human selenoprotein - fugu selenoprotein pairs (including 17 known human-fugu selenoprotein pairs); 2) 9 human selenoprotein - fugu cysteine homolog pairs; and 3) 2 fugu selenoprotein - human cysteine homolog pairs.

Genomes

The following completely sequenced genomes from 1) *Drosophila melanogaster*; 2) *Caenorhabditis elegans*; 3) *Saccharomices cerevisae*; 4) *Schizosaccharomyces pombe*; 5) *Plasmodium falciparum*; and 6) *Arabidopsis thaliana* were queried by TBLASTN to identify sequences with homology in TGA-flanking region, containing either TGA (Sec codon) or TGT or TGC (Cys codons) in place of TGA. BLASTP searches against proteins annotated in these genomes were also carried out to identify cysteine-containing homologs. At the same time, partial sequenced genomes from 1) *Mus musculus*; 2) *Xenopus laevis*; 3) *Danio rerio*; 4) *Dictyostelium discoideum*; and 5) *Chlamydomonas reinhardtii* were also screened in the same way. These searches, allowed screening for new homolog sequences and reconstruction of Sec/Cys distribution across the eukaryotic lineage.

ESTs

NCBI EST database (dbEST, build of April 15, 2002) was queried to 1) check consistency of human and fugu genomic sequence at the Sec/Cys region; and 2) search for novel homologs for members of the 14 potential selenoprotein pairs and 3) define Sec/Cys distribution across the eukaryotic lineage.

Blast searches against dbEST discarded pairs with either 1) predicted gene structure incompatible with the exonic structure of identical EST sequences; or 2) TGA selenocysteine codon not supported by corresponding EST sequences, therefore, presumedly a genomic sequence error. This filtering step, apart from known human and fugu selenoproteins, resulted in two pairs containing both fugu selenoproteins and human cysteine homologs.

On the other hand, several Sec and Cys-containing SelU homologs were found (see below).

cDNAs

The TIGR collection of transcripts (cDNAs and ESTs, http://www.tigr.org) was screened to search for SelU orthologs. In this way, a cysteine-containing homolog was found for zebrafish (*Danio rerio*, TC173888) and japanese medaka (*Oryzias latipes*, TC21944).

Paralogs

The four sequences of the predicted two pairs, accounting for two fugu selenoproteins and two human cysteine homologs, were globally aligned with `clustalw` (Thompson *et al*., 1994). Their alignment clearly showed that, on basis of sequence similarity, they belong to the same protein family. This fact reinforced the likelihood of them belonging to a real selenoprotein family.

On the other hand, further `TBLASTN` searches were done against the human and fugu genomes to unveil unpredicted paralogous sequences. `BLASTP` searches against annotated proteins in these genomes were also accomplished. An additional fugu selenoprotein member of the SelU family and a human cysteine-homolog belonging also to this familiy were found.

Search for prokaryotic homologs

Fugu SelUa and human ENSG00000122378 proteins were blasted against 246 bacterial and 18 archaeal genomes available at NCBI (http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi). `TBLASTN` and `BLASTP` programs, against proteins from 177 annotated genomes, were used. No significant hits were found.

SelU distribution across the eukaryotic lineage

Searches above yielded SelU homologs all across the eukaryotic lineage. They can be divided into (common name given when known):

Sec-containing homologs were found in:

Fish: fugu (*Takifugu rubripes*), zebrafish (*Danio rerio*), japanese medaka (Oryzias latipes), catfish (*I. punctatus*), rainbow trout (*Oncorhynchus mykiss*), carp (*Cyprinus carpio*), three spined stickleback (*Gasterosteus aculeatus*)
Birds: chicken (Gallus gallus)
Echinoderms: sea urchin (*Strongylocentrotus purpuratus*)
Green algae: *Chlamydomonas reinhardtii*
Diatoms: *Thalasiosira pseudonana*

Cys-containing homologs were found in:
Mammals: human (*Homo sapiens*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), pig (*Sus scrofa*), cow (*Bos taurus*), dog (*Canis canis*), rabbit (*Oryctolagus cuniculus*).
Fish: fugu (*Takifugu rubripes*), zebrafish (*Danio rerio*), japanese medaka (Oryzias latipes)
Amphibians: frog (*Xenopus laevis*), frog (*Silurana tropicalis*)
Tunicates: *Ciona intestinalis*
Arthropods (insects): silkworm (*Bombix mori*)
Nematodes: *Caenorhabditis elegans*, *Caenorhabditis briggsae*, *Ancylostoma ceylanicum*, *Parastrongyloides trichosuri*, *Strongyloides stercoralis*, *Pristionchus pacificus, Toxocara canis*
Land plants: sweet orange (*Citrus sinensis*), barrel medic (*Medicago truncatula*), cabernet sauvignon (*Vitis vinifera*), sunflower (*Helianthus annuus*), barley (*Hordeum vulgare*), onion (*Allium cepa*), rape (Brassica napus), european aspen (*Populus tremula*), pepper (*Capsicum annuum*), sorghum (Sorghum bicolor)
Green algae: *Chlamydomonas reinhardtii*
Slime molds: *Dictyostelium discoideum*

Arg-containing homologs were found in:
Nematodes: *Strongyloides ratti*

No homologs were found in (complete genome sequence):
Arthropods (insects): fly (*Drosophila melanogaster*), mosquito (*Anopheles gambiae*)
Yeast: baker's yeast (*Saccharomices cerevisae*), fission's yeast (*Schizosaccharomyces pombe*)

Apicomplexa: malaria parasite (*Plasmodium falciparum*)

Prediction of protein secondary structure

The crystal structure of an eukaryotic selenocysteine, the bovine glutathione peroxidase, has been resolved at 0.2 nm resolution (Epp *et al*., 1983). The catalytic site of this enzyme is characterized by a beta-sheet—turn—alpha-helix structural motif, with the selenocysteine residue lying within the turn. Secondary structure predictions around the selenocysteine residue of most known selenoproteins, obtained using the program `Predator` (Frishman and Argos, 1997; Castellano *et al*., 2001), essentially conformed to this structure (data not shown). Fugu SelU selenoproteins also stick to this pattern when predicted with the `Predator` program.

Prediction of SECIS elements

SECIS elements were predicted in selected selenoprotein genes with the `SECISearch` program (Kryukov *et al.*, 2003). This program is available as a web server resource at http://genome.unl.edu/SECISearch.html. Given that predictions are only done in short genomic regions, false positive are not a concern, therefore a loose SECIS pattern can be used to permit identification of SECIS variants. The whole range of SECIS patterns provided by SECISearch were used. However, only canonical SECIS were found in *T.rubripes* (fugu, puffer fish), *D. rerio* (zebrafish) and *G.gallus* (chicken).

Search for fossil SECIS

Annotated UTR regions were extracted from Ensembl (www.ensembl.org) for human, mouse and rat SelU homologs. The IDs for the three sets of SelU orthologous genes are: 1) ENSG00000122378, ENSMUSG00000021792, ENSRNOG00000011140; 2) ENSG00000157870, ENSMUSG00000029059, ENSRNOG00000013468; 3) ENSG00000158122, ENSMUSG00000021482, ENSRNOG00000018886. However, most of these annotated UTRs were uncomplete. Possibly, because of the lack of EST sequences. In addition, UTR regions for SelU Cys-homologs from *Takifugu rubripes*, *Danio rerio*, *Oryzias latipes*, *Xenopus laevis*, *Ciona intestinalis*, *Caenorhabditis elegans*, *Caenorhabditis briggsae* and *Dictyostelium discoideum* were extracted from the TIGR collection of transcripts (cDNAs and ESTs, http://www.tigr.org) and, if needed, from the original genomic sequence.

In these UTR regions two analysis were performed:

1. Fish and chicken SECIS sequences were blasted against these UTRs in the search for similarity. No significant hits were found. However, while SECIS elements share a high degree of sequence identity among mammals (Kryukov *et al.*, 2003), this is not necessarily the case for functional and vestigial SECIS between, for example, fish, chicken and mammalian SECIS.
2. SECISearch was run on these UTRs with canonical and non-canonical patterns. No hits were found. Furthermore, the program PatScan (Dsouza *et al.*, 1997) was used to run even more degenerated patterns. However, matches were unclear. Specially, because no similar hits were found between human and rodent UTRs.

In any case, the lack of a potential fossil SECIS does not yet discard the hypothesis of a Sec to Cys mutation, because the UTRs under study could have accumulated enough mutations to fade the SECIS phylogenetic signal.

In addition, SECIS similarity searches were run on the whole TIGR collection of transcripts (cDNAs and ESTs, http://www.tigr.org). The rational behind this was, again, to find vestigial SECIS elements through sequence similarity. In the hope that they are still recognizable, that is, change from Sec to Cys is either quite recent or the mutation rate is low enough, we could expect still some phylogenetic footprint. However, because even functional SECIS diverge, a negative result is likely and, at the same time, inconclusive respect to clarify evolutionary events. Searches were done on the Eukaryotic Gene Ortholog (EGO) database at TIGR. It is a collection of partial and full length cDNAs from 61 different eukaryotic organisms. Again, results were not convincing.

References

Altschul, S. F., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Borodovsky, M. and McIninch J. (1993) GenMark: Parallel gene recognition for both DNA strands. *Computer and Chemistry*, **17**, 123-134.

Castellano, S., Morozova, N., Morey, M., Berry, M. J., Serras, F., Corominas, M., and Guigó, R. (2001) *in silico* identification of novel selenoproteins in the *Drosophila melanogaster* genome. *EMBO reports*, **2**, 697–702.

Dsouza, M., Larsen, N. and Overweek, R. (1997) Searching for patterns in genomic data. *Trends Genet.*, **13**, 497-498.

Epp, O., Ladenstein, R., and Wendel, A. (1983) The refined structures of the selenoenzyme glutathione peroxidase at 0.2-nm resolution. *Eur. J. Biochem.*, **133**, 51.

Frishman, D. and Argos, P. (1997) Seventy-five percent accuracy in protein secondary structure prediction. *Proteins*, **27**, 329–335.

Guigó, R. (1999) DNA composition, codon usage and exon prediction. In Bishop, M., editor, *Genetic Databases*, pages 53–80. Academic Press, San Diego, California.

Kryukov, G.V., Castellano, S., Novoselov, S.V., Lobanov, A.V., Zehtab, O., Guigó, R. and Gladyshev V.N. (2003) Characterization of Mammalian Selenoproteomes. *Science*, **300**, 1439-1443.

Parra, G., Blanco, E., and Guigó, R. (2000) Geneid in *Drosophila*. *Genome Research*, **10**, 511–515.

Thompson, J., Higgins, D., and Gibson, T. (1994) CLUSTAL_W: improving the sensitivity of progressive sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic. Acids Res.*, **22**, 4673–4680.