Supplementary Data to:

Selenoprofiles: profile-based scanning of eukaryotic genome sequences for selenoprotein genes

Marco Mariotti and Roderic Guigó

Section S1: patterns used with SECISearch
We report here the patterns used with SECISearch in the current implementation of selenoprofiles. The syntax is the one used by PatScan, which is run under the hood by SECISearch. We are currently working to improve the patterns in terms of both specificity and sensitivity, so these may change soon.

Standard:
r1={au,ua,gc,cg,gu,ug} NNNNNNNNNN p1=7...7 3...13 ATGAN p2=10...13 AA (4...12 | 0...3 p3=3...6 3...6 r1~p3 0...3) (r1~p2[2,1,1] NGAN | r1~p2[2,1,0] NNGAN) 3...10 r1~p1[1,1,1] NNNNNNNNNN

Non-Standard:
r1={au,ua,gc,cg,gu,ug} NNNNNNNNNN p1=7...7 3...13 NNGAN p2=10...13 NN (4...13 | 0...2 p3=3...4 3...4 r1~p3 0...2) (r1~p2[1,1,1] NGAN | r1~p2[1,1,0] NNGAN) 3...10 r1~p1[1,1,1] NNNNNNNNNN

Twilight:
r1={au,ua,gc,cg,gu,ug} NNNNNNNNNN p1=7...7 3...13 NTGAN p2=10...13 (AR | CC) (4...12 | 0...3 p3=3...6 3...6 r1~p3 0...3) (r1~p2[2,1,1] NGAN | r1~p2[2,1,0] NNGAN) 3...10 r1~p1[1,1,1] NNNNNNNNNN

Table S2: List of releases of the Ensembl core databases used in this work. The genome release is 52 for all species except Vicugna Pacos for which is 51.

Species name	Ensembl core database release
Aedes aegypti	aedes_aegypti_core_52_1d
Anopheles gambiae	anopheles_gambiae_core_52_3k
Bos taurus	bos_taurus_core_52_4b
Caenorhabditis elegans	caenorhabditis_elegans_core_52_190
Canis familiaris	canis_familiaris_core_52_2j
Cavia porcellus	cavia_porcellus_core_52_3a
Ciona intestinalis	ciona_intestinalis_core_52_2l
Ciona savignyi	ciona_savignyi_core_52_2h
Danio rerio	danio_rerio_core_52_7e
Dasypus novemcinctus	dasypus_novemcinctus_core_52_1h
Dipodomys ordii	dipodomys_ordii_core_52_1a
Drosophila melanogaster	drosophila_melanogaster_core_52_54a
Echinops telfairi	echinops_telfairi_core_52_1g
Equus caballus	equus_caballus_core_52_2b
Erinaceus europaeus	erinaceus_europaeus_core_52_1e
Felis catus	felis_catus_core_52_1f
Gallus gallus	gallus_gallus_core_52_2j
Gasterosteus aculeatus	gasterosteus_aculeatus_core_52_1i
Gorilla gorilla	gorilla_gorilla_core_52_1
Homo sapiens	homo_sapiens_core_52_36n
Loxodonta africana	loxodonta_africana_core_52_1g
Macaca mulatta	macaca_mulatta_core_52_10j
Microcebus murinus	microcebus_murinus_core_52_1b
Monodelphis domestica	monodelphis_domestica_core_52_5g
Mus musculus	mus_musculus_core_52_37e
Myotis lucifugus	myotis_lucifugus_core_52_1g
Ochotona princeps	ochotona_princeps_core_52_1c
Ornithorhynchus anatinus	ornithorhynchus_anatinus_core_52_1i
Oryctolagus cuniculus	oryctolagus_cuniculus_core_52_1h
Oryzias latipes	oryzias_latipes_core_52_1h
Otolemur garnettii	otolemur_garnettii_core_52_1e
Pan troglodytes	pan_troglodytes_core_52_21j
Pongo pygmaeus	pongo_pygmaeus_core_52_1c
Procavia capensis	procavia_capensis_core_52_1a
Pteropus vampyrus	pteropus_vampyrus_core_52_1a
Rattus norvegicus	rattus_norvegicus_core_52_34u
Saccharomyces cerevisiae	saccharomyces_cerevisiae_core_52_1i
Sorex araneus	sorex_araneus_core_52_1e
Spermophilus tridecemlineatus	spermophilus_tridecemlineatus_core_52_1g
Takifugu rubripes	takifugu_rubripes_core_52_4k
Tarsius syrichta	tarsius_syrichta_core_52_1a
Tetraodon nigroviridis	tetraodon_nigroviridis_core_52_8b
Tupaia belangeri	tupaia_belangeri_core_52_1f
Tursiops truncatus	tursiops_truncatus_core_52_1a
Vicugna pacos	vicugna_pacos_core_51_1
Xenopus tropicalis	xenopus_tropicalis_core_52_41l

Table S3: Performances indices of selenoprofiles testing on human, drosophila and yeast genome. All families cited in the main article plus MsrA were considered. As reference, we considered the exonic structures annotated in Ensembl Core database, fetching the most similar to each selenoprofiles prediction. All annotations fetched in this way were then checked manually and compared with SelenoDB to make sure that both the selenoproteins were correctly annotated and that all genes were considered. In a some cases (drosophila SelK, SelH, SPS2 and human SelK, SelH, SelS, SelT, SelV, SelW1, TR1, TR2 and TR3) the fetched annotation was not carrying the selenocysteine residue, therefore it was modified to respect the annotation in SelenoDB. For machinery proteins not included in SelenoDB (SecS, PSTK, secp43), the annotations were selected among the selenoprofiles candidates analyzing the gene description in Ensembl. For some drosophila genes no description was available and the gene was selected after a manual sequence analysis. The annotations are split in three sets: selenoproteins, non-Sec homologues and machinery proteins. The selenoprotein set was compared with all selenoprofiles predictions with label "selenocysteine", while the homologues set was compared with the predictions with any other label. The machinery set was compared with all selenoprofiles predictions for machinery protein families.
Sensitivity (SN) and specificity (SP) were computed at the gene, exon, and nucleotide level. At the gene level, the number of false positives (FP) is reported instead of specificity. The exon level indexes are computed considering only the genes that were correctly paired between the predictions and the annotations, while the nucleotide indexes are computed considering everything. The average indexes at the end of the table are computed pulling together all genes for each set.

gene	level	exon	level	nucleotide	level	family, class, gene numbers
SN	FP	SN	SP	SN	SP	
						Homo sapiens
1	0	0	0	1	1	sps-selenoproteins: 1 gene
1	0	0.57	0.75	0.89	1	GPx-selenoproteins: 5 genes
1	0	0.63	0.71	0.98	0.97	DI-selenoproteins: 3 genes
1	0	1	1	1	1	15-kDa-selenoproteins: 1 gene
1	0	1	1	1	1	SelM-selenoproteins: 1 gene
1	0	1	1	1	1	SelH-selenoproteins: 1 gene
1	0	0.9	0.9	1	0.97	SelI-selenoproteins: 1 gene
1	1	0.6	0.75	1	0.5	SelK-selenoproteins: 1 gene
1	0	0.83	0.91	0.89	1	SelN-selenoproteins: 1 gene
1	0	1	1	1	1	SelO-selenoproteins: 1 gene
1	0	1	1	1	1	SelP-selenoproteins: 1 gene
1	0	1	1	1	1	SelR-selenoproteins: 1 gene
1	2	1	1	1	0.46	SelS-selenoproteins: 1 gene
1	1	0.8	0.8	0.96	0.53	SelT-selenoproteins: 1 gene
0.5	1	0.8	0.67	0.74	0.79	SelV-selenoproteins: 2 genes
1	0	0.91	0.89	0.99	0.92	TR-selenoproteins: 3 genes
1	2	0.88	0.88	0.96	0.4	sps-homologues: 1 gene
1	0	0.45	0.56	0.72	0.99	GPx-homologues: 3 genes
1	0	1	1	1	1	MsrA-homologues: 1 gene
/	2	/	/	/	/	SelJ-homologues: 0 genes
/	2	/	/	/	/	SelK-homologues: 0 genes
1	0	0.82	0.9	0.86	1	SelR-homologues: 2 genes
/	1	/	/	/	/	SelT-homologues: 0 genes
1	0	0.78	0.78	0.99	0.95	SelU-homologues: 3 genes
0	0	0	0	0	0	SelV-homologues: 1 gene
/	2	/	/	/	/	TR-homologues: 0 genes
1	1	0.76	0.81	0.99	0.43	sbp2-machinery: 1 gene
1	0	0.5	0.5	0.79	0.81	pstk-machinery: 1 gene
1	0	0.22	0.5	0.32	0.93	secp43-machinery: 1 gene
1	0	1	1	1	1	SecS-machinery: 1 gene
1	0	1	1	1	1	eEFsec-machinery: 1 gene

						Drosophila melanogaster
1	0	0.25	0.25	0.91	1	sps-selenoproteins: 1 gene
1	0	0	0	0.58	0.89	SelH-selenoproteins: 1 gene
1	0	1	1	1	1	SelK_insect-selenoproteins: 1 gene
1	0	0	0	0.99	1	sps-homologues: 1 gene
1	1	0.33	0.5	0.68	0.51	GPx-homologues: 1 gene
1	0	0	0	0.3	0.95	MsrA-homologues: 1 gene
1	0	0.33	0.5	0.92	1	15-kDa-homologues: 1 gene
/	1	/	/	/	/	SelM-homologues: 0 genes
1	0	0	0	0.92	0.88	SelH-homologues: 2 genes
1	3	0.5	0.4	0.88	0.33	SelI-homologues: 1 gene
/	1	/	/	/	/	SelK-homologues: 0 genes
0	0	0	0	0	0	SelK_insect-homologues: 1 gene
1	0	0.75	0.75	1	0.95	SelR-homologues: 1 gene
1	0	1	1	1	1	SelT-homologues: 1 gene
/	1	/	/	/	/	SelV-homologues: 0 genes
1	2	0.6	0.6	0.92	0.71	TR-homologues: 2 genes
1	0	1	1	1	1	sbp2-machinery: 1 gene
1	1	0	0	1	0.54	pstk-machinery: 1 gene
1	1	0.5	0.33	0.94	0.54	secp43-machinery: 1 gene
1	0	0.5	0.5	1	0.95	SecS-machinery: 1 gene
1	0	1	1	1	1	eEFsec-machinery: 1 gene
						Saccharomyces cerevisiae
1	0	0	0	0.97	1	GPx-homologues: 3 genes
1	0	0	0	0.61	1	MsrA-homologues: 1 gene
1	0	0	0	0.26	1	SelO-homologues: 1 gene
1	1	0	0	0.62	0.39	SelR-homologues: 1 gene
/	3	/	/	/	/	TR-homologues: 0 genes
/	1	/	/	/	/	pstk-machinery: 0 genes
Average (FP column refers to the total number)						
0.96	5	0.81	0.85	0.94	0.91	selenoproteins
0.97	22	0.57	0.6	0.8	0.58	homologues
1	4	0.71	0.77	0.93	0.68	machinery

Section S4: Exonerate vs genewise

In the following table, we report the global performance indices when we force the pipeline to choose always the exonerate or always the genewise prediction. When the standard routine of selenoprofiles is used (one of the two predictions is chosen according to the criteria detailed in the text) the indices improve or are the same.

gene	level	exon	level	nucleotide	level	class
SN	FP	SN	SP	SN	SP	
Average (FP column refers to the total number) choosing EXONERATE						
0.89	3	0.78	0.83	0.86	0.93	selenoproteins
0.9	14	0.6	0.63	0.73	0.65	homologues
0.9	4	0.74	0.72	0.91	0.68	machinery
Average (FP column refers to the total number) choosing GENEWISE						
0.96	5	0.8	0.85	0.94	0.91	selenoproteins
0.93	20	0.5	0.56	0.76	0.59	homologues
0.9	4	0.67	0.76	0.82	0.67	machinery

We observe that genewise is generally performing better than exonerate. Nonetheless, genewise is much slower than exonerate (it would not be feasible to use the cyclic procedure for genewise), so we believe that the best way to combine them is to use exonerate to outline the gene boundaries and genewise to refine the prediction. Anyway, since genewise appears to be more sensitive than exonerate, we created the genewise_to_be_sure routine (see text in the main manuscript) to ensure that we do not lose any potential candidates that would be missed by exonerate but caught by genewise. Also, in our experience genewise crashes systematically for some predictions (although it never crashed for the predictions in the testing set). We believe this is due to the fact that it was never tested with our particular scoring scheme, which may confound its computation. When this happens, selenoprofiles uses exonerate prediction instead, and this is another advantage of having two predictions available.

Section S5: Discussion of false positives

1. <u>Selenocysteine labelled</u>

In the human genome, 5 genes for which no annotation was found were predicted and labelled as "selenocysteine". One belongs to the SelT family. This is characterized by a single-exon structure, and no potential SECIS was identified downstream. An additional analysis revealed that the conservation of the coding sequence extends in the 5' side for an additional portion respect to selenoprofiles prediction. This extension contains a frameshift. All these facts make us believe that this is a recent retro-transcribed pseudogene.

Two selenocysteine containing SelS genes were predicted. In both cases a poor scoring SECIS element was found downstream of the predicted coding sequence. The SelS family is characterized by domains of repetitive sequences, rich in lysine, glutamic acid and glycine. These domains causes the profile to hit the genome in a lot of locations. In both predicted genes, the conservation with the profile is too poor to conclude that these are real genes: excluding the regions of repetitive sequence, we found no significant similarity with any other known protein. It is very likely that these predictions have said selenoprotein features just by chance.

Then, a selenocysteine containing SelK gene was predicted. This gene is characterized by a single-exon structure, and two poor scoring SECIS elements were found downstream. No annotation corresponding to this gene was found in Ensembl. Nonetheless, a search with blast found an human hypothetical protein (gi code: 169213282), matching with 100% identity the selenoprofiles prediction but stopping at the UGA position. A blast search in ncbi human EST dataset resulted in no perfect matches, suggesting that this genomic region is not transcribed. The single exon structure and the absence of transcription suggest the occurrence of a retro-transcribed pseudogene.

Lastly, a selenocysteine containing SelV gene was predicted, consisting of two exons with two poor scoring SECIS elements downstream. This corresponds to the Ensembl pseudogene ENSG00000215900. Searching ncbi human ESTs, we found no evidence of transcription. We think that this is most likely a pseudogene, too.

2. <u>Selenocysteine machinery proteins</u>

For these proteins, 4 false positives were predicted in total in the human, fly and yeast genome by selenoprofiles. Two false PSTKs were predicted, one in drosophila and one in yeast. The PSTK proteins share a domain with high similarity with another protein family, KTI12, and this causes selenoprofiles to find also KTI12 proteins when searching the PSTK profile in genomes.

One false SECP43 protein was predicted in drosophila. This is actually a portion of the protein Rox8 (or RE71384p), since it shares a nucleotide binding domain with SECP43.

Lastly, the human protein SBP2-like is found using the SBP2 profile. These two proteins diverged recently, during vertebrate evolution (see Donovan et al, "Evolutionary history of selenocysteine incorporation from the perspective of SECIS binding proteins", BMC evolutionary biology, 2009). They share high sequence similarity and, possibly, they are also functionally linked.