May 2003

Characterization of Mammalian Selenoproteomes

Gregory V. Kryukov
University of Nebraska-Lincoln

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

Sergey V. Novoselov
University of Nebraska-Lincoln

Alexey V. Lobanov
University of Nebraska-Lincoln

Omid Zehtab
University of Nebraska-Lincoln

See next page for additional authors

Characterization of Mammalian Selenoproteomes

Gregory V. Kryukov,[1] Sergi Castellano,[2] Sergey V. Novoselov,[1] Alexey V. Lobanov,[1] Omid Zehtab,[1] Roderic Guigó,[2] Vadim N. Gladyshev[1*]

[1] Department of Biochemistry, University of Nebraska, Lincoln, NE 68588–0664, USA

[2] Grup de Recerca en Informàtica Biomèdica, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, Centre de Regulació Genòmica, Doctor Aiguader 80, 08003 Barcelona, Catalonia, Spain

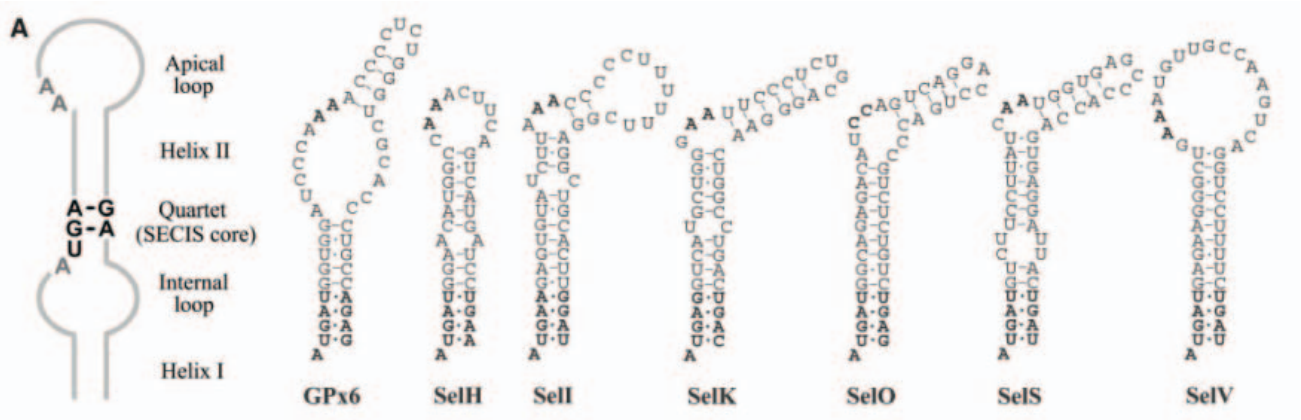
[*] Corresponding author. E-mail: vgladyshev1@unl.edu

In the genetic code, UGA serves as a stop signal and a selenocysteine codon, but no computational methods for identifying its coding function are available. Consequently, most selenoprotein genes are misannotated. We identified selenoprotein genes in sequenced mammalian genomes by methods that rely on identification of selenocysteine insertion RNA structures, the coding potential of UGA codons, and the presence of cysteine-containing homologs. The human selenoproteome consists of 25 selenoproteins.

Submitted February 14, 2003; accepted for publication April 24, 2003.

In the universal genetic code, 61 codons encode 20 amino acids, and 3 codons are terminators. However, the UGA codon has a dual function in that it signals both the termination of protein synthesis and incorporation of the amino acid selenocysteine (Sec) (*1–3*). Available computational tools lack the ability to correctly assign UGA function. Consequently, there are numerous examples of misinterpretations of UGA codons as both Sec codons (*4*) and terminators (*5, 6*), including annotations of the human genome (*7, 8*), where no selenoproteins have been correctly predicted. With 18 human selenoprotein genes previously discovered (*3*), the estimates of the actual number of such genes vary greatly (*9*). All previously characterized selenoproteins except selenoprotein P (*10*) contain single Sec residues that are located in enzyme-active sites and are essential for their activity. Thus, misidentification of UGA codons leads to a loss of crucial biological and functional information. Sec is cotranslationally incorporated into nascent polypeptides in response to UGA codons when a specific stem-loop structure, designated the Sec insertion sequence (SECIS) element, is present in the 3' untranslated regions (UTRs) in eukaryotes and in archaea, or immediately downstream of UGA in bacteria (*1, 11–13*). Trans-acting factors, including Sec tRNA, Sec-specific elongation factor, selenophosphate synthetase (SPS), Sec synthase, and a SECIS-binding protein, are also required for Sec biosynthesis and insertion (*1, 3, 13–15*). Most known selenoprotein genes have homologs, in which Sec is replaced with cysteine (Cys). However, these proteins are poor catalysts as compared with selenoproteins (*3*).

We hypothesized that the UGA dual-function problem could be solved by identifying selenoprotein genes in sequenced genomes and assigning terminator functions to the remaining in-frame UGAs. The requirement of SECIS elements for Sec insertion and the presence of Cys-containing homologs of selenoproteins suggested two independent bioinformatics methods for selenoprotein identification. In addition, we used an observation that the strong codon bias characteristic of protein-coding



B

Selenoprotein	Chromosomal location (number of exons)	Sec location in protein (length of protein)	Selenoprotein structure
15kDa	1p22.3 (5)	93 (162)	
DI1	1p32.3 (4)	126 (249)	
DI2	14q31.1 (2)	133 (265)	
DI3	14q32	144 (278)	
GPx1	3p21.31 (2)	47 (201)	
GPx2	14q23.3 (2)	40 (190)	
GPx3	5q33.1 (5)	73 (226)	
GPx4	19p13.3 (7)	73 (197)	
GPx6	**6p22.1 (5)**	**73 (221)**	
H	**11q12.1 (4)**	**44 (122)**	
I	**2p23.3 (10)**	**387 (397)**	
K	**3p21.31 (5)**	**92 (94)**	
M	22q12.2 (5)	48 (145)	
N	1p36.11 (12)	428 (556)	
O	**22q13.33 (9)**	**667 (669)**	
P	5p12 (4)	59, 300,318,330, 345, 352, 367, 369, 376, 378 (381)	
R	16p13.3 (4)	95 (116)	
S	**15q26.3 (6)**	**188 (189)**	
SPS2	-	60 (448)	
T	3q24 (6)	36 (182)	
TR1	12q23.3 (15)	498 (499)	
TR2	3q21.2 (16)	655 (656)	
TR3	22q11.21 (18)	522 (523)	
V	**19q13.13 (6)**	**273 (346)**	
W	19q13.32 (6)	13 (87)	

Figure 1. (A.) Mammalian selenoprotein genes. Mammalian SECIS element consensus and SECIS elements in newly unidentified human selenoprotein genes. Only the upper portions of SECIS elements are shown. (**B.**) Mammalian selenoprotein genes. Human selenoprotein genes. Proteins are shown in alphabetical order and the newly identified genes are highlighted. On the right, relative lengths of selenoproteins are shown and Sec locations within the proteins are indicated by red vertical lines. The regions in selenoproteins that correspond to downstream αhelices are highlighted.

regions extends beyond the UGA codon in selenoprotein gene. We previously developed two computer programs, SECISearch 1.0 and geneid, which were used to identify several new selenoprotein sequences (*16–18*), and related approaches have also been developed (*19*). However, these methods were insufficient in identifying selenoprotein genes in mammalian genomes because of their size and complexity.

Our SECIS-based method, as applied to mammalian genomes (Figure S1), consisted of the following principal steps (*20*): (i) We identified candidate SECIS elements in the human genome with SECISearch 2.0. This program analyzed structural and thermodynamic features of SECIS elements and was about 10 times more selective (with the same specificity) than the original version of SECISearch (*16*). (ii) We identified human/mouse and human/rat SECIS pairs with SECISblastn, a program that analyzed evolutionary conservation of mammalian SECIS elements. This program was based on our observation that human, mouse, and rat SECIS elements in orthologous selenoprotein genes exhibited detectable sequence similarity. SECISblastn provided an increase of about 100-fold in the specificity of genomic searches. (iii) We analyzed genomic sequences upstream of candidate SECIS elements with geneid (*18*), a gene prediction program that identified open reading frames (ORFs) that had high coding potential and that contained in-frame TGA codons. (iv) We analyzed predicted human selenoprotein genes with mammalian selenoprotein gene signature (MSGS) criteria (*21*), which screened selenoprotein homologs for the presence and conservation of ORFs, in-frame TGA codons, and SECIS elements.

Primary sequences of more than 95% previously characterized mammalian SECIS elements contain an adenosine that precedes the quartet of non–Watson-Crick base pairs, a TGA_GA motif in the quartet, and two adenosines in the apical loop or bulge (*12*) (the ATGA_AA_GA pattern) (Figure 1A). In addition, in mammalian SelM SECIS elements, AA is replaced with CC (*22*) (the ATGA_CC_GA

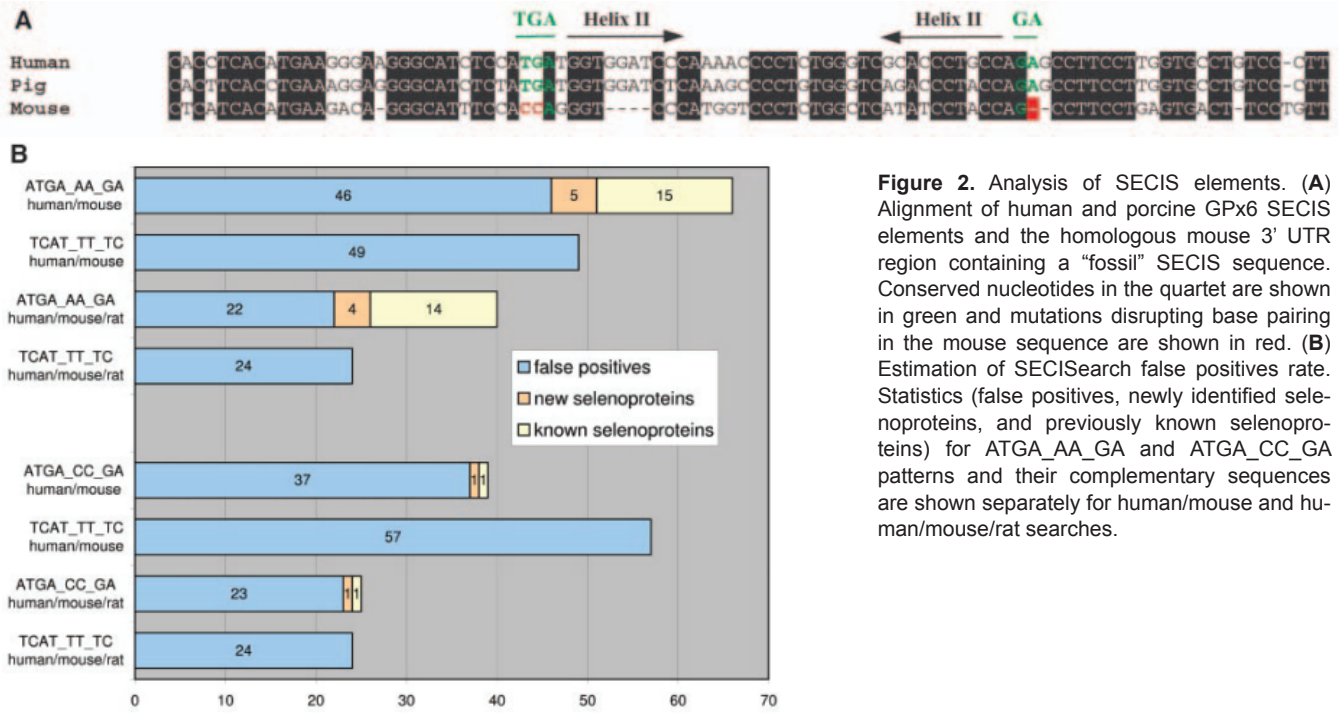


Figure 2. Analysis of SECIS elements. (**A**) Alignment of human and porcine GPx6 SECIS elements and the homologous mouse 3' UTR region containing a "fossil" SECIS sequence. Conserved nucleotides in the quartet are shown in green and mutations disrupting base pairing in the mouse sequence are shown in red. (**B**) Estimation of SECISearch false positives rate. Statistics (false positives, newly identified selenoproteins, and previously known selenoproteins) for ATGA_AA_GA and ATGA_CC_GA patterns and their complementary sequences are shown separately for human/mouse and human/mouse/rat searches.

pattern). The SECISearch 2.0 screen of mammalian genomes using the ATGA_AA_GA pattern resulted in 7146 human structures. The SECISblastn analysis reduced the number of structures to 1031 human/mouse and 276 human/rat pairs, and subsequent use of contamination, shotgun redundancy, and repetitive element filters resulted in 56 unique human/mouse and 58 unique human/rat pairs, including 40 structures that were common to all three organisms. The geneid analyses of sequences upstream of candidate SECIS elements and a subsequent analysis with MSGS criteria reduced the set to 20 hits. Among these, 15 were already known human selenoproteins and 5 were novel selenoproteins, designated as SelH, SelI, SelK, SelS, and SelV (Figure 1B, figs. S2 to S6, and figs. S10 and S11).

A similar computational screen using the ATGA_CC_GA pattern (*23*) detected a single true positive selenoprotein (SelM) and one novel selenoprotein (SelO) (Figure 1A, and 1B; Figure S7; and figs. S10 and S11). Only two known human selenoprotein genes were not identified by these procedures: The *SPS2* gene was absent in the human genome assembly, whereas the thioredoxin reductase 2 (TR2) gene contained a SECIS element with a thymidine preceding the quartet, a structure that does not correspond to other known SECIS elements.

The 24 mammalian selenoproteins were subsequently examined for the presence of homologs. This analysis identified a 25th human selenoprotein, designated glutathione peroxidase 6 (GPx6) (figs. S8, S10, and S11), a close homolog of plasma GPx3. GPx6 was not identified in the SECISearch-based computational screen, because its mouse and rat orthologs had Cys in place of Sec and the corresponding genes lacked SECIS elements. Rat GPx6 was previously cloned as rat odorant-metabolizing protein (*24*). Homology analyses revealed a "fossil," nonfunctional SECIS element in the 3' UTR of the mouse GPx6

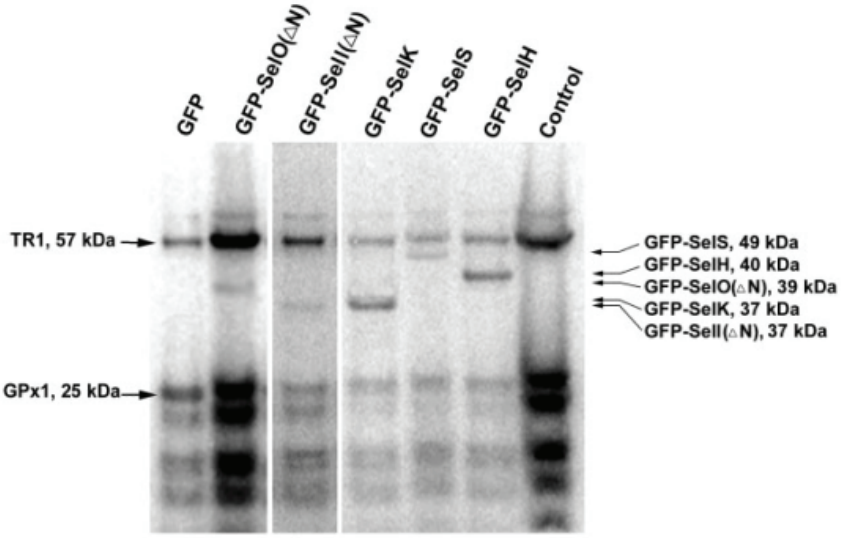


Figure 3. Incorporation of selenium into newly identified mammalian selenoproteins. GFP-selenoprotein constructs were used for convenient visualization of signals, wherein the fusion proteins differed in size from endogenous selenoproteins. Also for convenient visualization, the N-terminal regions of SelO and SelI were deleted. After transfection into CV-1 cells, transfected and control cells were incubated with ^{75}Se[selenite] for 24 hours, the extracts were resolved by SDS–polyacrylamide gel electrophoresis, and the labeled selenoproteins were visualized with a PhosphorImager. Locations of transfected selenoproteins are indicated on the right, and locations of major endogenous selenoproteins (TR1 and GPx1) are on the left. The left lane (GFP) shows control transfection with GFP alone. The right lane (control) shows untransfected CV-1 cells. The five middle lanes show experiments with indicated selenoproteins. All five showed ^{75}Se-labeled bands of the size expected if TGA encoded Sec.

gene, which contained mutations that disrupted the quartet and secondary structure (Figure 2A). We also cloned the gene encoding porcine GPx6 and found that it had a SECIS element and encoded a selenoprotein. These data revealed that Sec, which was initially present in the mammalian GPx family, was replaced by Cys in rodent genes for GPx6.

To estimate the number of false positives in the set of hits selected by SECISearch and SECISblastn, searches were performed using patterns that were complementary to the conserved SECIS sequences. The false positive rate with such patterns should be similar to that in the SECIS patterns, but the true positive rate with the complementary patterns should be zero. The difference between the number of SECIS candidates conforming to the major SECIS pattern, ATGA_AA_GA, and that of the complementary pattern corresponded approximately to the number of identified selenoprotein genes (Figure 2B). Thus, the ability of our SECIS-based method to recognize known mammalian selenoproteins and to complete analyses of all other candidates indicates that all or almost all selenoproteins common to human and rodent genomes were identified by our procedures. In addition, neither the SECISearch analyses of human and mouse dbEST and pair-wise searches of human/mouse genomes with altered SECIS patterns (*23*), nor the SECIS-independent searches for Sec/Cys pairs in homologous sequences (see below), revealed additional mammalian selenoproteins. The seven new human selenoproteins were either incorrectly predicted or not detected at all in Celera (*8*), National Center for Biotechnology Information (*7*), and Golden Path (*25*) human genome assemblies and annotations. In new as well as in known selenoproteins, Sec was located either upstream of an αhelix or very close to the C terminus (Figure 1B).

When the SECISearch-based method was applied to other eukaryotic genomes, we found neither selenoprotein genes nor Sec insertion machinery genes in yeast *Saccharomyces cerevisiae* or *Schizosaccharomyces pombe*, or in plant *Arabidopsis thaliana* genomes, whereas we could find only one and three already known selenoproteins in *Caenorhabditis elegans* and *Drosophila melanogaster* genomes, respectively (*26*) (Figure S12).

GPx6 and SelV were homologs of the previously characterized selenoproteins

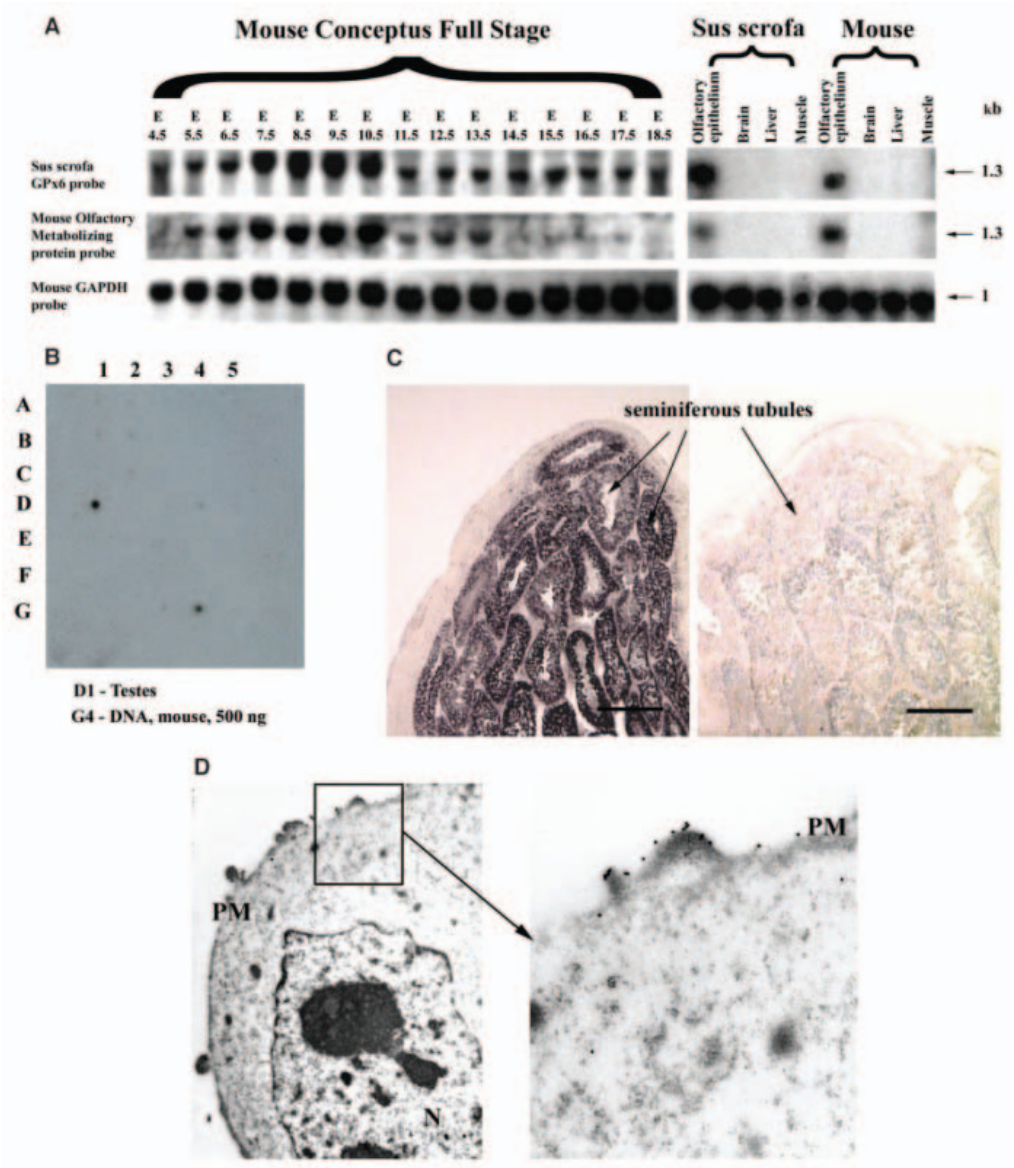


Figure 4. Expression of mammalian selenoproteins. (**A**) GPx6 mRNA is expressed in embryos and olfactory epithelium. On the left, a mouse full-stage conceptus Northern blot (See-Gene, Del Mar, CA) was probed with pig GPx6, mouse GPx6, and glyceraldehyde-3-phosphate dehydrogenase cDNA probes. On the right, mRNA isolated from indicated mouse and pig tissues was probed as above. We observed no significant cross-hybridization with other GPx mRNAs, which also migrated differently than the 1.3-kb GPx6 mRNA on these northern blots. (**B**) SelV mRNA is expressed in testes. A mouse multiple-tissue blot was developed with a mouse SelV mRNA probe. Northern blots also revealed testes-specific expression (*23*). (**C**) In situ hybridization of SelV mRNA in seminiferous tubules. On the left, a SelV sense probe was used. On the right, a SelV antisense probe (control) was used. (**D**) SelS and SelK are plasma membrane proteins. A construct encoding SelS-GFP fusion protein was generated and transfected into NIH 3T3 cells, and the expressed protein was detected with antibodies to GFP by means of electron microscopy.

GPx1 and SelW, respectively, and shared a conserved Sec with these proteins. To validate the remaining five new selenoproteins, we demonstrated the incorporation of selenium into these proteins by metabolic ^{75}Se labeling of CV-1 cells that were transfected with selenoprotein constructs (Figure 3). Analysis of the expression patterns of these selenoprotein genes revealed that SelH, SelI, SelO, SelS, and SelK mRNAs were present in a variety of tissues and cell types (*23*). However, the GPx6 mRNA was only detected in embryos and olfactory epithelium (Figure 4A), and expression of SelV mRNA was restricted to testes (Figure 4B), where it occurred in seminiferous tubules (Figure 4C). The secondary structure and protein organization predictions suggested that, like all previously characterized mammalian selenoproteins, GPx6, SelH, SelO, and SelV were globular proteins. However, SelK and SelS were predicted membrane proteins. We expressed fusions of SelK (*23*) and SelS (Figure 4D) containing a C-terminal green fluorescent protein (GFP) tag in CV-1 cells and found that the fusion products did reside on the plasma membrane. Thus, SelK and SelS are the first known plasma membrane selenoproteins.

We next applied the Sec/Cys homology method to the human genome in two different ways. First, we predicted with geneid, and regardless of SECIS elements, all possible human genes that were interrupted by in-frame TGA codons. The predicted ORFs were extended from TGA to the next terminator signal and were analyzed by BLASTP and TBLASTN against all proteins predicted in completely sequenced eukaryotic genomes. This procedure was designed to identify sequences with homology in TGA-flanking regions, which either conserve TGA or replace TGA with TGC or TGT (Cyst codons). Second, we analyzed by TBLASTN all human proteins against all human expressed sequence tags to identify paralogs that contain TGA in place of a Cys codon. These two Sec/Cys homology approaches recognized the majority of selenoprotein genes that were found through SECIS elements but did not identify additional selenoproteins (*23*), providing additional evidence that all or virtually all mammalian selenoproteins have been identified in our work.

Dietary selenium plays an important role in cancer prevention (*27*), immune function (*28*), aging (*17*), male reproduction (*28*), and other physiological and pathophysiological processes (*29*). Selenoproteins are thought to be responsible for most biomedical effects of dietary selenium and are essential to mammals. Information on a set of human and mouse selenoproteins should provide the basis for future systematic analysis of mammalian selenoprotein functions.

References

1. A. Bock, *Biofactors* **11**, 77 (2000).
2. S. C. Low, M. J. Berry, *Trends. Biochem. Sci.* **21**, 203 (1996).
3. D. L. Hatfield, V. N. Gladyshev, *Mol. Cell. Biol.* **22**, 3565 (2002).
4. L. Cataldo *et al., Mol. Reprod. Dev.* **45**, 320 (1996).
5. V. N. Gladyshev, K.-T. Jeang, T. C. Stadtman, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 6146 (1996).
6. M. J. Guimaraes *et al., Proc. Natl. Acad. Sci. U.S.A.* **93**, 15086 (1996).
7. E. S. Lander *et al., Nature* **409**, 860 (2001).
8. J. C. Venter *et al., Science* **291**, 1304 (2001).
9. D. Behne *et al., Biol. Trace Elem. Res.* **55**, 99 (1996).
10. R. F. Burk, K. E. Hill, *Bioessays* **21**, 231 (1999).
11. M. J. Berry *et al*., *Nature* **353**, 273 (1991).
12. R. Walczak, E. Westhof, P. Carbon, A. Krol, *RNA* **2**, 367 (1996).
13. R. M. Tujebajeva *et al., EMBO Rep.* **1**, 158 (2000).
14. D. Fagegaltier *et al., EMBO J.* **19**, 4796 (2000).
15. P. R. Copeland *et al., EMBO J.* **19**, 306 (2000).
16. G. V. Kryukov, V. M. Kryukov, V. N. Gladyshev, *J. Biol. Chem.* **274**, 33888 (1999).
17. M. J. Martin-Romeo *et al., J. Biol. Chem.* **276**, 29798 (2001).
18. S. Castellano *et al*., *EMBO Rep.* **2**, 697 (2001).
19. A. Lescure, D. Gautheret, P. Carbon, A. Krol, *J. Biol. Chem.* **274**, 38147 (1999).
20. Materials and methods are available as supporting material on *Science* Online.
21. G. V. Kryukov, V. N. Gladyshev, *Methods Enzymol.* **347**, 84 (2002).
22. K. V. Korotkov, S. V. Novoselov, D. L. Hatfield, V. N. Gladyshev, *Mol. Cell. Biol.* **22**, 1402 (2002).
23. G. V. Kryukov *et al*., data not shown.
24. T. N. Dear, K. Campbell, T. H. Rabbitts, *Biochemistry* **30**, 10376 (1991).
25. J. W. Kent, D. Haussler, *Genome Res.* **11**, 1541 (2001).
26. G. V. Kryukov, V. N. Gladyshev, unpublished data.
27. G. F. Combs Jr., L. C. Clark, B. W. Turnbull, *Biofactors* **14**, 153 (2001).
28. F. Ursini *et al*., *Science* **285**, 1393 (1999).
29. M. P. Rayman, *Lancet* **356**, 233 (2000).

30. We thank D. L. Hatfield for helpful discussions and Y. Zhou for assistance with microscopy. Supported by NIH GM61603 (to V.N.G.) and Ministerio de Ciencia y Tecnologia BIO2000-1358-C02-02 (to R.G.). S.C. is the recipient of a predoctoral fellowship from Generalitat de Catalunya.

Supporting Material is attached (it is also online @ http://www.sciencemag.org/cgi/content/full/300/5624/1439/DC1).

Supporting Material

Databases

The 08/06/01 "GoldenPath" draft assembly of the human genome that was masked for repetitive elements with RepeatMasker was used in the present study. Mouse and rat genome shotgun sequencing data, completely and incompletely sequenced genomes of eukaryotes, archaea and bacteria, and EST databases were obtained from NCBI, TIGR or sources indicated in the text.

SECIS-based identification of selenoprotein genes in eukaryotic genomes

SECISearch 2.0: genome-wide identification of SECIS elements

SECISearch 2.0 can identify candidate SECIS elements in nucleotide sequence databases on the basis of their primary sequences, secondary structures and predicted free energy criteria. This program has major improvements over its initial version (*1*) both at the level of individual modules and the overall composition of the program. An on-line version of SECISearch (Supporting Fig. S13) is available at http://genome.unl.edu/SECISearch.html and allows a user to choose among three patterns of different stringency and manually adjust free energy parameters. Several fine structural filters are also optional. Investigators interested in the source code are encouraged to contact the authors (E-mail: vgladyshev1@unl.edu).

Both SECISearch 2.0 and its on-line version contain three modules. The first module is based on the PatScan program (http://www-unix.mcs.anl.gov/compbio/PatScan/HTML/patscan.html) and searches for RNA structures that match the SECIS element primary sequence and the secondary structure consensus. The second module, based on the RNAfold program from Vienna RNA package (http://www.tbi.univie.ac.at/~ivo/RNA) (*2*), predicts secondary structure and calculates free energy for the entire SECIS element and separately for its core structure composed of quartet, Helix II and Apical loop. The program imposes three constraints: 1) pairing of the quartet nucleotides; 2) the presence of an unpaired nucleotide in the 5' proximal position to the quartet; and 3) the presence of two unpaired nucleotides that correspond to the AA motif in the Apical loop or bulge in the SECIS element consensus. Predicted RNA structures, whose calculated free energies are above thresholds determined from the analysis of known SECIS elements, are excluded from further analysis by the second module. The third module of SECISearch 2.0 imports SECIS candidates that are generated by the first two modules and filters out structures that possess features not found in any known eukaryotic SECIS elements. Specifically, these fine structural requirements remove SECIS candidates that 1) are Y-shaped, 2) contain >2 adjacent unpaired nucleotides among 7 nucleotides in Helix II that are proximal to the quartet, 3) contain <8 base pairs in the SECIS segment composed of Helix II and Apical loop; and 4) contain >2 unpaired nucleotides on the 5' side than on the 3' side. For convenient visualization and examination of the data, we developed an RNAnice program,

which can draw SECIS elements in proper orientation, with annotation and highlighted features.

Eukaryotic SECIS elements are usually classified as Type I and Type II structures (*3*). Type I SECIS elements have a fully unpaired Apical loop, whereas Type II SECIS elements possess an additional minihelix within the Apical loop. Both structures are interconvertible by mutations in the minihelix (*3*) and do not differ in their predicted free energy values (*1*). SECISearch 2.0 is able to identify both SECIS types using the same set of parameters. SECISearch 2.0 parameters were tuned using a set of 75 eukaryotic SECIS elements that were extracted from non-redundant and EST databases. This set included SECIS elements from all previously known human and mouse selenoprotein genes and also contained 37 SECIS elements from 11 other species.

SECISblastn: analysis of evolutionary conservation of predicted candidate SECIS elements

Since SECIS elements are essential for Sec insertion (and therefore for selenoprotein function), they are subject to natural selection pressure. We have found that not only secondary structures of SECIS elements are conserved, but that all known human SECIS elements exhibit nucleotide sequence similarity to SECIS elements in orthologous mammalian selenoprotein genes. In contrast, non-orthologous SECIS elements have no detectable sequence homology. This finding allowed us to greatly reduce the number of false positives by requiring that each human candidate SECIS element have a homologous SECIS element in rat, mouse or both rat and mouse genomes.

Blast (*4*) databases were generated from human sets of candidate SECIS elements generated by SECISearch and the mouse and rat sets of SECIS candidates were searched against these databases using SECISblastn. This blastn-based program has been optimized for comparison of short segments of 3'-UTR regions (cost to open a gap is 3, cost to extend a gap is 1, reward for nucleotide match is 2, and low complexity sequence filtering with DUST is off). Mouse or rat candidate SECIS elements were discarded if no hits in the human database were found with an expectation value below $1e^{-10}$ (this threshold was determined from homology analyses of known SECIS elements in human and mouse orthologous selenoprotein genes). SECISblastn allowed more than 100-fold reduction in the number of false positives.

Shotgun redundancy filter

Intrinsic redundancy of mouse and rat shotgun genome sequence data resulted in redundancy of the set of identified putative SECIS elements and was removed by the redundancy filter that was developed using String::Approx Perl module for approximate string matching. All candidate SECIS elements in the mouse and rat sets with identity of $\geq$95% (measured as Levenshtein edit distance) to each other were replaced by first representative hits.

Human contamination filter

Our preliminary searches indicated that the current rat and mouse shotgun sequence data are contaminated with human sequence entries. To remove human sequences, we utilized a "cleaning" procedure – each rodent shotgun sequence entry that contained a putative SECIS element was compared with non-masked human genome using blastn program. Entries with ≥96% homology in regions longer than 500 nucleotides were removed from further analysis, and those that produced hits with a length *l* and identity level *I* were removed from further analysis if *I* exceeded *l*(1.142-0.005769*l)*. The fact that no known selenoprotein genes were lost during this procedure suggested the legitimate choice of criteria. A set of human candidate SECIS elements that corresponded to the remaining mouse and rat hits was extracted for further analysis of upstream genome regions with the geneid program.

geneid: a gene structure prediction program

geneid is a program that predicts protein coding genes in anonymous eukaryotic sequences (*5*; program documentation is available at http://www1.imim.es/geneid). We have modified geneid for predicting selenoprotein genes. The new version of the program recognizes TGA as both a stop codon and as a sense codon for Sec. Thus, coding exons with in-frame TGA can be reliably predicted as long as they maintain high coding potential in sequences downstream of the TGA. In a single prediction on a given genome, the modified version of geneid is able to predict both standard genes and selenoprotein genes. For each candidate human SECIS element, flanking 1 Mb sequence regions on each side were extracted. Selenoprotein gene prediction was performed, admitting genes interrupted by in-frame TGA codons with an additional requirement that SECIS structures be located less than 6,000 nucleotides downstream of the predicted stop codons.

Mammalian Selenoprotein Gene Signature: analysis of evolutionary conservation of predicted selenoprotein genes

Mammalian Selenoprotein Gene Signature (MSGS) is a set of criteria that describe features common to mammalian (and possibly eukaryotic) selenoprotein genes (*6*):

1) TGA-encoded Sec should be conserved and Sec-flanking protein sequences should be homologous for mammalian orthologous selenoprotein genes.
2) The SECIS element should be conserved and located in the 3'-UTRs of mammalian orthologous selenoprotein genes.
3) Distinct Cys- and/or Sec-containing homologs should exist, i.e., the occurrence of genes containing a Cys codon in place of TGA (or occurrence of distinct homologous genes that conserve TGA).

Predicted amino acid sequences of geneid-predicted selenoproteins were analyzed for the presence of paralogs in the human genome and homologs in other species in non-redundant and EST databases with blast programs. Six predicted selenoproteins that had both selenoprotein homologs (which contained SECIS elements) and cysteine-containing homologs (which had no SECIS elements) were considered to be true positives (GPx6, SelH, SelK, SelO, SelS and SelV). The remaining new selenoprotein, SelI, had no cysteine

homologs, but its orthologs in frogs, fish and other mammals had SECIS elements (Supporting Fig. S9). Thus, this protein was also classified as a true positive.

Identification of human selenoprotein genes by searching for Sec/Sec and Sec/Cys pars in homologous sequences (SECIS-independent methods)

Comparative selenoprotein gene prediction

SECIS-independent selenoprotein gene searches were performed on the 08/06/01 "GoldenPath" human genome assembly. The procedure employed was based on identification of in-frame TGA codons regardless of the presence of downstream SECIS elements, therefore addressing the issue of non-canonical SECIS elements in the human genome. This procedure also addressed the issue of potential occurrence of selenoproteins specific to the human genome. In the SECIS-independent searches for new selenoproteins, sensitivity was preferred to specificity, thus the chance of missing yet unknown selenoproteins was minimized. The *ab initio* gene prediction yielded 50,126 potential human genes, of which 27,605 had a TGA in-frame. This latter set included 21 out of 24 true selenoprotein genes that were identified by the SECISearch/SECISblastn/geneid/MSGS procedure. The set of 27,605 genes was further analyzed as follows:

1) The human 27,605 sequences were analyzed by blastp against a corresponding set of *Takifugu rubripes* proteins interrupted by TGA codons. The genome of this puffer fish (10/25/01 JGI draft assembly) encodes selenoprotein homologs of all 25 human selenoproteins, although the number of proteins in each selenoprotein family is different between human and puffer fish genomes. The *ab initio* geneid analysis of the puffer fish genome yielded 33,126 genes, of which 28,603 had a TGA in-frame, including 16 true selenoproteins corresponding to all but three human families. Human and fish proteins were then analyzed to identify potential human-fish selenoprotein orthologs containing in-frame TGA codons. This analysis identified 351 candidate orthologs.

2) The 27,605 human sequences were analyzed by blastp against a set of predicted *Takifugu rubripes* standard proteins. The *ab initio* geneid analysis of the puffer fish genome yielded 41,127 standard genes. Human and fish proteins were then analyzed to identify potential human-fish selenoprotein orthologs containing cysteine in fish. This analysis identified 296 candidate orthologs.

3) The sequences of these two sets of human candidate selenoproteins (351 + 296) were analyzed by blastp and tblastn against several completely sequenced eukaryotic genomes as well as against proteins predicted in these genomes (*Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisae* and *Arabidopsis thaliana*). The incompletely sequenced genomes were also analyzed (*Mus musculus*, *Xenopus laevis* and *Danio rerio*) to identify sequences with homology in TGA-flanking regions, containing either TGA (Sec codon) or TGT or TGC (Cys codons) in place of TGA. This analysis resulted in 32 human selenoprotein candidates with selenoprotein counterparts in fish and 58 human selenoprotein candidates with cysteine counterparts in fish.

4) After filtering proteins that had been previously characterized, the set contained only known selenoproteins and 12 other candidates. However, comparisons of these twelve sequences with corresponding EST sequences discarded potential in-frame TGAs due to either 1) predicted gene structure incompatible with the exonic structure of identical ESTs; or 2) TGA codon not supported by ESTs sequences (therefore, these were probable sequencing errors which produced false TGA codons in place of correct cysteine codons). Thus, SECIS-independent searches did not add new human selenoproteins to the set of selenoprotein predicted by the SECIS-dependent prediction.

Selenoprotein homology search: cysteine homolog approach

80% (20 out of 25) human selenoproteins have known homologs that contain cysteine in place of selenocysteine. Therefore, cysteine-containing homologs of most mammalian selenoproteins are likely already annotated in public databases and can be used to unveil their selenoprotein counterparts, providing a third independent approach to selenoprotein identification. 29,076 standard human genes (Ensembl protein annotation on the 12/22/01 "GoldenPath" draft assembly) were analyzed by tblastn against all human ESTs (EMBL, Rel. 69). This set contained seven cysteine paralogs of known selenoprotein families: GPx (ENSP00000229441, ENSP00000262661, ENSP00000296734, ENSP00000244392), SelR (ENSP00000286571, ENSP00000277598) and SelW (ENSP00000269578).

In order to pinpoint novel human selenoproteins the following procedure was carried out: 1) selection of Ensembl proteins with at least 5 human ESTs containing a TGA codon in place of a given cysteine position; and 2) selection of Ensembl proteins with an unknown or unclear function that might correspond to a selenoprotein. The final set contained only the seven paralogs of already known human selenoproteins.

A similar procedure was carried out for 4,380 potential novel human proteins obtained from sgp2 predictions (*7*). sgp2 is a program to predict genes by comparing anonymous genomic sequences from two different species. It combines tblastx (WU-Blast), a sequence similarity search program, with geneid, an *ab initio* gene prediction program. In this way, 4,380 new human proteins with a reliable mouse ortholog were obtained. Because of the novelty of these sequences, not many ESTs may be available. For this reason, proteins with as less as 2 human ESTs containing a TGA codon in place of a given cysteine position were selected for analysis. Four human candidates were further studied, though given the high error rate in EST sequencing, these proteins had low supporting evidence. No other homology support was found in screened genomes, and these ESTs were considered to have sequencing errors. Therefore, no novel human selenoproteins were discovered by this approach.

The overall data from the independent approaches (SECIS prediction, in-frame TGA prediction and Sec/Cys homology approaches) argue that we have identified all or almost all selenoprotein genes in the human genome. Thus, the remaining in-frame TGA codons may be interpreted as terminator signals.

Newly identified mammalian selenoproteins

Among the new proteins, SelV was composed of a ~25 kDa N-terminal proline/threonine-rich domain of unknown function and a ~10 kDa C-terminal Sec-containing domain homologous to SelW (Supporting Fig. S4). SelH was an ~13 kDa protein containing Sec within a putative redox motif CxxU (Sec separated from Cys by two other residues) and was a homolog of an N-terminal region of *Drosophila* BthD (Supporting Fig. S2). The ~9 kDa SelK had a predicted N-terminal trans-membrane region followed by an unstructured region that included a C-terminal penultimate Sec (Supporting Fig. S5). Similar protein organization and Sec location were observed for SelS, although at ~21 kDa, it was a larger protein than SelK (Supporting Fig. S3). SelI was a ~45 kDa protein homologous to yeast and human choline/ethanolaminephosphotransferases, except that it had a C-terminal Sec-containing extension (Supporting Fig. S6). Choline/ ethanolaminephosphotransferases are plasma membrane proteins containing 7 transmembrane regions. SelO was an ~73 kDa protein, the largest eukaryotic selenoprotein (Supporting Fig. S7). Sec was present in this protein as a C-terminal penultimate residue, and no homologs of known function were detected for SelO.

Expression of mammalian selenoprotein genes

To assess GPx6 mRNA expression, total RNA was isolated from indicated pig and mouse tissues with a RNAqueous Kit (Ambion), applied on a denaturing agarose gel and transferred onto a Zeta-Probe Blotting membrane (Bio-Rad). This membrane, as well as Mouse Conceptus Full Stage membrane (See-gene) were probed individually with a 1.3-kb ^{32}P-labeled fragment of pig GPx6, 0.7 kb ^{32}P-labeled fragment of mouse GPx6 and a ^{32}P-labeled mouse glyceraldehydes-3-phosphate dehydrogenase probe as a control. All probes were generated by a Rediprime II random prime labeling system (Amersham Pharmacia Biotech). To assay SelV mRNA expression, mouse RNA Master Blot (Clontech), which contained mRNA samples isolated from 22 mouse tissues, was probed with a full length SelV probe, also generated by a Rediprime II random prime labeling system.

To localize SelV gene expression in mouse testes, a 160 bp fragment of the mouse SelV gene was amplified with 5'-TATGAAGCTTAAGTCCCTAACCCTGTTCCAATC-3' and 5'-TCAAGAATTCGATCTTAGGAAAGACCCGACCTAG-3' primers and cloned into *Hind*III/*EcoR*I sites pGEM-3Z(+) vector (Promega). 8 μm thick slides of mouse testes were probed with sense or antisense SelV RNA probes, which were obtained by *in vitro* transcription using the DIG RNA Labeling Kit (Roche Molecular Biochemicals). The probe was visualized using BCIP/NBT substrate and AP conjugated anti-DIG IgG (Roche Molecular Biochemicals).

Electron microscopy

A 300 bp coding region of SelK protein was amplified with primers 5'-ATCCCTCGAGTCTCTGTCGCTAGGAAGCAGGCAAC-3' and 5'-AATCGGATCCTTCCGTCCACCAGCCATTGG-3' and cloned into *Xho*I/*BamH*I sites of pEGFP-N2 vector (Clontech). NIH 3T3 cells were transfected with a SelK-GFP construct using Lipofectamine Plus reagent (Invitrogene), fixed, embedded into LR-white resin and

sectioned. Ultrathin slides were treated with anti-GFP serum (Invitrogene) and anti-rabbit gold-labeled secondary antibodies (Jackson ImmunoResearch).

Metabolic labeling of proteins with ^{75}Se

All constructs that were used for metabolic labeling of new selenoproteins with ^{75}Se were developed using expression vector pEGFP-C3 (Clontech). Entire coding regions and 3'-UTRs of selenoproteins K, H and S, and 3'-UTRs and regions coding for C-terminal portions of Selenoproteins I and O, were amplified, respectively, with
K-5prime 5'-ACTCCCGAATTCTGGTTTACATCTCGAATGGTCAGG-3' and
K-3prime 5'-CGAACCGGATCCATGAGAGCAAAGTAACAGTGAGCAG-3',
H–5prime 5'-CTAAGAATTCATGGCCCCCCACGGAAGAAAG-3' and
H-3prime 5'-CTATGGATCCTTATGAAAGGTACTTCTTCAATTCTTC-3',
S-5prime 5'-AACTGACTCGAGATGGATCGCGATGAGGAACCTC-3' and
S-3prime 5'-TCCAGAGGATCCGTTTACTGTCTGACAAAGTCAAGCTCAC-3',
I-5prime 5'-AAGACTCGAGAGCAGCACGCGGTGCCCGAC-3' and
I-3prime 5'-CAGGGAATTCGGGCTTATCTTCGACAGCCTGGAC-3',
O-5prime 5'-GACCGTCTCGAGCTACAGGAATACAGAGACCGTCTC-3' and
O-3prime 5'-CATTGAATTCGCACACACAGGCCACAAGGCTTAC-3' primers, and cloned into *Eco*RI/*BamH*I (SelK and SelH), *Xho*I/*BamH*I (SelS) and *Xho*I/*Eco*RI (SelI and SelO) sites of pEGFP-C3. Transfections of CV-1 cells were carried out using Lipofectamine and 4-5 μg of DNA. The samples were analyzed on SDS-10% NuPAGE gels (Invitrogene). ^{75}Se-labeled proteins were visualized with a Storm PhosphorImager system (Molecular Dynamics).

Figure S1. Computational search for mammalian selenoprotein genes. A search using the ATGA_AA_GA pattern is shown. See Supporting Material text for details.

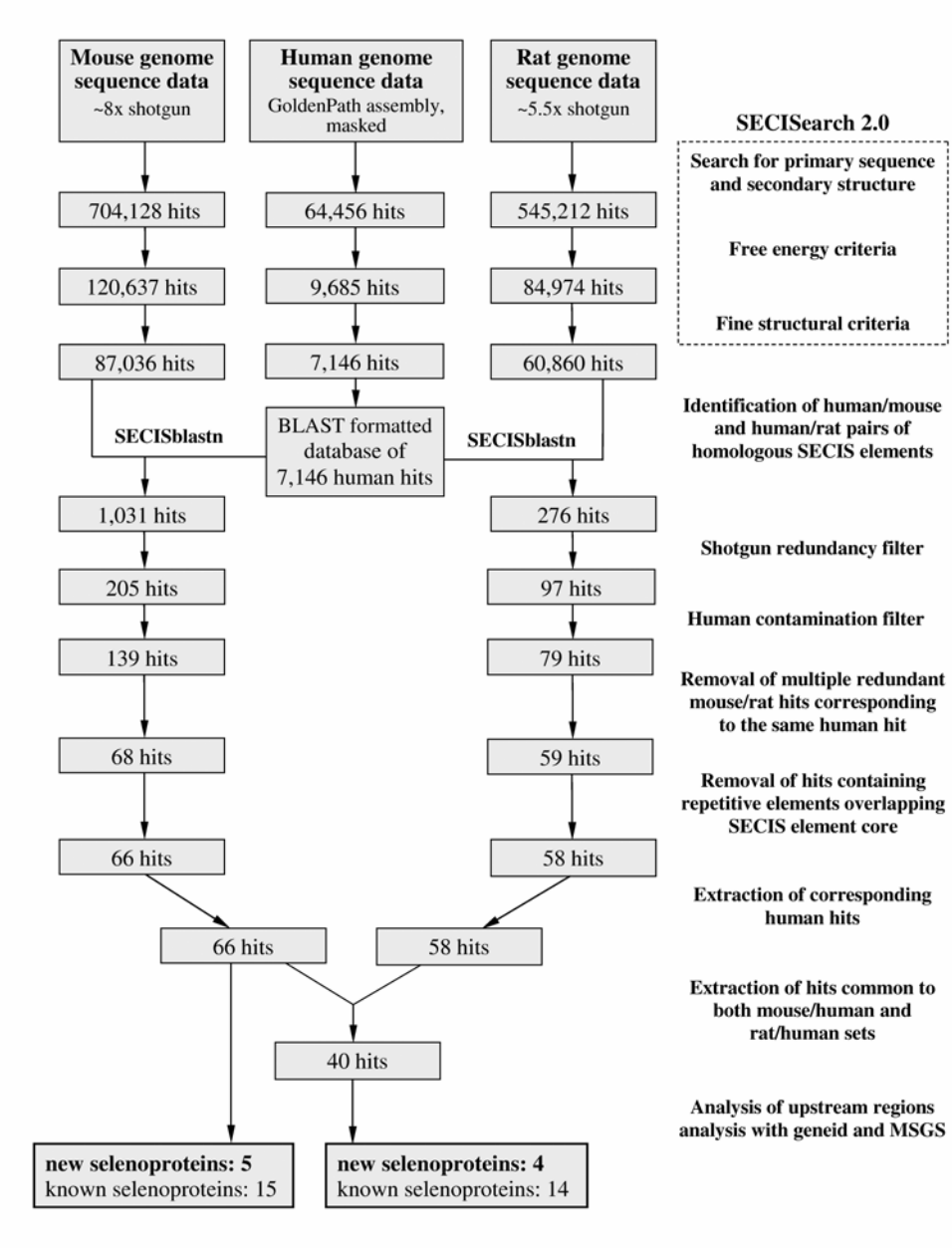


Figure S2. Selenoprotein H (SelH) alignment. Accession numbers for sequences are as follows: *Homo sapiens* (this study), *Mus musculus* (this study), *Oryzias latipes* (BJ007554) and *Drosophila melanogaster* (NP_572903). Amino acid sequence alignments in Fig. S2-S8 were generated with PileUp program from GCG package and shaded by BoxShade program v3.21. Selenocysteine is shown in red as U.

```
Homo sapiens               1 ~~~~MAPRGRKRKAEAAVVA.VAEKREKLANGGEGMEEA..T.......VVIEHCTSURVYGRNAAALSQALRLEA.PEL
Mus musculus               1 ~~~~MAPHGRKRKAGAAPME.TVDKREKLAEG......A..T.......VVIEHCTSURVYGRHAAALSQALQLEA.PEL
Oryzias latipes            1 MASKAGRRGTKRKVEAKKEEDKTSTEEKKARGENAHEEAGLK.......VLIEHCKSURVYGRNAEEVKSALLAAR.PEL
Drosophila melanogaster    1 ~~~~~~~~~~~~~~~~~~~~~MPPKRNKKAEAPIAERDAGEELDPNAPVLYVEHCRSURVFRRRAEELHSALRERGLQQL

Homo sapiens              66 PVKVNP.TKPRRGSFEVTLLRPDGSS...AELWTGIKKGPPRKLKFPEPQEVVEELKKYLS~~~~~~~~~~~~~~~~~~~
Mus musculus              60 PVQVNP.SKPRRGSFEVTLLRSDNSR...VELWTGIKKGPPRKLKFPEPQEVVEELKKYLS~~~~~~~~~~~~~~~~~~~
Oryzias latipes           73 TVVCNP.EKPRRNSFEITLL..DGAK..ETSLWTGIKKGPPRKLKFPQPDDVVAAFKDALKTE~~~~~~~~~~~~~~~~~
Drosophila melanogaster   60 QLQLNALGAPRRGAFELSLSAGGMGKQEQVALWSGLKRGPPRARKFPTVEEVYDQIVGILGDQQESKEQTNTQKSSKIDL

Homo sapiens             123 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Mus musculus             117 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Oryzias latipes          131 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Drosophila melanogaster  140 PGSEAIASPKKSESTEEAQENKAPTSTSTSRKSKKEQKSEEEPTQVDSKEAKQSKELVKTKRQPKAQKKQAKASESQEEV

Homo sapiens             123 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Mus musculus             117 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Oryzias latipes          131 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Drosophila melanogaster  220 AEDKPPSSQKRKRTTRSSTDEATAGAKRRR
```

Figure S3. Selelenoprotein S (SelS) alignment. Accession numbers for sequences are as follows: *Homo sapiens* (this study), *Mus musculus* (this study), *Rattus norvegicus* (AAL59556) and *Ciona intestinalis* (AV972635).

Figure S4. Selenoprotein V (SelV) alignment. Accession numbers for sequences are as follows: *Homo sapiens* SelV (this study), *Mus musculus* SelV (this study), *Homo sapiens* SelW (O15532) and *Mus musculus* SelW (P49904).

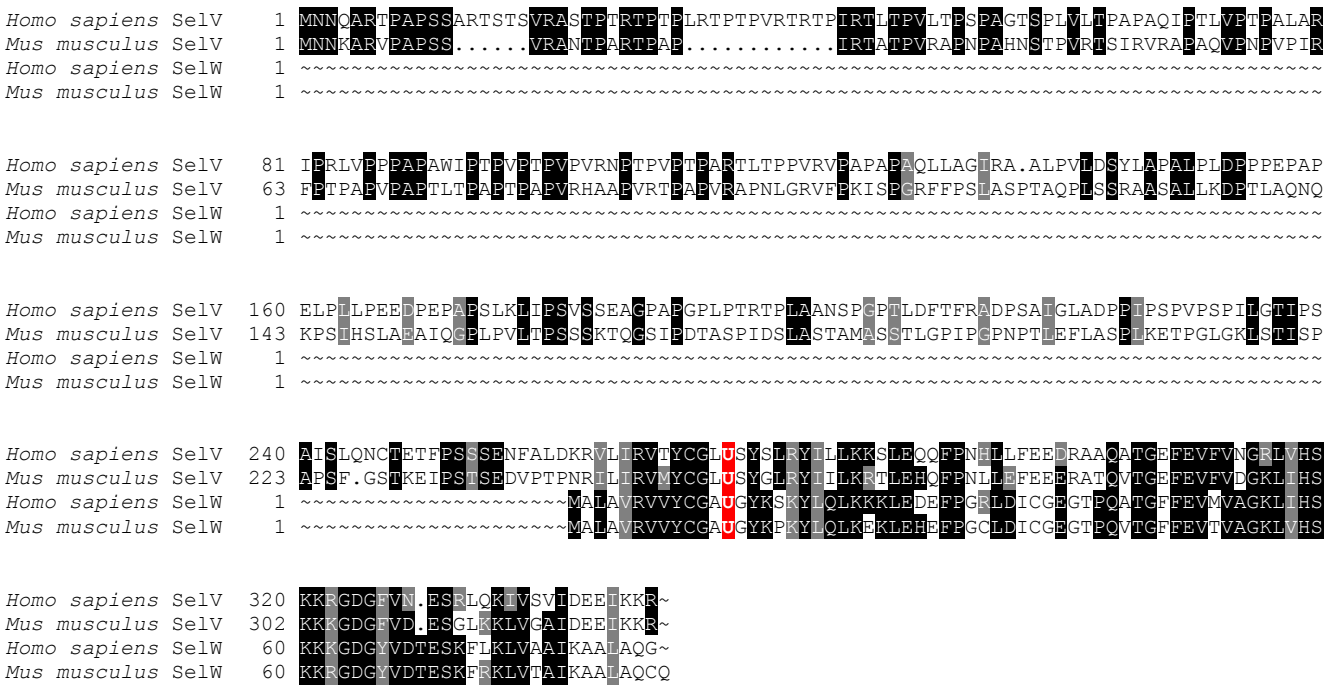


Figure S5. Selenoprotein K (SelK) alignment. Accession numbers for sequences are as follows: *Homo sapiens* (this study), *Mus musculus* (this study), *Fugu rubripes* (this study), *Oryzias latipes* (BJ003636, BJ004144 and BJ017876), *Drosophila melanogaster* (Sec-containing homolog) (AAK72981), *Drosophila melanogaster* (Cys-containing homolog) (AAF48112), *Arabidopsis thaliana* (AY072406), and *Physcomitrella patens* (BJ162647, BJ203491 and BJ193522).

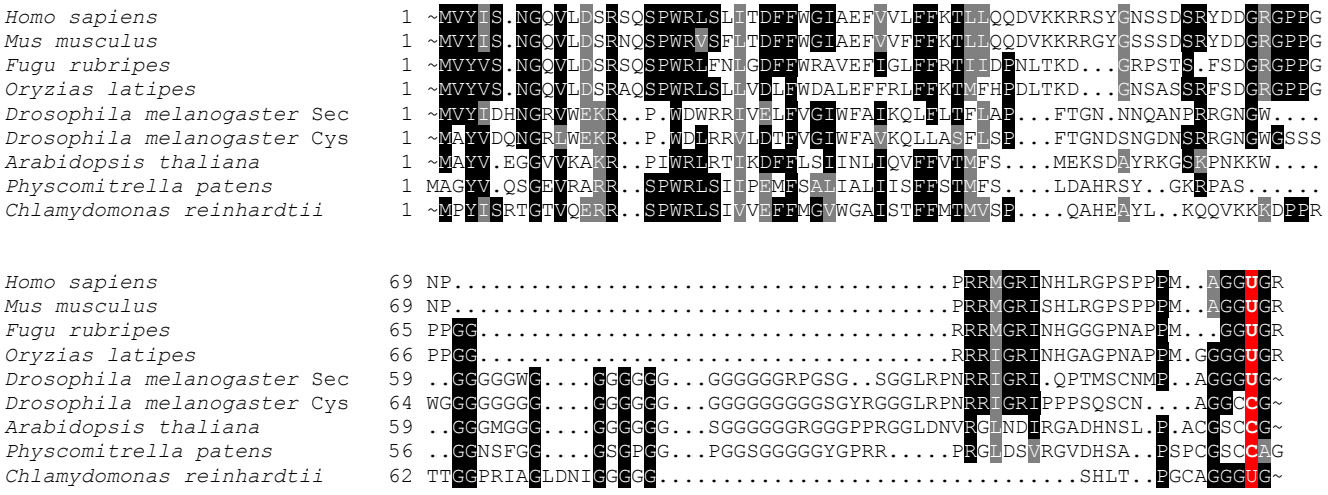
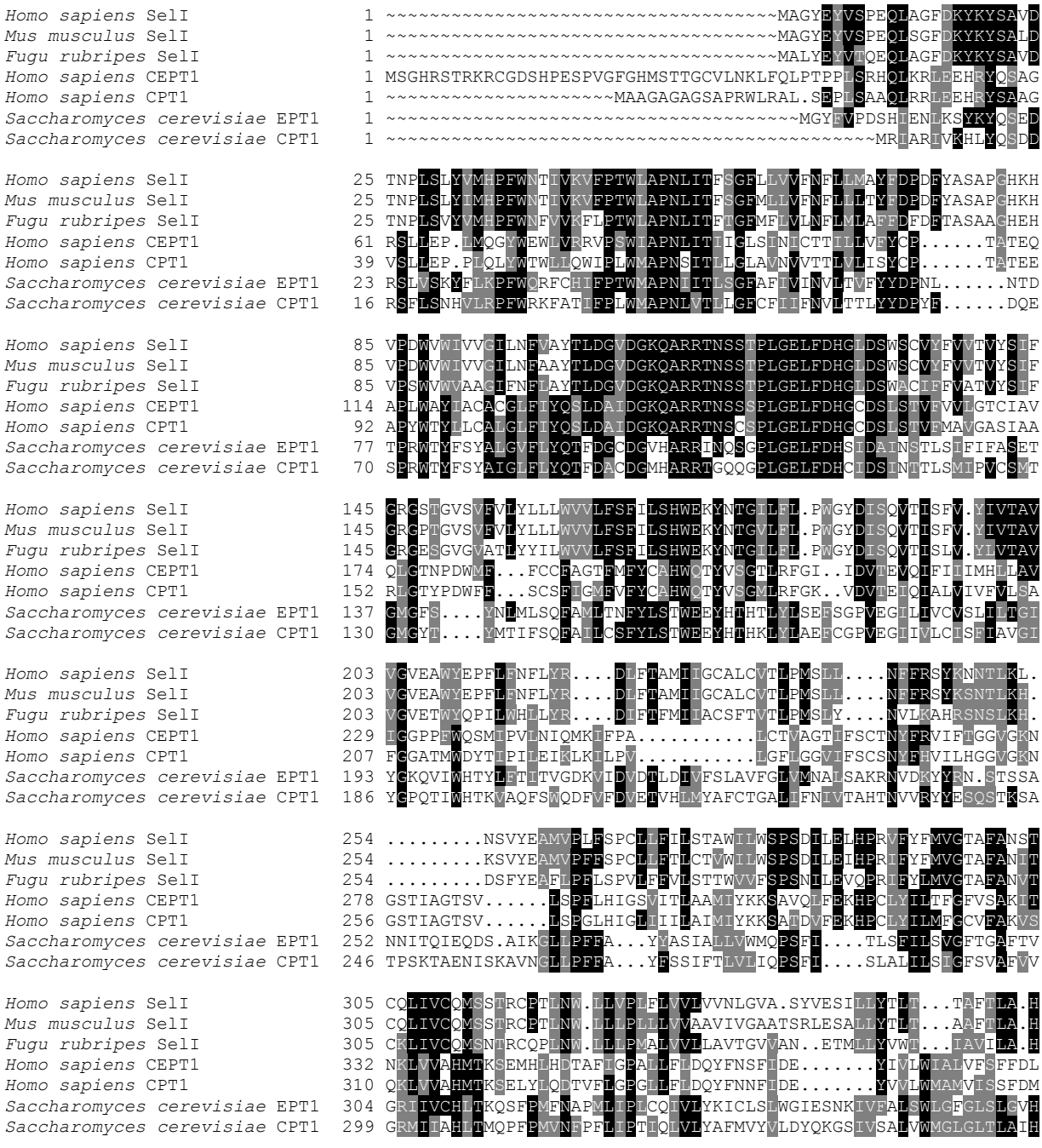


Figure S6. Selenoprotein I (SelI) alignment. Accession numbers for sequences are as follows: *Homo sapiens* (this study), *Mus musculus* (this study), *Fugu rubripes* (this study), *Homo sapiens* ethanolamine- and cholinephosphotranferase CEPT1 (AAD25170), *Homo sapiens* cholinephosphotransferase CPT1 (NP_064629), *Saccharomyces cerevisiae* ethanolamine- and cholinephosphotranferase EPT1 (NP_011991) and *Saccharomyces cerevisiae* cholinephosphotransferase CPT1 (NP_014269).



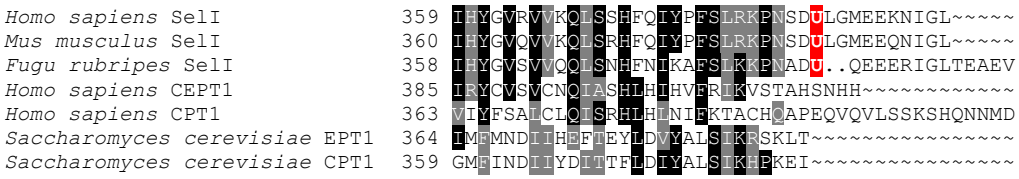
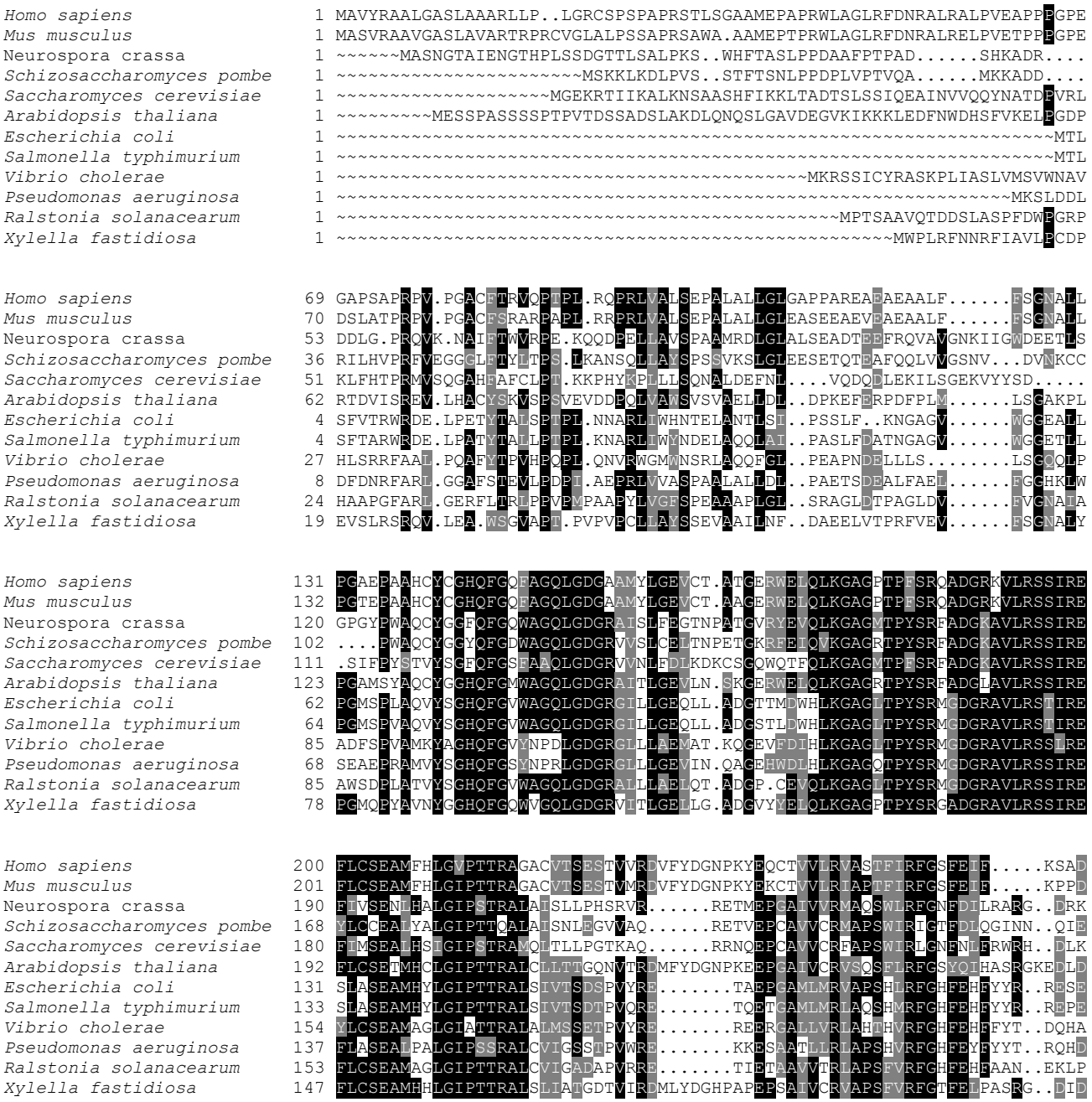
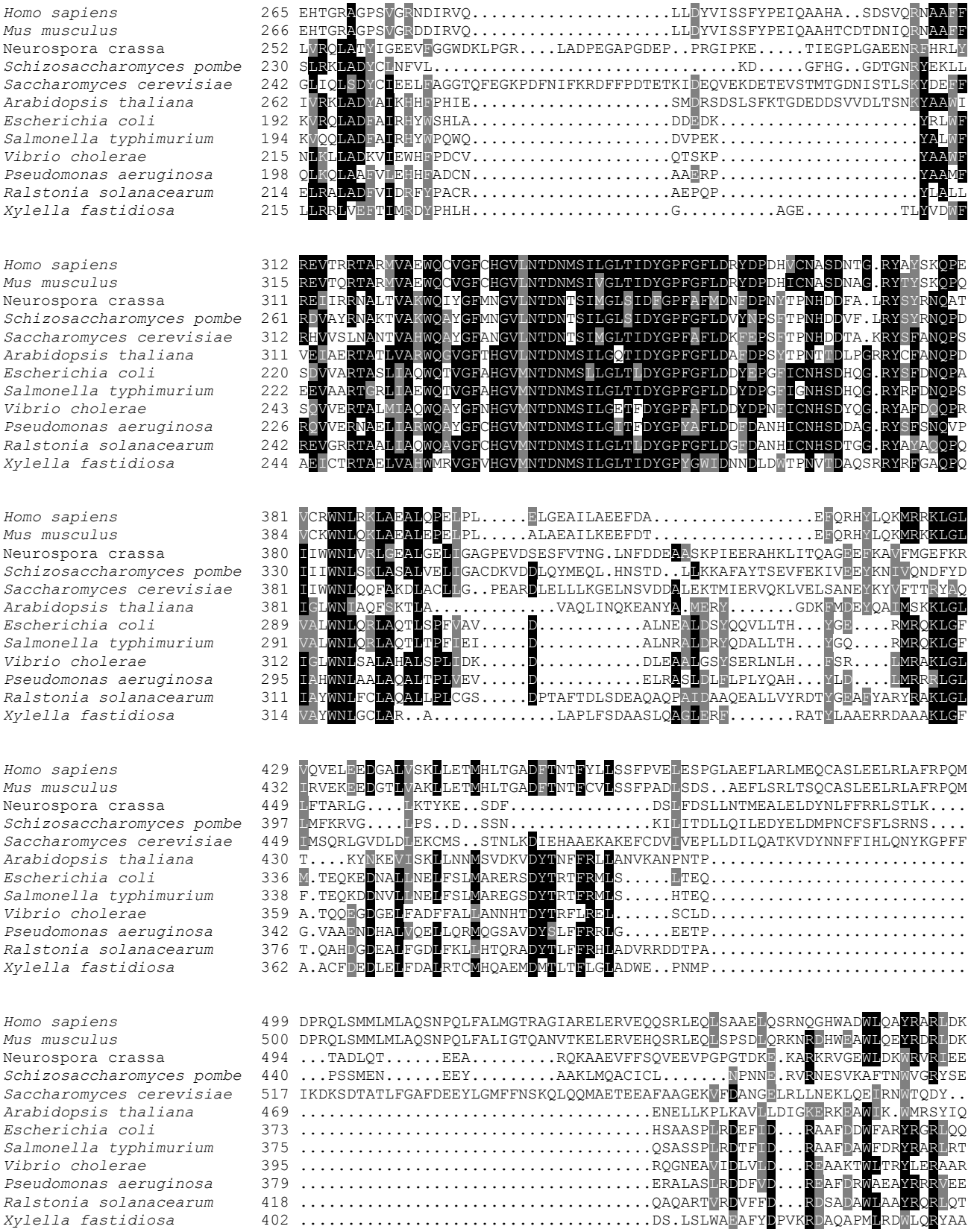
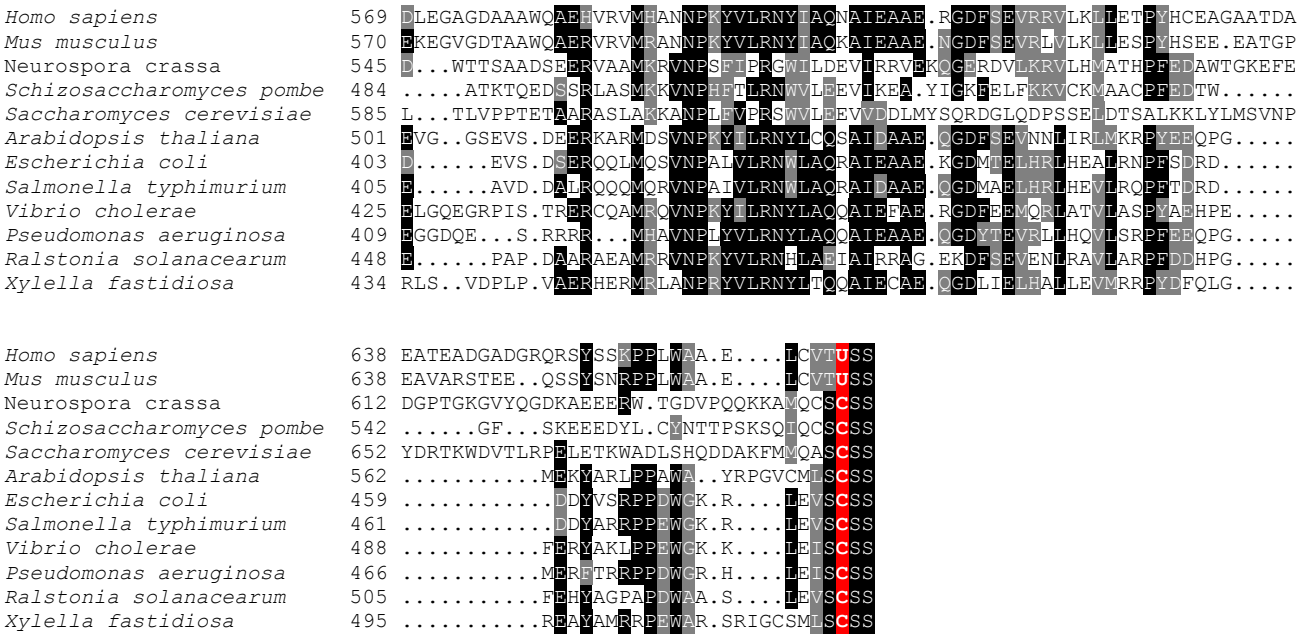

Homo sapiens SelI 359 IHYGVRVVKQLSSHFQIYPFSLRKPNSDULGMEEKNIGL~~~~~
Mus musculus SelI 360 IHYGVQVVKQLSRHFQIYPFSLRKPNSDULGMEEQNIGL~~~~~
Fugu rubripes SelI 358 IHYGVSVVQQLSNHFNIKAFSLKKPNADU..QEEERIGLTEAEV
Homo sapiens CEPT1 385 IRYCVSVCNQIASHLHIHVFRIKVSTAHSNHH~~~~~~~~~~~
Homo sapiens CPT1 363 VIYFSALCLQISRHLHLNIFKTACHQAPEQVQVLSSKSHQNNMD
Saccharomyces cerevisiae EPT1 364 IMFMNDIIHEFTEYLDVYALSIKRSKLT~~~~~~~~~~~~~~~~
Saccharomyces cerevisiae CPT1 359 GMFINDIIYDITTFLDIYALSIKHPKEI~~~~~~~~~~~~~~~~

Figure S7. Selenoprotein O (SelO) alignment. Accession numbers for sequences are as follows: *Homo sapiens* (this study), *Mus musculus* (this study), *Neurospora crassa* (CAB91237), *Schizosaccharomyces pombe* (O13890); *Saccharomyces cerevisiae* (NP_015102), *Arabidopsis thaliana* (AAK25868), *Escherichia coli* (NP_416221), *Salmonella typhimurium* (NP_460311), *Vibrio cholerae* (NP_231565), *Pseudomonas aeruginosa* (NP_253710), *Ralstonia solanacearum* (NP_519869) and *Xylella fastidiosa* NP_299896.




Homo sapiens 265 EHTGRAGPSVGRNDIRVQ.....LLDYVISSFYPEIQAAHA..SDSVQRNAAFF
Mus musculus 266 EHTGRAGPSVGRDDIRVQ.....LLDYVISSFYPEIQAAHTCDTDNIQRNAAFF
Neurospora crassa 252 LVRQLATYIGEEVEGGWDKLPGR....LADPEGAPGDEP..PRGIPKE.....TIEGPLGAEENREHRLY
Schizosaccharomyces pombe 230 SLRKLADYCLNFVL.....KD.....GFHG..GDTGNRYEKLL
Saccharomyces cerevisiae 242 GLIQLSDYCIEELEAGGTQFEGKPDFNIFKRDFFPDTETKIDEQVEKDETEVSTMTGDNISTLSKYDEFF
Arabidopsis thaliana 262 IVRKLADYAIKEHEPHIE.....SMDRSDSLSFKTGDEDDSVVDLTSNKYAAWI
Escherichia coli 192 KVRQLADFAIREYWSHLA.....DDEDK.....YRLWF
Salmonella typhimurium 194 KVQQLADFAIREYWPQWQ.....DVPEK.....YALWF
Vibrio cholerae 215 NLKLLADKVIEWHEPDCV.....QTSKP.....YAAWF
Pseudomonas aeruginosa 198 QLKQLAAFVLEHHEADCN.....AAERP.....YAAMF
Ralstonia solanacearum 214 ELRALADFVIDRFYPACR.....AEPQP.....YLALL
Xylella fastidiosa 215 LLRRLVEFTIMRDYPHLH.....G.....AGE.....TLYVDWF
Homo sapiens 312 REVTRRTARMVAEWQCVGFCHGVLNTDNMSILGLTIDYGPFGFLDRYDPDHVCNASDNTG.RYAYSKQPE
Mus musculus 315 REVTQRTARMVAEWQCVGFCHGVLNTDNMSIVGLTIDYGPFGFLDRYDPDHICNASDNAG.RYTYSKQPQ
Neurospora crassa 311 REIIRRNALTVAKWQIYGFMNGVLNTDNTSIMGLSIDFGPFAFMDNFDPNYTPNHDDFA.LRYSYRNQAT
Schizosaccharomyces pombe 261 RDVAYRNAKTVAKWQAYGFMNGVLNTDNTSILGLSIDYGPFGFLDVYNPSFTPNHDDVF.LRYSYRNQPD
Saccharomyces cerevisiae 312 RHVVSLNANTVAHWQAYGFANGVLNTDNTSIMGLTIDYGPFAFLDKFEPSFTPNHDDTA.KRYSFANQPS
Arabidopsis thaliana 311 VEIAERTATLVARWQGVGFTHGVLNTDNMSILGQTIDYGPFGFLDAFDPSYTPNTIDLPGRRYCFANQPD
Escherichia coli 220 SDVVARTASLIAQWQTVGFAHGVMNTDNMSLLGLTLDYGPFGFLDDYEPGFICNHSDHQG.RYSFDNQPA
Salmonella typhimurium 222 EEVAARTGRLIAEWQTVGFAHGVMNTDNMSILGLTIDYGPFGFLDDYDPGFIGNHSDHQG.RYRFDNQPS
Vibrio cholerae 243 SQVVERTALMIAQWQAYGFNHGVMNTDNMSILGETFDYGPFAFLDDYDPNFICNHSDYQG.RYAFDQQPR
Pseudomonas aeruginosa 226 RQVVERNAELIARWQAYGFCHGVMNTDNMSILGITFDYGPYAFLDDFDANHICNHSDDAG.RYSFSNQVP
Ralstonia solanacearum 242 REVGRRTAALIAQWQAVGFCHGVMNTDNMSILGLTLDYGPFGFLDGFDANHICNHSDTGG.RYAYAQQPQ
Xylella fastidiosa 244 AEICTRTAELVAHWMRVGFVHGVMNTDNMSILGLTIDYGPYGWIDNNDLDWTPNVTDAQSRRYRFGAQPQ
Homo sapiens 381 VCRWNLRKLAEALQPELPL.....ELGEAILAEEFDA.....EFQRHYLQKMRRKLGL
Mus musculus 384 VCKWNLQKLAEALEPELPL.....ALAEAILKEEFDT.....EFQRHYLQKMRKKLGL
Neurospora crassa 380 IIWWNLVRLGEALGELIGAGPEVDSESFVTNG.LNFDDEAASKPIEERAHKLITQAGEEFKAVFMGEFKR
Schizosaccharomyces pombe 330 IIIWNLSKLASALVELIGACDKVDDLQYMEQL.HNSTD..LLKKAFAYTSEVFEKIVEEYKNIVQNDFYD
Saccharomyces cerevisiae 381 IIWWNLQQFAKDLACLLG..PEARDLELLLKGELNSVDDALEKTMIERVQKLVELSANEYKYVFTTRYAQ
Arabidopsis thaliana 381 IGLWNIAQFSKTLA.....VAQLINQKEANYA.MERY.....GDKFMEYQAIMSKKLGL
Escherichia coli 289 VALWNLQRLAQTLSPFVAV.....D.....ALNEALDSYQQVLLTH...YGE....RMRQKLGF
Salmonella typhimurium 291 VALWNLQRLAQTLTPFIEI.....D.....ALNRALDRYQDALLTH...YGQ....RMRQKLGF
Vibrio cholerae 312 IGLWNLSALAHALSPLIDK.....D.....DLEAALGSYSERLNLH...FSR....LMRAKLGL
Pseudomonas aeruginosa 295 IAHWNLAALAQALTPLVEV.....D.....ELRASLDLFLPLYQAH...YLD....LMRRKLGL
Ralstonia solanacearum 311 IAYWNLFCLAQALLPLCGS.....DPTAFTDLSDEAQAQPAIDAAQEALLVYRDTYGEAFYARYRAKLGL
Xylella fastidiosa 314 VAYWNLGCLAR..A.....LAPLFSDAASLQAGLERE.....RATYLAAERRDAAAKLGF
Homo sapiens 429 VQVELEEDGALVSKLLETMHLTGADFTNTFYLLSSFPVELESPGLAEFLARLMEQCASLEELRLAFRPQM
Mus musculus 432 IRVEKEEDGTLVAKLLETMHLTGADFTNTFCVLSSFPADLSDS..AEFLSRLTSQCASLEELRLAFRPQM
Neurospora crassa 449 LFTARLG....LKTYKE..SDF.....DSLFDSLLNTMEALELDYNLFFRRLSTLK....
Schizosaccharomyces pombe 397 LMFKRVG....LPS..D..SSN.....KILITDLLQILEDYELDMPNCFSFLSRNS....
Saccharomyces cerevisiae 449 LMSQRLGVDLDLEKCMS..STNLKDIEHAAEKAKEFCDVIVEPLLDILQATKVDYNNFFIHLQNYKGPFF
Arabidopsis thaliana 430 T....KYNKEVISKLLNNMSVDKVDYTNFFRLLANVKANPNTP.....
Escherichia coli 336 M.TEQKEDNALLNELFSLMARERSDYTRTFRMLS.....LTEQ.....
Salmonella typhimurium 338 F.TEQKDDNVLLNELFSLMAREGSDYTRTFRMLS.....HTEQ.....
Vibrio cholerae 359 A.TQQEGDGELFADFFALLANNHTDYTRFLRELL.....SCLD.....
Pseudomonas aeruginosa 342 G.VAAENDHALVQELLQRMQGSAVDYSLFFRRLG.....EETP.....
Ralstonia solanacearum 376 T.QAHDGDEALFGDLFKLLHTQRADYTLFFRHLADVRRDDTPA.....
Xylella fastidiosa 362 A.ACFDEDLELFDALRTCMHQAEMDMTLTFLGLADWE..PNMP.....
Homo sapiens 499 DPRQLSMMLMLAQSNPQLFALMGTRAGIARELERVEQQSRLEQLSAAELQSRNQGHWADWLQAYRARLDK
Mus musculus 500 DPRQLSMMLMLAQSNPQLFALIGTQANVTKELERVEHQSRLEQLSPSDLQRKNRDHWEAWLQEYRDRLDK
Neurospora crassa 494 ...TADLQT.....EEA.....RQKAAEVFFSQVEEVPGPGTDKE.KARKRVGEWLDKWRVRLEE
Schizosaccharomyces pombe 440 ...PSSMEN.....EEY.....AAKLMQACICL.....NPNNE.RVRNESVKAFTNWVGRYSE
Saccharomyces cerevisiae 517 IKDKSDTATLFGAFDEEYLGMFFNSKQLQQMAETEEAFAAGEKVFDANGELRLLNEKLQELRNWTQDY..
Arabidopsis thaliana 469 .....ENELLKPLKAVLLDIGKERKDAWIK.WMRSYIQ
Escherichia coli 373 .....HSAASPLRDEFID...RAAFLDWFARYRGRLQQ
Salmonella typhimurium 375 .....QSASSPLRDTFID...RAAFLAWFDRYRARLRT
Vibrio cholerae 395 .....RQGNEAVLDLVLD...REAAKTWLTRYLERAAR
Pseudomonas aeruginosa 379 .....ERALASLRDDFVD...REAFLRWAEAYRRVEE
Ralstonia solanacearum 418 .....QAQARTVRDVFFD...RDSALAWLAAYRQRLQT
Xylella fastidiosa 402 .....DS.LSLWAEAFYDPVKRDAQAPMLRDWLQRYAA


Homo sapiens 569 DLEGAGDAAAWQAEHVRVMHANNPKYVLRNYLAQNAIEAAE.RGDFSEVRRVLKLLETPYHCEAGAATDA
Mus musculus 570 EKEGVGDTAAWQAERVRVMRANNPKYVLRNYLAQKAIEAAE.NGDFSEVRLVLKLLESPYHSEE.EATGP
Neurospora crassa 545 D...WTTSAADSEERVAAMKRVNPSFIPRGWILDEVIRRVEKQGERDVLKRVLHMATHPFEDAWTGKEFE
Schizosaccharomyces pombe 484 .....ATKTQEDSSRLASMKKVNPHFTLRNWVLEEVIKEA.YIGKFELFKKVCKMAAGPFEDTW......
Saccharomyces cerevisiae 585 L...TLVPPTETAARASLAKKANPLFVPRSWVLEEVVDDLMYSQRDGLQDPSSELDTSALKKLYLMSVNP
Arabidopsis thaliana 501 EVG..GSEVS.DEERKARMDSVNPKYILRNYLCQSAIDAAE.QGDFSEVNNLIRLMKRPYEEQPG.....
Escherichia coli 403 D.....EVS.DSERQQLMQSVNPALVLRNWLAQRAIEAAE.KGDMTELHRLHEALRNPFSDRD......
Salmonella typhimurium 405 E.....AVD.DALRQQQMQRVNPAIVLRNWLAQRAIDAAE.QGDMAELHRLHEVLRQPFTDRD......
Vibrio cholerae 425 ELGQEGRPIS.TRERCQAMRQVNPKYILRNYLAQQAIEFAE.RGDFEEMQRLATVLASPYAEHPE.....
Pseudomonas aeruginosa 409 EGGDQE...S.RRRR...MHAVNPLYVLRNYLAQQAIEAAE.QGDYTEVRLLHQVLSRPFEEQPG.....
Ralstonia solanacearum 448 E.....PAP.DAARAEAMRRVNPKYVLRNHLAPIAIRRAG.EKDFSEVENLRAVLARPFDDHPG.....
Xylella fastidiosa 434 RLS..VDPLP.VAERHERMRLANPRYVLRNYLTQQAIECAE.QGDLIELHALLEVMRRPYDFQLG.....

Homo sapiens 638 EATEADGADGRQRSYSSKPPLWAA.E....LCVTUSS
Mus musculus 638 EAVARSTEE..QSSYSNRPPLWAA.E....LCVTUSS
Neurospora crassa 612 DGPTGKGVYQGDKAEEERW.TGDVPQQKKAMQCSCSS
Schizosaccharomyces pombe 542 ......GF...SKEEEDYL.CYNTTPSKSQIQCSCSS
Saccharomyces cerevisiae 652 YDRTKWDVTLRPELETKWADLSHQDDAKFMMQASCSS
Arabidopsis thaliana 562 ..........MEKYARLPPAWA..YRPGVCMLSCSS
Escherichia coli 459 ..........DDYVSRPPDWCK.R....LEVSCSS
Salmonella typhimurium 461 ..........DDYARRPPEWCK.R....LEVSCSS
Vibrio cholerae 488 ..........FERYAKLPPEWCK.K....LEISCSS
Pseudomonas aeruginosa 466 ..........MERFTRRPPDWCR.H....LEISCSS
Ralstonia solanacearum 505 ..........FEHYAGPAPDWAA.S....LEVSCSS
Xylella fastidiosa 495 ..........REAYAMRRPEWAR.SRIGCSMLSCSS

Figure S8. Glutathione peroxidase 6 (GPx6) alignment. Accession numbers for sequences are as follows: *Homo sapiens* (this study), *Sus scrofa* (this study), *Mus musculus* (AAH13526) and *Rattus norvegicus* (AAA42094).

Figure S9. SECIS elements in human SelI gene and orthologous vertebrate genes. Structures of SelI SECIS elements from indicated organisms were generated with SECISearch and visualized with RNAnice. Conserved nucleotides in the quartet and Apical loop are shown in bold.

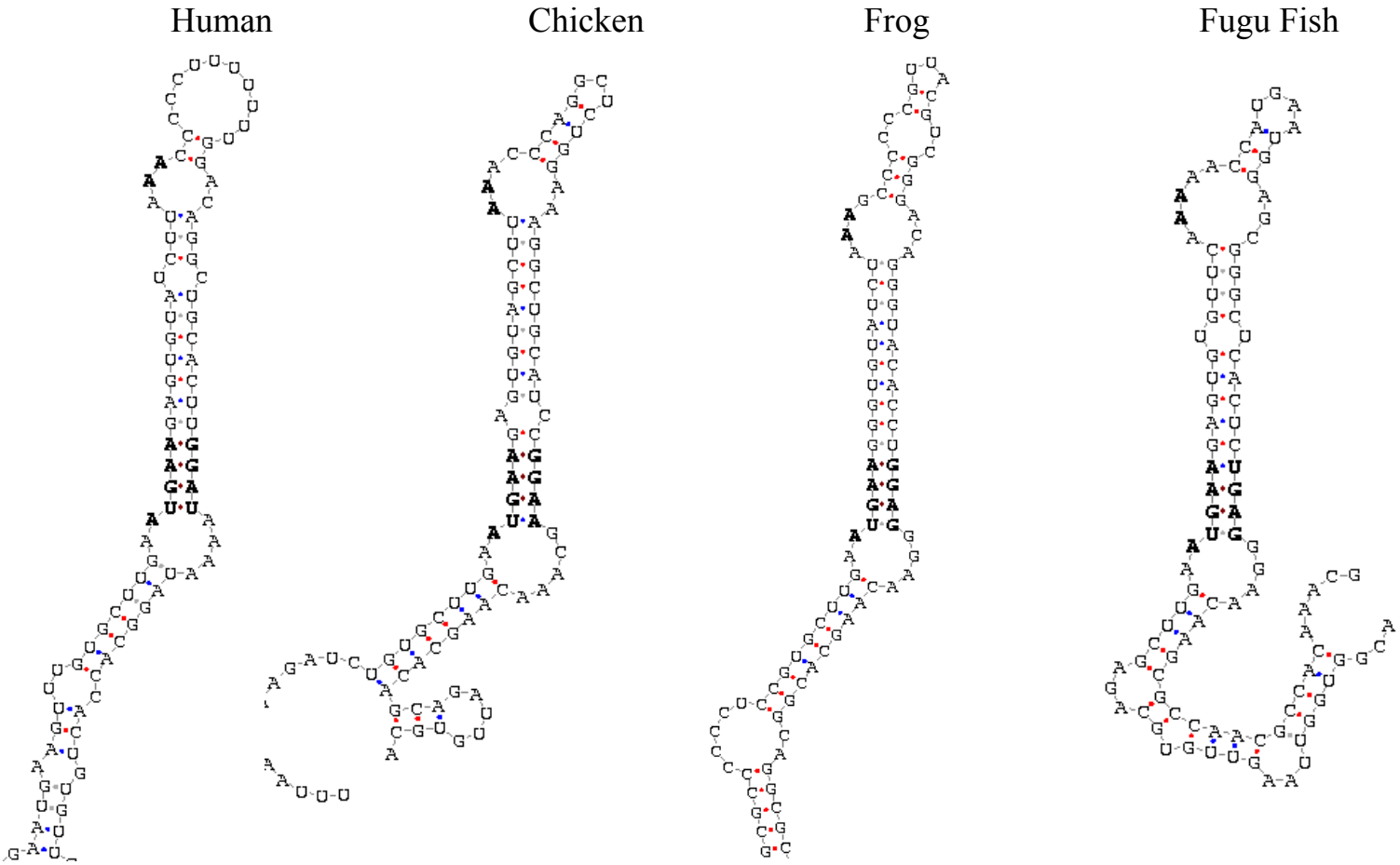


Figure S10. Structures of newly identified human selenoprotein genes. Structures and chromosomal locations of newly identified selenoprotein genes were obtained by aligning selenoprotein cDNA sequences to the GoldenPath human genome assembly (Aug 2001 release) using BLAT program (http://genome.ucsc.edu/cgi-bin/hgBlat?command=start) (*9*). Introns are shown by horizontal lines, coding regions by filled boxes and untranslated regions by open boxes. Chromosomal locations are indicated below exons and locations of SECIS elements and initiation, selenocysteine and termination codons are shown above the sequences.

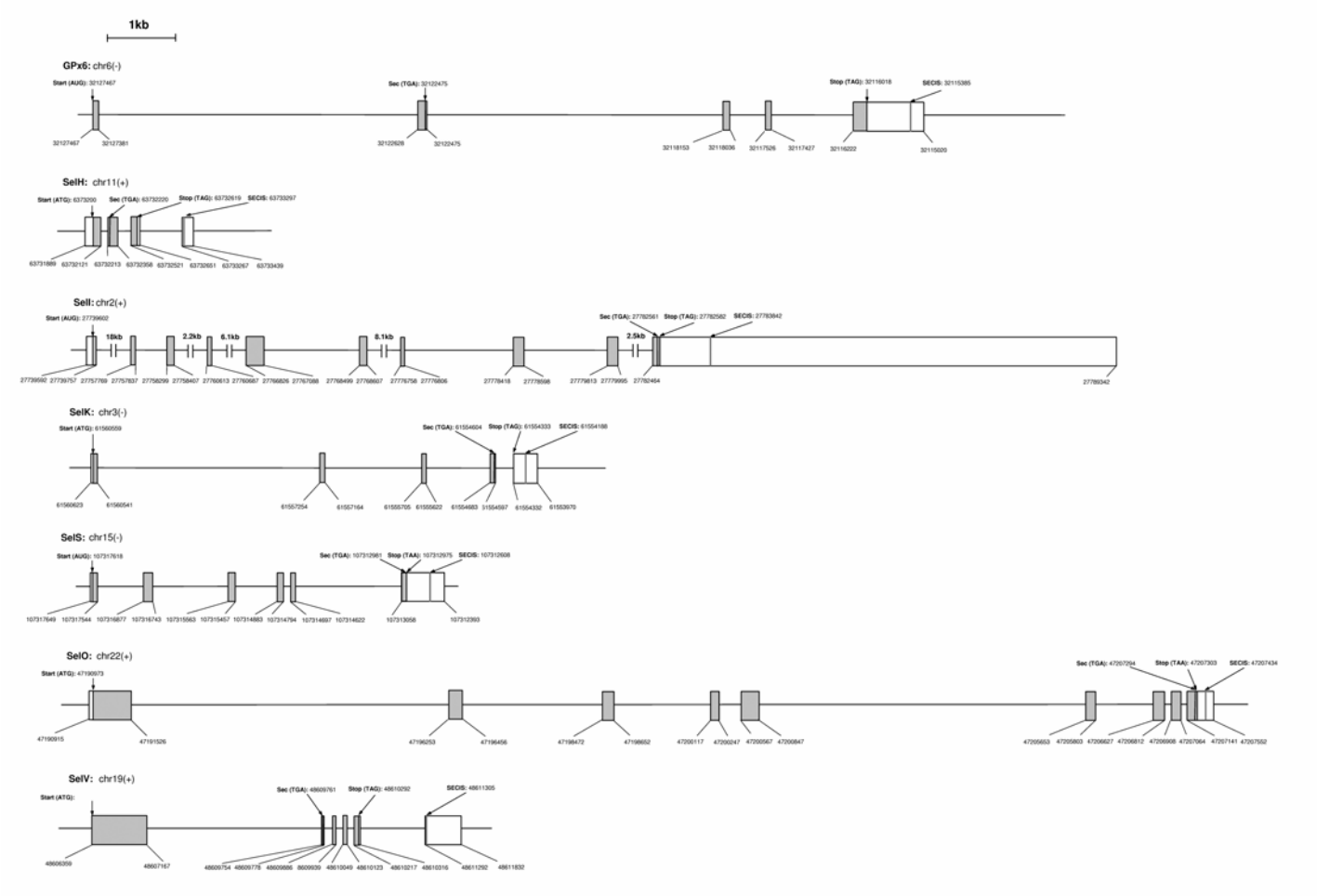


Figure S11. Alignment of SECIS elements in human selenoprotein genes. The human genome has 26 SECIS elements, including 17 structures in 17 previously identified genes, 7 structures in the 7 selenoprotein genes identified in the present study, and two elements in the SelP gene. The 26 SECIS elements were manually aligned on the basis of their primary sequence and secondary structure features. Nucleotides that are strictly conserved in all SECIS elements and nucleotides that are mostly conserved are shown by black and grey backgrounds, respectively.

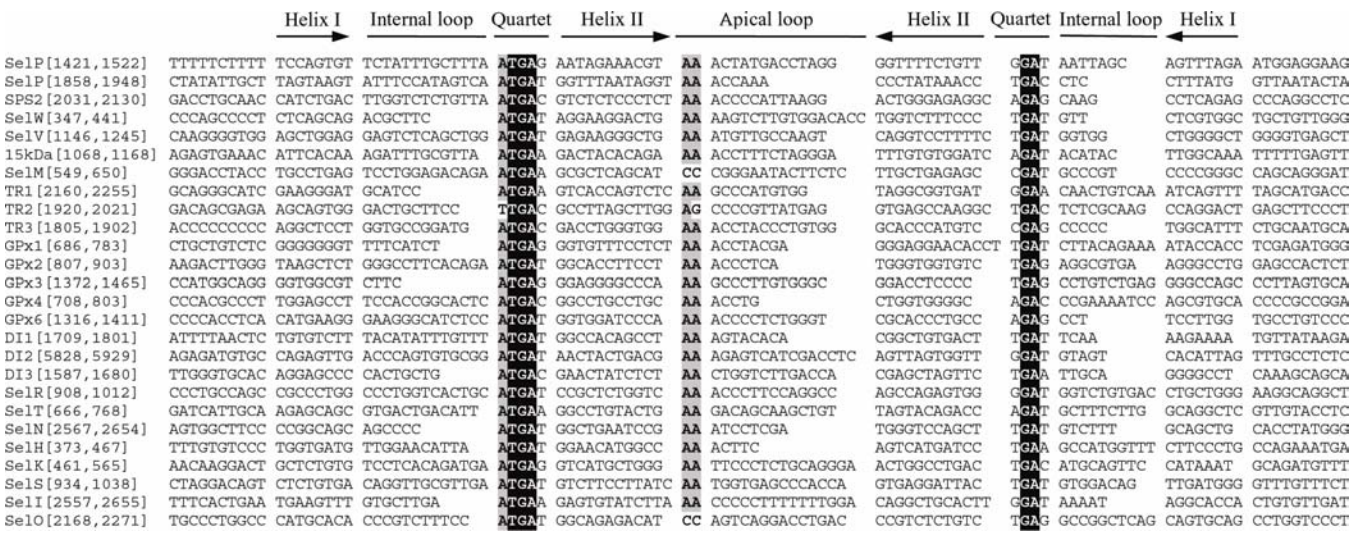


Figure S12. Selenoprotein genes in completely sequenced eukaryotic genomes.

Organism	Genome size	Estimated number of genes	Number of selenoprotein genes
Homo sapiens	3,400,000,000	~40,000	25
Mus musculus	3,454,200,000	~40,000	24
Drosophila melanogaster	137,000,000	14,331	3
Caenorhabditis elegans	97,000,000	20,206	1
Arabidopsis thaliana	100,000,000	~25,000	0
Saccharomyces cerevisiae	12,067,280	6,312	0
Schizosaccharomyces pombe	13,800,000	4,824	0

Figure S13A. Web-based version of SECISearch. Input page of the program (available at http://genome.unl.edu/SECISearch.html).

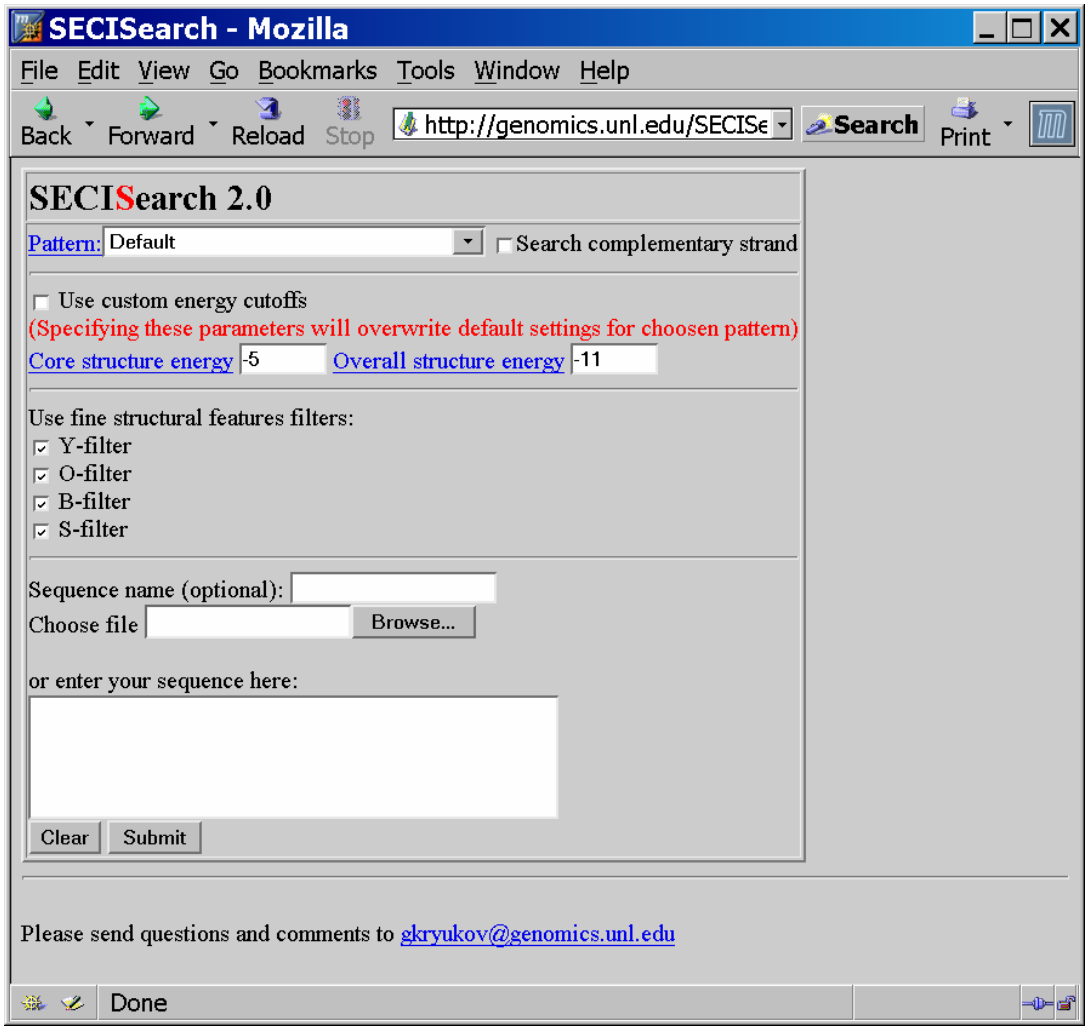
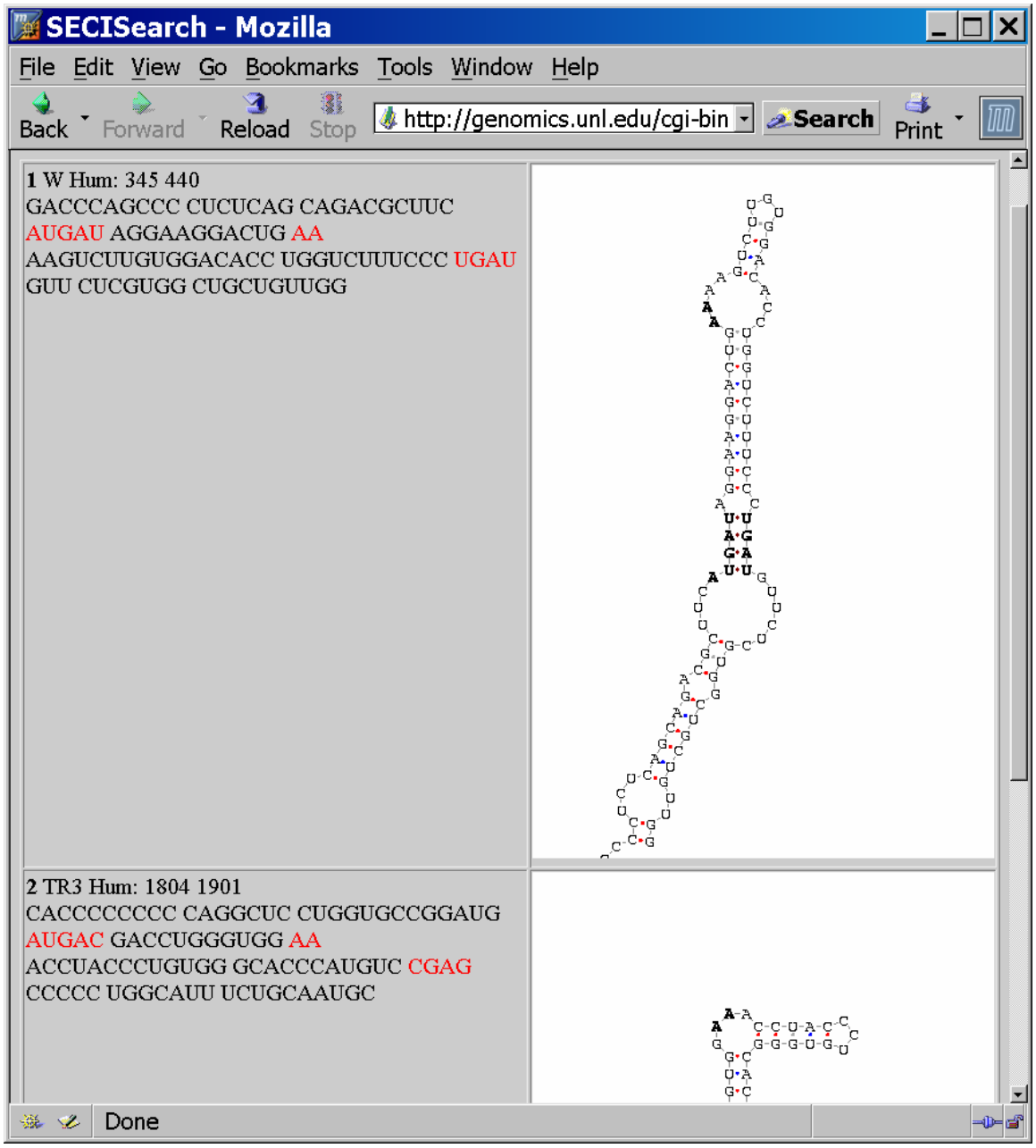


Figure S13B. Web-based version of SECISearch. Output of the program. The output shows locations of SECIS elements in query sequences and visualizes SECIS elements with RNAnice.



Supporting references

1. G. V. Kryukov, V. M. Kryukov, V. N. Gladyshev, *J. Biol. Chem.* **274**, 33888 (1999).
2. I. L. Hofacker *et al*., *Monatshefte f. Chemie* **125**, 167 (1994).
3. E. Grundner-Culemann *et al*., *RNA* **5**, 625 (1999).
4. S. F. Altschul *et al*., *J. Mol. Biol.* **215**, 403 (1990).
5. G. Parra, E. Blanco, R. Guigó, *Genome Res.* **10**, 511 (2000).
6. G. V. Kryukov, V. N. Gladyshev, *Methods Enzymol.* **347**, 84 (2002).
7. G. Parra, *Genome Res.* **13**, 108 (2003).