594–606 Nucleic Acids Research, 2019, Vol. 47, No. 2 Published online 7 December 2018
doi: 10.1093/nar/gky1234

RefSeq curation and annotation of stop codon recoding in vertebrates

Bhanu Rajput, Kim D. Pruitt and Terence D. Murphy *

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received August 17, 2018; Revised November 20, 2018; Editorial Decision November 27, 2018; Accepted December 03, 2018

ABSTRACT

Recoding of stop codons as amino acid-specifying codons is a co-translational event that enables C-terminal extension of a protein. Synthesis of selenoproteins requires recoding of internal UGA stop codons to the 21st non-standard amino acid selenocysteine (Sec) and plays a vital role in human health and disease. Separately, canonical stop codons can be recoded to specify standard amino acids in a process known as stop codon readthrough (SCR), producing extended protein isoforms with potential novel functions. Conventional computational tools cannot distinguish between the dual functionality of stop codons as stop signals and sense codons, resulting in misannotation of selenoprotein gene products and failure to predict SCR. Manual curation is therefore required to correctly represent recoded gene products and their functions. Our goal was to provide accurately curated and annotated datasets of selenoprotein and SCR transcript and protein records to serve as annotation standards and to promote basic and biomedical research. Gene annotations were curated in nine vertebrate model organisms and integrated into NCBI's Reference Sequence (RefSeq) dataset, resulting in 247 selenoprotein genes encoding 322 selenoproteins, and 93 genes exhibiting SCR encoding 94 SCR isoforms.

INTRODUCTION

Recoding is an mRNA-specific, non-standard decoding of the genetic code that is stimulated by cis-acting signals. The genetic information transmitted to mRNA via transcription can be redefined during translation by several types of recoding events, including ribosomal frameshifting and translational or stop codon readthrough. During translational readthrough, a stop codon is co-translationally recoded to specify an amino acid, which results in translation termination at an in-frame downstream stop codon and extension of the protein. Two types of stop codon recoding events are known to occur in eukaryotes (Table 1): selenocysteine (Sec) insertion, which involves recoding a UGA stop codon to specify Sec for the synthesis of full-length Sec-containing proteins, selenoproteins (Figure 1A); and stop codon readthrough (SCR), which involves recoding any of the three stop codons (UAA, UAG or UGA) to specify a standard amino acid to generate C-terminally extended protein isoforms (Figure 1B).

Selenoproteins are found in all three domains of life: bacteria, archaea and eukaryota. They contain the essential trace element selenium (Se), which is of fundamental importance in humans and animals (1). Most of the biological effects of dietary Se are mediated by Sec, a Se-containing amino acid, which is co-translationally inserted into nascent polypeptides by recoding UGA codon via a sophisticated machinery involving cis- and trans- acting factors. A conserved stem-loop structure, the Sec insertion sequence (SECIS) element, present in the 3′ UTRs of selenoprotein mRNAs in eukaryotes (Figure 1A), is essential for the recognition of UGA as a Sec codon rather than a stop signal. The requirement of the SECIS element for Sec insertion and its conservation have been exploited to identify selenoprotein genes in sequenced vertebrate genomes. The first complete set of selenoproteins (i.e. selenoproteome) in vertebrates was determined for human and consists of 25 selenoproteins (2). Subsequently, selenoprotein genes in 44 vertebrate species were identified and analyzed to study the composition and evolution of selenoproteomes (3). The size of the selenoproteome varies greatly among eukaryotes; larger selenoproteomes have been observed in aquatic organisms, such as fish, which contain several selenoproteins not found in mammals. Except for the characteristic presence of Sec in all selenoproteins, they are a diverse group with varied sequences and involved in a variety of processes, ranging from antioxidant defense, fertility, immune function, muscle development, thyroid hormone metabolism, and in cancer prevention and promotion (1,4,5). The specific functions of many selenoproteins are not known. Many of the best characterized selenoproteins, such as the members of the DIO, GPX and TXNRD gene families, function as enzymes containing a Sec residue at the catalytic site, which is

*To whom correspondence should be addressed. Tel: +1 301 402 0990; Fax: +1 301 480 2918; Email: murphyte@ncbi.nlm.nih.gov

Published by Oxford University Press on behalf of Nucleic Acids Research 2018.

Table 1. Comparison of Sec insertion and SCR

Features	Sec insertion	SCR
Upstream stop codon	UGA	UGA>UAG>UAA[a]
Downstream stop codon	UAA, UAG, UGA	UAA, UAG, UGA
Amino acid specified	Sec	Several standard amino acids
Stimulatory signal	SECIS element	CUAG, stem loop, other
End product	Selenoprotein	C-terminally extended isoform

[a] Frequency of occurrence as readthrough codon.

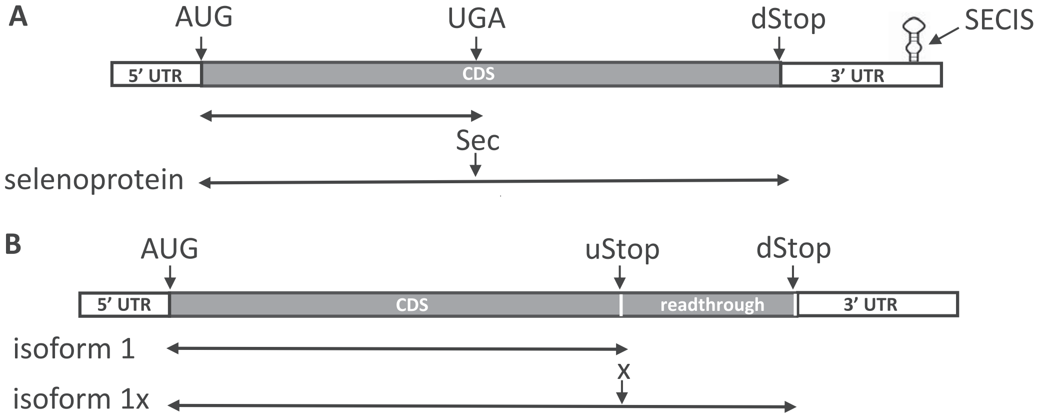


Figure 1. Schematic representing Sec insertion (**A**) and SCR (**B**). The coding regions are shown as filled black bars, and the 5′ and 3′ UTRs as open bars. (**A**) Selenoprotein mRNA containing an AUG start codon, in-frame UGA and downstream stop (dStop) codons, and SECIS element is shown. Double arrowed lines indicate proteins resulting from translation termination at UGA (top) or dStop following recoding of UGA as Sec (bottom). (**B**) SCR candidate mRNA containing an AUG start codon, canonical upstream stop (uStop) and in-frame downstream stop (dStop) codons, and inter-stop readthrough region encoding C-terminal extension is shown. Double arrowed lines indicate isoforms resulting from translation termination at uStop (top) or dStop following recoding of uStop to specify a standard amino acid (denoted as 'x') (bottom).

thought to participate in redox reactions. GPX4 is unusual in that it has a dual function in male reproduction: as a peroxidase in spermatids, and as a structural protein in mature spermatozoa (6). SELENOP, another well studied selenoprotein with multiple Sec residues, is implicated in selenium transport (7). SELENON has a role in myogenesis; mutations in this gene have been associated with several muscle disorders, collectively known as SEPN1-related myopathies (8).

SCR is used extensively in viruses to expand the proteomes encoded by their small genomes (9). However, SCR in eukaryotes was known only for a few cellular genes (10–12), until recently when two complementary approaches, a phylogenetic analysis using protein-coding evolutionary signatures (13) and ribosome profiling (14), identified abundant SCR in Drosophila, and to a lesser degree in yeast and mammals. Together, these and subsequent studies (15–19) predicted the presence of over 50 SCR candidate genes in mammals and showed that the C-terminal extensions in several mammalian candidate genes were also conserved in other vertebrates. Experimental evidence for readthrough, however, exists for just a few genes. It is believed that unlike selenoproteins for which the property of Sec is functionally important, for SCR it is the C-terminal extension and not the identity of the specified amino acid that is functionally relevant in the majority of cases. Beier and Grimm (20) showed that the misreading of stop codons in eukaryotes is achieved by a variety of naturally occurring tRNAs that recognize near-cognate codons. Thus, each stop codon has been reported to specify more than one amino acid: Arg, Cys, Trp, Ser for UGA; Gln, Tyr, Leu, Lys for UAG; and Gln and Tyr for UAA (10,13,20). Studies have shown that the stop codon context, including the stop codon itself, plays a significant role in influencing SCR efficiency. The hierarchy of the leakiest SCR-prone codon observed is UGA>UAG>UAA, and the leakiest downstream base is C>U>G>A, suggesting that the UGA-C context is most effective in enabling SCR in eukaryotes (13). Other elements, including adjacent short sequence motifs (16,21), 3′ *cis*-acting elements (15), and stem loop structures (22,23), have also been implicated in stimulating SCR. A trans-acting factor has also been reported to promote SCR (15). In experimentally verified genes, those with the highest readthrough efficiency had a UGA stop codon immediately followed by a short motif, CUAG (16). Little is known about the functional importance of the readthrough isoforms; however, recent studies have shown that the extended C-terminus may expose additional signals or domains, which may modulate their localization, activity, and/or stability. For instance, the extended isoforms of *LDHB* and *MDH1* were targeted to the peroxisomes compared to the cytosolic canonical isoforms (18,19), the extended isoform of *VEGFA* was reported to have antiangiogenic activity (15) but this has been disputed recently (24), and the extended isoforms of *MPZ* (12), *AQP4* (25) and *VDR* (17) exhibit different properties that may modulate the original gene function.

Recoding of stop codons to sense codons bends the rules of standard decoding of the genetic code. Conventional computational tools cannot distinguish between the dual functionality of UGA codon as a stop signal and as Sec, resulting in misannotation of selenoprotein coding sequences (CDSes) on transcript and genomic sequences, and loss of crucial biological information. Predicting SCR systematically is even more challenging, given three potential readthrough stop codons and the diversity of stimulatory signals promoting readthrough. Manual curation is thus essential for accurate representation of these recoding events and their gene products. Our goal was to provide an accurately curated and annotated set of selenoprotein transcript and protein records and represent experimentally validated cases of SCR in several vertebrate model organisms in NCBI's Reference Sequence (RefSeq) database (26). The RefSeq database is a comprehensive collection of genomic, transcript, and protein sequences spanning the tree of life. The RefSeq collection is based on both automated and manual curation approaches, and the highly curated subset is vital for additional resources at NCBI, including Gene and Genomes, and helps guide NCBI's evidence-based eukaryotic genome annotation pipeline. RefSeq records are also widely used by the scientific community for basic and biomedical research, as well as for large scale genome annotation and comparative analyses.

MATERIALS AND METHODS

Several different approaches were used to identify selenoprotein genes in vertebrates similar to the methods described before for antizyme genes (27). Briefly, the earliest selenoprotein records in the RefSeq database were generated by automatic processes: provisional RefSeq records were created by automatic processing based on primary sequence data submitted to the International Nucleotide Sequence Database Collaboration (INSDC) (28); and predicted RefSeq records were generated by the NCBI's eukaryotic genome annotation pipeline. Many of the provisional and predicted RefSeq records contained misannotated selenoprotein CDS propagated from the primary sequence data or from computational prediction. These records were subjected to manual review as guided by our in-house QA analyses, which for instance would flag candidates for nonsense-mediated mRNA decay (NMD) (29). The manually curated records from well-studied organisms, such as human and mouse, facilitated prediction of orthologous models by the NCBI's eukaryotic genome annotation pipeline, which then became targets for manual review. Selenoprotein genes were also identified by literature search, search of specialized databases, such as SelenoDB 2.0 (30) and Recode-2 (31), and BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) search of publicly available databases at NCBI with related sequences. The SECISearch3 tool (32) was used to search for SECIS elements in the RefSeq transcript sequences. Selenoprotein gene nomenclature was updated where necessary as per recommendation by Gladyshev *et al.* (33), in collaboration with organism-specific nomenclature authorities (34–39). A brief RefSeq summary describing the salient features of a gene and its function (if known), a biological attribute ('protein contains selenocysteine') with supporting evidence for the recoding event, and an INSDC-approved regulatory feature annotation with a 'recoding_stimulatory_region' qualifier for the SECIS element were added to the curated RefSeq records. Additional phylogenetically conserved elements shown to modulate Sec insertion, such as the stop-codon redefinition element (SRE) in *SELENON* mRNA (40) and the two stem-loop structures (stem–loop 1 and 2) in *SELENOS* mRNA (41), were also annotated on the respective RefSeq transcript records. Separate RefSeq records were created for alternatively spliced transcript variants supported by transcript, EST, RNA-seq data or publications. Each transcript variant record is accompanied by a brief text highlighting its uniqueness and difference from the predominant variant.

SCR genes were selected for inclusion in the RefSeq database based on experimental support demonstrating readthrough and conservation of the C-terminal extension; readthrough candidate genes identified only by systematic bioinformatic or computational analysis were not included in this review. The RefSeq data model for eukaryotes allows only one transcript explicitly linked to one protein; therefore, to represent SCR, a second RefSeq transcript identical to the one encoding the canonical isoform was created and manually annotated with a CDS terminating at an alternative in-frame, downstream stop codon. As the identity of the amino acid specified by a readthrough stop codon is often not known, the amino acid inserted at the readthrough site was generally denoted as an undefined amino acid ('X'). In rare instances, when the specified amino acid was identified as a single species, as for VEGFA (15), then that specific amino acid (serine, S) was annotated at the readthrough site. The names of C-terminally extended isoforms were generally appended with an 'x' (for extended, as per convention in this field) to distinguish them from their standard counterparts (e.g. isoform 1 and isoform 1x), except when an alternate name was reported in literature (e.g. isoform MPZ and isoform L-MPZ) (12). The reviewed SCR records also included a RefSeq summary, transcript variant text, attribute ('stop codon readthrough') and a regulatory feature for SCR signal (if known) as described above for selenoprotein records.

The analysis, assembly and annotation of RefSeq records was performed using NCBI's Genome Workbench application (Gbench; http://www.ncbi.nlm.nih.gov/tools/gbench) as described before (27). RefSeqs are generally created from a combination of conventional transcript sequences (INSDC transcripts and ESTs), transcriptome shotgun assemblies (TSAs) and genomic sequence— the latter two are used when a gene or region(s) of a gene lacks conventional transcript support. Assembling a RefSeq transcript for the chicken *SEPHS2* gene posed a challenge (see Results) as none of the sources of sequence data mentioned above was available for a portion of the 3′ UTR of the *SEPHS2* transcript; therefore, a full-length RefSeq transcript (NM_001366334.2) was assembled using ESTs and two (18 and 52 bp) inserts based on RNA-seq reads available in SRA. The details of the assembly components are shown on the public NM_001366334.2 record under the COMMENT section.

RESULTS

Curation and annotation of selenoprotein genes in vertebrates

The dual role of the UGA codon as a stop signal and as encoding Sec poses a problem for automatic annotation of selenoprotein genes by standard computational tools, which often results in misannotation of selenoprotein CDSes on transcripts and genomic sequences, as illustrated by the mouse *Dio1* gene (Figure 2A and B). INSDC sequences AK290780.1 (human) and U49861.1 (mouse) were misannotated with a CDS terminating at TGA and encoding a C-terminally truncated protein of 125 amino acid residues (aa), rendering these transcripts candidates for NMD. A similar misannotated CDS was found on an Ensembl transcript (ENSMUST00000150974.1/Dio1-208). The misannotation was also automatically propagated to NM_007860.1, a provisional RefSeq transcript based on U49861.1. Subsequent manual review led to the replacement of the erroneous record with an updated version (NM_007860.4) containing the accurately annotated CDS encoding the full-length selenoprotein of 257 aa (NP_031886.3). A translation exception (/transl_except) qualifier was manually added to the CDS feature indicating the position of the TGA-specifying Sec using feature annotation details approved by the INSDC (Figure 2C). This allowed annotation of the full-length CDS and inclusion of Sec as 'U' in the selenoprotein sequence. The 'protein contains selenocysteine' attribute (Figure 2D) and feature annotations, including the locations of Sec and SECIS element (Figure 2E), were also added to the RefSeq transcript record to highlight the functionally relevant features of the selenoprotein transcript. Although UGA recoding to Sec has been reported to be inefficient (42) and modulated by Se availability (43), which can result in the production of both the full-length and truncated isoforms of selenoproteins, we have opted to only represent CDSes for the full-length gene products in accordance with RefSeq's emphasis on representing full-length, functional products.

Alignment of manually curated DIO1 orthologs across phylogenetically distant model organisms showed conservation in selenoprotein length, location of Sec, and sequence, especially around the Sec residue (found at the catalytic site) and the thioredoxin_like super family domain (Figure 2F). The high sequence conservation of selenoproteins is used to predict orthologous models in other vertebrate species by the NCBI's eukaryotic genome annotation pipeline.

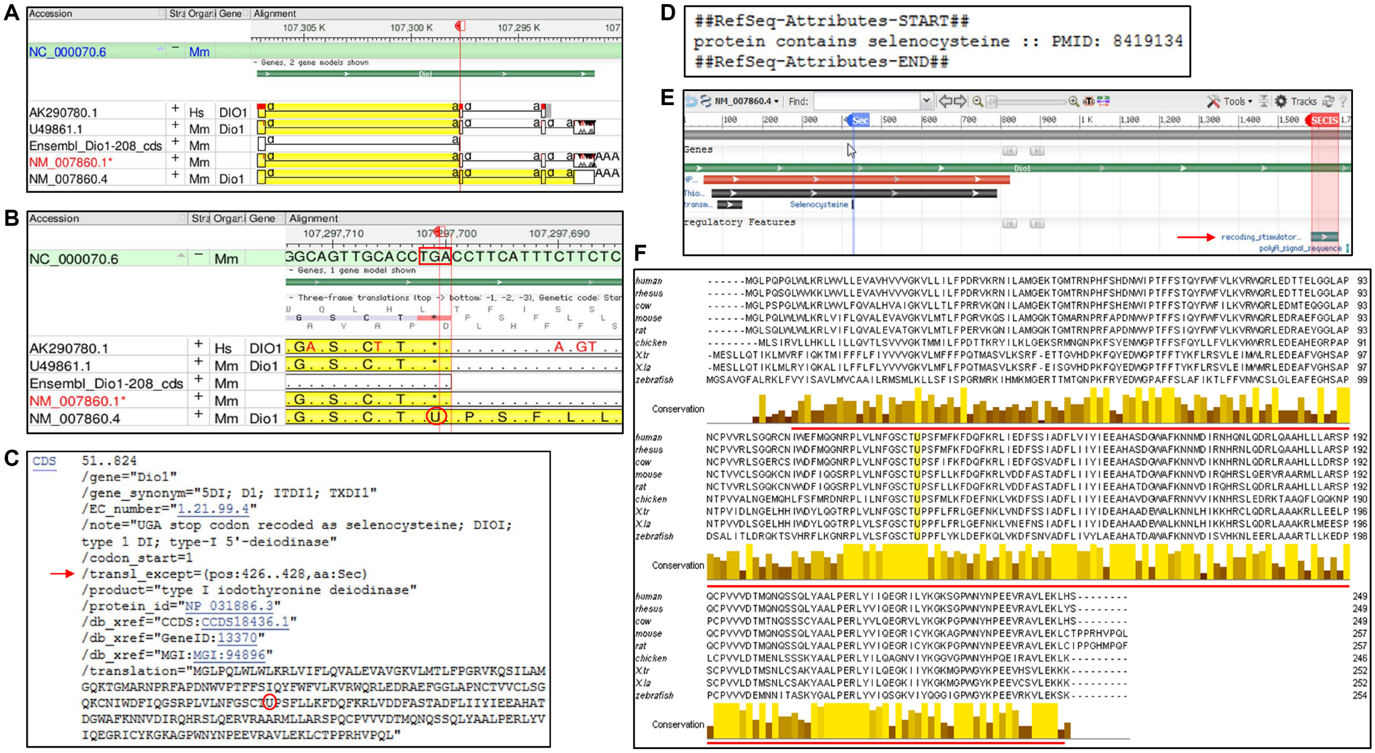
Selenoprotein genes and selenoproteomes of vertebrate model organisms

Curation of 247 selenoprotein genes in nine model organisms resulted in 361 curated RefSeq records, including 115 alternatively spliced transcript variants consisting of 76 protein coding variants with Sec, 10 without Sec, and 29 non-protein coding variants (designated by NR_ accessions). The Sec-containing protein coding variants retained the conserved Sec-UGA codon and SECIS element. The non-Sec containing protein coding variants included those that lacked the exon containing the conserved Sec-UGA codon (e.g. NM_001039715.2 for DIO1) or the SECIS-containing 3′ exon (e.g. NM_203472.2 for SELENOS (41); thus, they were not expected to incorporate Sec. Most of the non-protein coding variants were likely candidates for NMD (e.g. NR_136692.1 for DIO1). The GeneID, RefSeq identifiers, transcript variant and isoform names, and other related information are provided in Supplementary Table 1. Comparative analysis of selenoproteins in nine model organisms is shown in Table 2.

The selenoproteomes of human, rhesus macaque and cow consist of the same 25 selenoproteins. A common core of 21 selenoproteins was also found in the remaining model organisms reviewed here, which all lack GPX6—the only selenoprotein missing in mouse and rat. All non-mammalian vertebrate models also lack SELENOV. Chicken additionally has a non-mammalian paralog (SELENOP2) and selenoprotein (SELENOU). *X. tropicalis* lacks SELENOH, and has a non-mammalian paralog, SELENOW2. *X. laevis* has the same set of 23 selenoproteins as *X. tropicalis*; however, because of its tetraploid genome (44) *X. laevis* has two homeologs (L and S) for 16 and a single homolog for seven selenoproteins. Zebrafish lacks TXNRD1 and has duplicates (a and b) of five selenoproteins, as well as additional non-mammalian paralogs (three singles, one duplicate) and selenoproteins (three singles, one duplicate), some of which were found only in zebrafish. The duplication of several selenoprotein genes in zebrafish was likely the result of whole genome duplication in the early evolution of bony fishes (45).

Our selenoprotein results in five mammalian model species and chicken (more below) are consistent with previous reports (2,3). However, our estimate of the number of selenoproteins in *X. tropicalis* and zebrafish of 23 and 37, respectively, was lesser by one than previously reported (3). We could not find *SELENOH* in *X. tropicalis* nor *X. laevis*. Together, blast and orthology (https://www.ncbi.nlm.nih.gov/gene/?Term=ortholog_gene_280636[group]) results showed the presence of SELENOH in many different vertebrate lineages, including birds, fishes, and two frog species (*Nanorana parkeri* and *Rana catesbeiana*), but not in Xenopus, suggesting that the lack of SELENOH is restricted to a subset of amphibians. SelenoDB 2.0 also does not list SELENOH for *X. tropicalis*. Zebrafish was previously reported to have 38 selenoprotein genes including two GPX3 paralogs, *gpx3a* and *gpx3b* (3); however, we found only a single copy of *gpx3* gene (GeneID:798788) on chromosome 14. SelenoDB 2.0 has entries for both *gpx3a* and *gpx3b* genes. Blast results showed that the *gpx3a* sequence corresponded to *gpx3* gene on chromosome 14; whereas, the *gpx3b* sequence aligned to a locus on chromosome 12, previously annotated as *si:ch73-111m19.2* and very recently renamed *gpx9* (GeneID:794084, ZDB id: ZDB-GENE-130603-13). The full extent of the *gpx3b* transcript and the predicted gpx3b protein was not clear, and the N- and C-termini of gpx3b lacked protein support; therefore, it was not certain if GeneID:794084 on chromosome 12 represents the *gpx3b* paralog or even encodes a selenoprotein.

The human selenoproteome (Figure 3) exemplified some characteristics observed across all model organisms. Selenoproteins varied in length and in the location of their Sec residues. Most selenoproteins contained a single Sec residue; SELENOP was a major exception containing multiple Sec residues (10 in human and between 10–18 in other



Accession	Stra	Organi	Gene
NC_000070.6	–	Mm	
AK290780.1	+	Hs	DIO1
U49861.1	+	Mm	Dio1
Ensembl_Dio1-208_cds	+	Mm	
NM_007860.1*	+	Mm	
NM_007860.4	+	Mm	Dio1

Accession	Stra	Organi	Gene	Alignment
NC_000070.6	–	Mm		GGCAGTTGCACCTGACCTTCATTTCTTCTC
AK290780.1	+	Hs	DIO1	.GA.S..CT.T..*...........A.GT..
U49861.1	+	Mm	Dio1	.G..S..C..T..*..................
Ensembl_Dio1-208_cds	+	Mm		....
NM_007860.1*	+	Mm		.G..S..C..T..*..................
NM_007860.4	+	Mm	Dio1	.G..S..C..T..U..P..S..F..L..L.

```
CDS             51..824
                /gene="Dio1"
                /gene_synonym="5DI; D1; ITDI1; TXDI1"
                /EC_number="1.21.99.4"
                /note="UGA stop codon recoded as selenocysteine; DIOI;
                type 1 DI; type-I 5'-deiodinase"
                /codon_start=1
                /transl_except=(pos:426..428,aa:Sec)
                /product="type I iodothyronine deiodinase"
                /protein_id="NP_031886.3"
                /db_xref="CCDS:CCDS18436.1"
                /db_xref="GeneID:13370"
                /db_xref="MGI:MGI:94896"
                /translation="MGLPQLWLWLKRLVIFLQVALEVAVGKVLMTLFPGRVKQSILAM
                GQKTGMARNPRFAPDNWVPTFFSIQYFWFVLKVRWQRLEDRAEFGGLAPNCTVVCLSG
                QKCNIWDFIQGSRPLVLNFGSCTUPSFLLKFDQFKRLVDDFASTADFLIIYIEEAHAT
                DGWAFKNNVDIRQHRSLQERVRAARMLLARSPQCPVVVDTMQNQSSQLYAALPERLYV
                IQEGRICYKGKAGPWNYNPEEVRAVLEKLCTPPRHVPQL"
```

```
##RefSeq-Attributes-START##
protein contains selenocysteine :: PMID: 8419134
##RefSeq-Attributes-END##
```

```
human      ------MGLPQPGLWLKRLWVLLEVAVHVVVGKVLLILFPDRVKRNILAMGEKTGMTRNPHFSHDNWIPTFFSTQYFWFVLKVRWQRLEDTTELGGLAP 93
rhesus     ------MGLPQSGLWVKKLWVLLEVAVHVVVGKVLLILFPDRVKRNILAMGEKTGMTRNPHFSHDNWIPTFFSTQYFWFVLKVRWQRLEDTTELGGLAP 93
cow        ------MGLPSPGLWLKRLWVLFQVALHVAIGKVLLTLFPRRVKQNILAMGEKTGMTRNPHFSHENWIPTFFSTQYFWFILKVRWQRLEDMTEQGGLAP 93
mouse      ------MGLPQLWLWLKRLVIFLQVALEVAVGKVLMTLFPGRVKQSILAMGQKTGMARNPRFAPDNWVPTFFSIQYFWFVLKVRWQRLEDRAEFGGLAP 93
rat        ------MGLSQLWLWLKRLVIFLQVALEVATGKVLMTLFPERVKQNILAMGQKTGMTRNPRFAPDNWVPTFFSIQYFWFVLKVRWQRLEDRAEYGGLAP 93
chicken    --------MLSIRVLLHKLLILLQVTLSVVVGKTMMILFPDTTKRYILKLGEKSRMNQNPKFSYENWGPTFFSFQYLLFVLKVKWRRLEDEAHEGRPAP 91
X.tr       -MESLLQTIKLMVRFIQKTMIFFFLFIYVVVGKVLMFFFPQTMASVLKSRF-ETTGVHDPKFQYEDWGPTFFTYKFLRSVLEIMWLRLEDEAFVGHSAP 97
X.la       -MESLLQTIKLMLRYIQKALILFFLFLYVVVGKVLMFLFPQTMASVLKSRF-EISGVHDPKFQYEDWGPTFFTYKFLRSVLEIMWMRLEDEAFVGHSAP 97
zebrafish  MGSAVGFALRKLFVYISAVLMVCAAILRMSMLKLLSFISPGRMRKIHMKMGERTTMTQNPKFRYEDWGPAFFSLAFIKTLFFVNWCSLGLEAFEGHSAP 99
Conservation

human      NCPVVRLSGQRCNIWEFMQGNRPLVLNFGSCTUPSFMFKFDQFKRLIEDFSSIADFLVIYIEEAHASDGWAFKNNMDIRNHQNLQDRLQAAHLLLARSP 192
rhesus     NCPVVRLSGQRCNIWDFMQGNRPLVLNFGSCTUPSFMFKFDQFKRLIEDFSSIADFLIIYIEEAHASDGWAFKNNMDIRNHRNLQDRLQAAHLLLARSP 192
cow        NCPVVRLSGERCSIWDFMQGNRPLVLNFGSCTUPSFIFKFDQFKRLIEDFGSVADFLIIYIEEAHASDGWAFKNNVDIKNHRNLQDRLRAAHLLLDRSP 192
mouse      NCTVVCLSGQKCNIWDFIQGSRPLVLNFGSCTUPSFLLKFDQFKRLVDDFASTADFLIIYIEEAHATDGWAFKNNVDIRQHRSLQERVRAARMLLARSP 192
rat        NCTVVRLSGQKCNWDFIQGSRPLVLNFGSCTUPSFLLKFDQFKRLVDDFASTADFLIIYIEEAHATDGWAFKNNVDIRQHRSLQDRLRAAHLLLARSP 192
chicken    NTPVVALNGEMQHLFSFMRDNRPLILNFGSCTUPSFMLKFDEFNKLVKDFSSIADFLIIYIEEAHAVDGWAFRNNVVIKNHRSLEDRKTAAQFLQQKNP 190
X.tr       NTPVIDLNGELHHIWDYLQGTRPLVLNFGSCTUPPFLFRLGEFNKLVNDFSSIADFLIIYIDEAHAADEWALKNNLHIKKHRCLQDRLAAAKRLLEELP 196
X.la       NTPVVDLSGELHHIWDYLQGTRPLVLSFGSCTUPPFLFRLGEFNKLVNEFNSIADFLIIYIDEAHAADEWALKNNLHIKKHRSLQDRLAAAKRLMEESP 196
zebrafish  DSALITLDRQKTSVHRFLKGNRPLVLSFGSCTUPPFLYKLDEFKQLVKDFSNVADFLIVYLAEAHATDAWAFKNNVDISVHKNLEERLAAARTLLKEDP 198
Conservation

human      QCPVVVDTMQNQSSQLYAALPERLYIIQEGRILYKGKSGPWNYNPEEVRAVLEKLHS--------        249
rhesus     QCPVVVDTMQNQSSQLYAALPERLYVIQEGRILYKGKSGPWNYNPEEVRAVLEKLYS--------        249
cow        PCPVVVDTMTNQSSSCYAALPERLYVLQEGRVLYKGKPGPWNYHPEEVRAVLEKLHS--------        249
mouse      QCPVVVDTMQNQSSQLYAALPERLYVIQEGRICYKGKAGPWNYNPEEVRAVLEKLCTPPRHVPQL        257
rat        QCPVVVDTMQNQSSQLYAALPERLYVIQEGRICYKGKPGPWNYNPEEVRAVLEKLCIPPGHMPQF        257
chicken    LCPVVLDTMENLSSSKYAALPERLYILQAGNVIYKGGVGPWNYHPQEIRAVLEKLK---------        246
X.tr       SCPVVLDTMSNLCSAKYAALPERLYILQEGKIIYKGKMGPWGYKPEEVHSVLEKKK---------        252
X.la       SCPVVLDTMSNLCSAKYAALPERLYILQEGKIIYKGKMGPWGYKPEEVCSVLEKKK---------        252
zebrafish  PCPVVVDEMNNITASKYGALPERLYVIQSGKVIYQGGIGPWGYKPEEVRKVLEKSK---------        254
Conservation
```

Figure 2. Manual curation, annotation and analysis of selenoprotein. (**A**) Genome Workbench display of alignment of a few select transcripts to the mouse genomic sequence (chr4, NC_000070.6) at the location of *Dio1*gene (GeneID:13370). The annotated CDS is highlighted in yellow; the red flag marks the location of TGA in exon 2. (**B**) Expansion of the region around the red marker shows translation of TGA (boxed in red) as a stop signal (*) or Sec (U, circled in red) in the -2 frame of the 3-frame translation track. (**C**) GenBank format display of manually added /transl_except qualifier (red arrow) and the incorporated Sec residue (U circled in red) on the public RefSeq record, NM_007860.4. (**D**) Display of a portion of the COMMENT section on NM_007860.4 showing manually added selenoprotein-specific attribute with supporting publication. (**E**) Display of the graphical view of NM_007860.4 configured to add two feature annotations highlighting the location of Sec (nt 426–428) and SECIS element (nt 1582–1651). (**F**) Alignment of type I iodothyronine deiodinase (DIO1) amino acid sequence from nine model organisms. Rows 1–9 represent human (NP_000783.2), rhesus (NP_001116124.1), cow (NP_001116065.1), mouse (NP_031886.3), rat (NP_067685.5), chicken (NP_001091083.1), *X. tropicalis* (NP_001243226.1), *X. laevis* (NP_001089136.1) and zebrafish (NP_001007284.3) RefSeq protein sequences. The position of Sec (U) is highlighted in yellow and the extent of the thioredoxin_like super family domain (aa 8–247, mouse NP_031886.3) is indicated by the red line. The color gradation from yellow to darkest brown in the conservation histogram indicates highest (identical amino acids) to lowest (least conserved) scoring columns. The multiple protein sequence alignment was performed using Clustal Omega, version 1.2.4 (https://www.ebi.ac.uk/Tools/msa/clustalo/) and the alignment was viewed with Jalview, version 2.10.4b1 (http://www.jalview.org). The thioredoxin_like super family domain was identified by searching NCBI's conserved domain database (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi).

model organisms). *SELENOP* was also unique in that most of its mRNA contained two SECIS elements, while other selenoprotein mRNAs contained only one. The use of an alternative polyadenylation (polyA) site located between the two SECIS elements has been reported to result in 10–25% of *SELENOP* mRNA lacking the second SECIS element (46). Both polyA sites were annotated on the human (NM_005410.3) and rhesus macaque (NM_001159490.2) RefSeq transcripts (the upstream polyA site was not found in other model organisms). The same study showed that the two SECIS elements might have distinctive roles in recoding the various UGA codons, which could produce different SELENOP forms in plasma. Other forms of SELENOP have also been reported: under low Se conditions, cysteine was found to be inserted at various UGA sites (47) and C-terminally truncated isoforms were shown to result from termination at a subset of UGA sites (48). Given the complexity of SELENOP regulation, and the lack of information on the full-length nature of many isoforms, we elected to represent only the canonical SELENOP with all UGA-encoded Sec residues. *DIO2* was also of interest because unlike the other two DIO family members (*DIO1* and *DIO3*) its mRNA contained two in-frame UGA codons upstream of an in-frame universal stop codon. It has been reported (in human) that the second UGA codon can function either as a Sec or a stop codon resulting in two isoforms with one or two Sec residues; however, only the upstream Sec (conserved with the single Sec residue found at the active site in DIO1 and DIO3) was essential for enzyme activity (49). Additionally, the observation that the protein extension past the second UGA codon to the downstream stop codon is not conserved (3) suggests that the one-Sec form represents the canonical isoform. Therefore, both the one-Sec (NP_001353425.1) and two-Sec (NP_054644.1) forms were represented for the human *DIO2* gene, and only the canonical one-Sec form for the *DIO2* gene in other model organisms. Feature annotations with relevant comments for the potentially longer isoform were added to the latter RefSeq transcript records at the location of the second UGA and the downstream stop codon (e.g. NM_010050.4).

Table 2. Seleoprotein genes in vertebrate model organisms

							Frog[a]		
Gene	Human	Rhesus macaque	Cow	Mouse	Rat	Chicken	X.tr	X.la	Zebrafish[b]
DIO1	+	+	+	+	+	+	+	+ (L)	+
DIO2	+	+	+	+	+	+	+	+ (L&S)	+
DIO3	+	+	+	+	+	+	+	+ (L&S)	+ (a&b)
GPX1	+	+	+	+	+	+	+	+ (L&S)	+ (a&b)
GPX2	+	+	+	+	+	+	+	+ (L&S)	+
GPX3	+	+	+	+	+	+	+	+ (L&S)	+
GPX4	+	+	+	+	+	+	+	+ (L&S)	+ (a&b)
GPX6	+	+	+	–	–	–	–	–	–
MSRB1	+	+	+	+	+	+	+	+ (L)	+ (a&b)
SELENOF	+	+	+	+	+	+	+	+ (L&S)	+
SELENOH	+	+	+	+	+	+	–	–	+
SELENOI	+	+	+	+	+	+	+	+ (L&S)	+
SELENOK	+	+	+	+	+	+	+	+ (L&S)	+
SELENOM	+	+	+	+	+	+	+	+ (L&S)	+
SELENON	+	+	+	+	+[c]	+	+	+ (L&S)	+
SELENOO	+	+	+	+	+	+	+	+ (L)	+
SELENOP	+	+	+	+	+	+	+	+ (L)	+
SELENOS	+	+	+	+	+	+	+	+ (L&S)	+
SELENOT	+	+	+	+	+	+	+	+ (L&S)	+ (a&b)
SELENOV	+	+	+	+	+	–	–	–	–
SELENOW	+	+	+	+	+	+	+	+ (L)	+
SEPHS2	+	+	+	+	+	+[c]	+	+ (L)	+
TXNRD1	+	+	+	+	+	+	+	+ (L&S)	–
TXNRD2	+	+	+	+	+	+	+	+ (S)	+
TXNRD3	+	+	+	+	+	+	+	+ (L&S)	+
Non-mammalian vertebrate selenoprotein genes									
SELENOE	–	–	–	–	–	–	–	–	+
SELENOJ	–	–	–	–	–	–	–	–	+
SELENOL	–	–	–	–	–	–	–	–	+
SELENOO2	–	–	–	–	–	–	–	–	+
SELENOP2	–	–	–	–	–	–	–	–	+
SELENOT2	–	–	–	–	–	–	–	–	+
SELENOU	–	–	–	–	–	–	–	–	+ (a&b)
SELENOW2	–	–	–	–	–	–	+	+ (L&S)	+ (a&b)
Seleno-proteome	**25**	**25**	**25**	**24**	**24**	**25**	**23**	**23(39)**	**30(37)**

[a] X.tr (*Xenopus tropicalis*); X.la (*Xenopus laevis*); L&S represent homeologs of X.la.
[b] a and b represent paralogs of zebrafish.
[c] rat *Selenon* and chicken *SEPHS2* genes have genome problems (see text).

Several genome and annotation problems were encountered during the curation of this dataset. First, our initial genome blast analysis seemed to indicate that the *SEPHS2* gene was missing in chicken (GRCg6a assembly) and many other avian genomes, corroborated by a similar conclusion in turkey (50). However, *SEPHS2* is an essential gene for Sec synthesis (required for translation of all selenoproteins) whose function cannot be complemented by its paralog, *SEPHS1* (51); therefore, a functional *SEPHS2* gene was expected to be present in chicken and other birds. Tblastn analysis using human SEPHS2 protein (NP_036380.2) revealed the presence of many ESTs in chicken, which allowed the construction of a full-length RefSeq (NM_001366334.2, see Methods for details) for the chicken *SEPHS2* gene (GeneID: 113219448). The encoded selenoprotein of 418 aa (NP_001353263.2) shared ∼70% and ∼80% identity with the mammalian and avian counterparts, respectively. NM_001366334.2 was overall very GC-rich and contained several tandem repeats in the 3′ region, which may contribute to the assembly problem seen with this gene. Similar problems observed in other avian species suggested a peculiarity of the genomic region containing the *SEPHS2* gene that complicates proper genome assembly. However, there were a couple of exceptions: a complete 8-exon *SEPHS2* gene could be identified in the genomes of *Pseudopodoces humilis* (Tibetan ground-tit, GeneID:102113712) and *Parus major* (Great Tit, GeneID:107198884). Curated RefSeqs for the *SEPHS2* gene in both these avian species were created and included in Supplementary Table 1 to facilitate detection of this gene in other birds in the future. Also of note is that the multi-exon *SEPHS2* gene found in non-mammalian vertebrates lacks synteny with the single-exon retrocopy found in placental mammals (3), suggesting that they are not true orthologs, though functionally alike. Second, the *Selenon* gene in rat (GeneID:362624) was found on chromosome 5, but because of gaps in the rat genome assembly (Rnor_6.0) in the 5′ region of this gene, and lack of suitable transcript data, it was not possible to provide a full-length *Selenon* RefSeq. The rat Selenon protein represented in SelenoDB 2.0 is N-terminally partial. Third, the mouse *Smcp* gene (GeneID: 17235) was previously thought to encode a selenoprotein due to the presence of three in-frame UGA codons in the 5′ region (52); however, this was subsequently shown to be incorrect (53).

Gene	Gene ID	Selenoprotein accession	length/Sec location
DIO1	1733	NP_000783.2	249/126
DIO2	1734	NP_054644.1	273/133, 266
DIO3	1735	NP_001353.4	304/170
GPX1	2876	NP_000572.2	203/49
GPX2	2877	NP_002074.2	190/40
GPX3	2878	NP_002075.2	226/73
GPX4	2879	NP_002076.2	197/73
GPX6	257202	NP_874360.1	221/73
MSRB1	51734	NP_057416.1	116/95
SELENOF	9403	NP_004252.2	165/96
SELENOH	280636	NP_734467.1	122/44
SELENOI	85465	NP_277040.1	397/387
SELENOK	58515	NP_067060.2	94/92
SELENOM	140606	NP_536355.1	145/48
SELENON	57190	NP_996809.1	556/428
SELENOO	83642	NP_113642.1	669/667
SELENOP	6414	NP_005401.3	381/59, 300, 318, 330, 345, 352, 367, 369, 376, 378
SELENOS	55829	NP_060915.2	189/188
SELENOT	51714	NP_057359.2	195/49
SELENOV	348303	NP_874363.1	346/273
SELENOW	6415	NP_003000.1	87/13
SEPHS2	22928	NP_036380.2	448/60
TXNRD1	7296	NP_877393.1	499/498
TXNRD2	10587	NP_006431.2	524/523
TXNRD3	114112	NP_443115.1	643/642

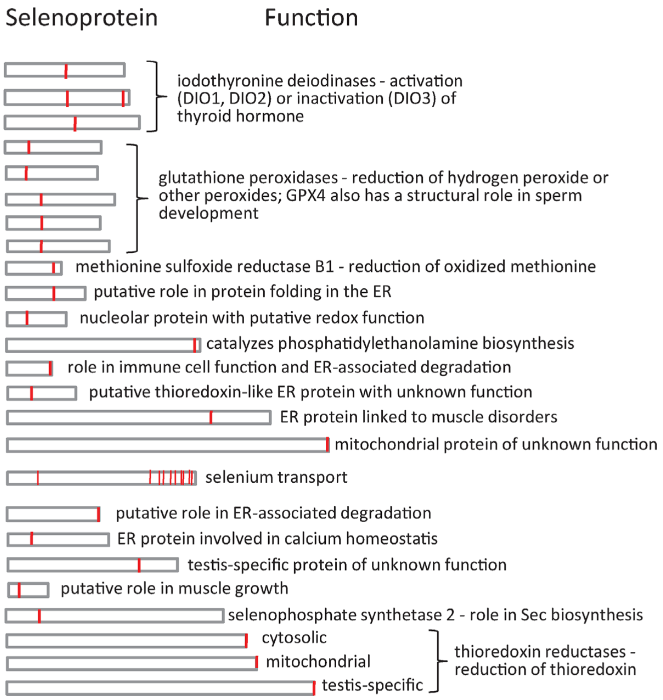

Selenoprotein
Function
iodothyronine deiodinases - activation (DIO1, DIO2) or inactivation (DIO3) of thyroid hormone
glutathione peroxidases - reduction of hydrogen peroxide or other peroxides; GPX4 also has a structural role in sperm development
methionine sulfoxide reductase B1 - reduction of oxidized methionine
putative role in protein folding in the ER
nucleolar protein with putative redox function
catalyzes phosphatidylethanolamine biosynthesis
role in immune cell function and ER-associated degradation
putative thioredoxin-like ER protein with unknown function
ER protein linked to muscle disorders
mitochondrial protein of unknown function
selenium transport
putative role in ER-associated degradation
ER protein involved in calcium homeostatis
testis-specific protein of unknown function
putative role in muscle growth
selenophosphate synthetase 2 - role in Sec biosynthesis
cytosolic
mitochondrial
testis-specific
thioredoxin reductases - reduction of thioredoxin

Figure 3. Human selenoproteome. The table on the left shows details of 25 selenoprotein genes, RefSeq accessions for the encoded selenoproteins, their lengths (in aa) and the locations of Sec residue(s) within. The schematic on the right shows relative lengths of selenoproteins, with the vertical red lines indicating the positions of Sec residue(s). A brief description of selenoprotein function (known or putative) is also shown.

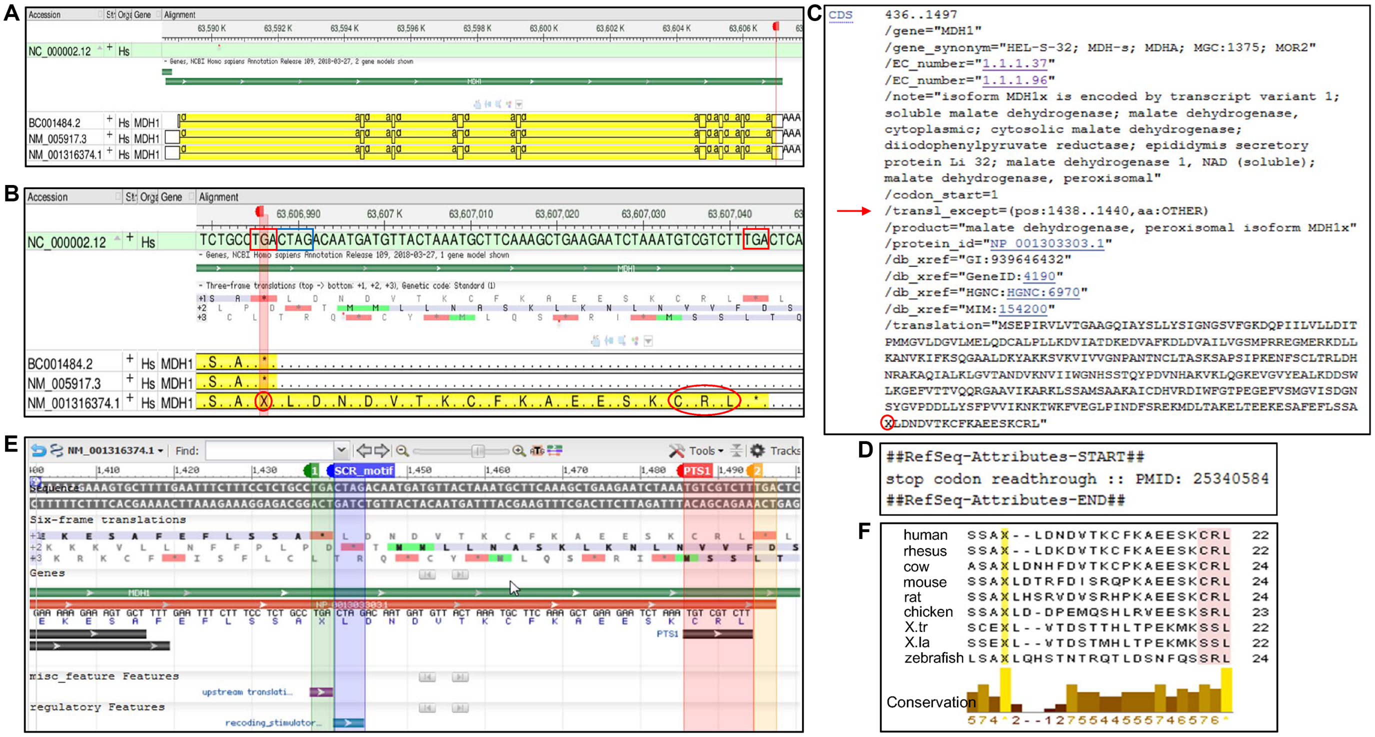
Based on the latter study, the previous mouse *Smcp* RefSeq (NM_008574.3) was updated to NM_008574.4 with a correct CDS encoding an N-terminally shorter protein of 143 aa (NP_032600.3), which is supported by the rat ortholog of similar length (NP_113724.1). Fourth, it was previously reported that fish have only two (*txnrd1* and *txnrd2*) of the three txnrd family members found in mammals (3). We also detected only two *txnrd* genes in zebrafish; however, our analysis showed that they correspond to the mammalian *TXNRD2* and *TXNRD3* genes. Zebrafish *txnrd2* (GeneID:798259) and *txnrd3* (GeneID: 352924) genes are located on chromosomes 5 (3′ neighboring genes, tbx1 and gnb1l) and 6 (3′ neighboring genes, plxna1b, chchd6), respectively, in regions syntenic with their human counterparts. Zebrafish txnrd3 also had a better protein match to human TXNRD3 (GeneID: 114112) than to TXNRD1 (GeneID: 7296). Based on our analysis and consultation with ZFIN, the nomenclature on GeneID:352924 was updated to *txnrd3*.

Curation and annotation of SCR in vertebrates

As with selenoprotein curation, representing recoding of stop codons as sense codons in cases of SCR required manual curation, illustrated by the human *MDH1* gene (Figure 4). INSDC sequence BC001484.2 and NM_005917.3 were annotated with a conventional CDS terminating at the upstream stop codon (TGA) and encoding the canonical isoform of 334 aa (Figure 4A and B). A second RefSeq transcript (NM_001316374.1) identical to NM_005917.3 was created and manually annotated with a CDS terminating at an alternative in-frame stop codon (TGA) 19 codons downstream, resulting in a C-terminally extended isoform of 353 aa (NP_001303303.1). To ensure accurate annotation of the C-terminally extended CDS, a translation exception qualifier was manually added to the CDS as described for selenoprotein curation, except specifying the inserted amino acid as 'OTHER' (Figure 4C). This allowed annotation of the longer CDS, with the inclusion of 'X' in the protein sequence at the location of the upstream stop codon. A 'stop codon readthrough' attribute with published evidence (Figure 4D) and additional feature annotations (Figure 4E) were also manually added to the RefSeq record to highlight the functionally relevant features of this transcript. *MDH1* contained a SCR-stimulatory motif (CTAG) adjacent to the upstream stop codon (TGA) and the encoded protein contained a tripeptide (CRL) at the C-terminus (Figure 4B and E), which has been shown to function as a peroxisomal targeting signal (PTS1) (19). The cryptic PTS1 signal in the C-terminal extension was conserved in mammals and other vertebrate model organisms (Figure 4F).

SCR in vertebrate model organisms

Thirteen genes experimentally verified as authentic SCR cases and the corresponding orthologous model genes with conserved C-terminal extensions were selected for inclusion in the RefSeq database. Review of 93 genes exhibiting SCR in nine model organisms resulted in 94 curated SCR RefSeq records. The rabbit beta-hemoglobin gene (GeneID: 100009084) was also reviewed for historical reasons, being the first mammalian gene experimentally verified to un-



```
CDS             436..1497
                /gene="MDH1"
                /gene_synonym="HEL-S-32; MDH-s; MDHA; MGC:1375; MOR2"
                /EC_number="1.1.1.37"
                /EC_number="1.1.1.96"
                /note="isoform MDH1x is encoded by transcript variant 1;
                soluble malate dehydrogenase; malate dehydrogenase,
                cytoplasmic; cytosolic malate dehydrogenase;
                diiodophenylpyruvate reductase; epididymis secretory
                protein Li 32; malate dehydrogenase 1, NAD (soluble);
                malate dehydrogenase, peroxisomal"
                /codon_start=1
            →   /transl_except=(pos:1438..1440,aa:OTHER)
                /product="malate dehydrogenase, peroxisomal isoform MDH1x"
                /protein_id="NP_001303303.1"
                /db_xref="GI:939646432"
                /db_xref="GeneID:4190"
                /db_xref="HGNC:HGNC:6970"
                /db_xref="MIM:154200"
                /translation="MSEPIRVLVTGAAGQIAYSLLYSIGNGSVFGKDQPIILVLLDIT
                PMMGVLDGVLMELQDCALPLLKDVIATDKEDVAFKDLDVAILVGSMPRREGMERKDLL
                KANVKIFKSQGAALDKYAKKSVKVIVVGNPANTNCLTASKSAPSIPKENFSCLTRLDH
                NRAKAQIALKLGVTANDVKNVIIWGNHSSTQYPDVNHAKVKLQGKEVGVYEALKDDSW
                LKGEFVTTVQQRGAAVIKARKLSSAMSAAKAICDHVRDIWFGTPEGEFVSMGVISDGN
                SYGVPDDLLYSFPVVIKNKTWKFVEGLPINDFSREKMDLTAKELTEEKESAFEFLSSA
                XLDNDVTKCFKAEESKCRL"
```

```
##RefSeq-Attributes-START##
stop codon readthrough :: PMID: 25340584
##RefSeq-Attributes-END##
```

```
human      SSAX--LDNDVTKCFKAEESKCRL   22
rhesus     SSAX--LDKDVTKCFKAEESKCRL   22
cow        ASAXLDNHFDVTKCPKAEESKCRL   24
mouse      SSAXLDTRFDISRQPKAEESKCRL   24
rat        SSAXLHSRVDVSRHPKAEESKCRL   24
chicken    SSAXLD-DPEMQSHLRVEESKSRL   23
X.tr       SCEXL--VTDSTTHLTPEKMKSSL   22
X.la       SSEXL--VTDSTMHLTPEKMKSSL   22
zebrafish  LSAXLQHSTNTRQTLDSNFQSSRL   24
Conservation
           574^2--127554455574657 6^
```

Figure 4. Manual curation, annotation and analysis of SCR. (**A**) Genome Workbench display of alignment of few select transcripts to the human genomic sequence (chr2, NC_000002.12) at the location of the *MDH1*gene (GeneID:4190). The annotated CDS is highlighted in yellow; the red flag marks the location of the upstream stop codon in the 3′ terminal exon. (**B**) Expansion of the region around the red marker between the two in-frame TGA stop codons (boxed in red) shows translation of upstream TGA as a stop signal (*) or an amino acid (X, circled in red) in the +1 frame of the 3-frame translation track. The CTAG SCR_motif (boxed in blue) and the tri-peptide CRL (circled in red) at the C-terminus of NM_001316374.1 are shown. (**C**) GenBank format display of manually added /transl_except qualifier (red arrow) and the TGA-specified amino acid (X circled in red) on the public RefSeq record, NM_001316374.1. (**D**) Display of a portion of the COMMENT section on NM_001316374.1 showing manually added SCR-specific attribute with supporting publication. (**E**) Display of the graphical view of NM_001316374.1 configured to add four feature annotations highlighting the location of the two in-frame stop codons (1 and 2, nt 1438–1440 and 1495–1497, respectively), SCR_motif (nt 1441–1444) and PTS1 (CRL, nt 1486–1494). (**F**) Amino acid (aa) sequence alignment of MDH1 C-terminal extension from nine model organisms. Rows 1–9 represent human (NP_001303303.1, aa 332–353), rhesus (NP_001307034.1, aa 332–353), cow (NP_001307248.1, aa 332–355), mouse (NP_001303604.1, aa 332–355), rat (NP_001303806.1, aa 332–355), chicken (NP_001303820.1, aa 332–354), *X. tropicalis* (NP_001303830.1, 332–353), *X. laevis* (NP_001342383.1, aa 332–353) and zebrafish (NP_001303854.1, aa 332–354). The positions of TGA-specified amino acid (X) and the C-terminal tripeptide are highlighted. The multiple protein sequence alignment was performed using Clustal Omega and viewed with Jalview as described earlier.

dergo SCR (10,11), although this readthrough appears to be a species-specific event not conserved in other mammals. The GeneID, RefSeq identifiers, transcript and isoform names, and other related information are provided in Supplementary Table 2. Only the twin pair of RefSeqs encoding the canonical and C-terminally extended isoform (with or without scr attribute) are shown here. Other alternatively spliced transcript variants that differ only in the N-terminus may also undergo readthrough, resulting in additional C-terminally extended isoforms. However, we represented only one readthrough isoform per gene, unless there was published evidence for more readthrough forms, as described for *AQP4* M1 and M23 isoforms (25). The summary of SCR curation in nine model organisms is shown in Table 3, and the alignment of C-terminal extensions in orthologs of various genes is shown in Supplementary Figure S1.

SCR in many of the genes reviewed here was phylogenetically conserved. In seven (*AQP4*, *MAPK10*, *MDH1*, *MPZ*, *OPRK1*, *OPRL1* and *SACM1L*) out of 13 genes, the C-terminal extension was conserved in vertebrates beyond mammals. For five genes (*ACP2*, *AGO1*, *LDHB*, *MTCH2* and *VEGFA*), readthrough was conserved only in mammals. SCR in the *VDR* gene was represented only in human and rhesus macaque as it appears to be conserved only in the Old World monkey clade, including apes and human (17). Note, the *X. tropicalis* genome assembly (Xtropicalis_v7) has gaps in the 5′ region of the *oprl1* gene (GeneID: 105945889) and in the absence of suitable transcript data, it was not possible to provide a full-length RefSeq for this gene. However, a predicted model (XM_012954875.1) had an intact 3′ coding region and 3′ UTR containing the conserved readthrough extension of 30 aa (see Supplementary Figure S1), so SCR in *OPRL1* gene also seems to be conserved in this frog species. Interestingly, rhesus macaque *VEGFA* gene (GeneID:574209) contains a sequence variant in the downstream stop codon, changing from TGA in other primates to AGA. The AGA codon was confirmed to be valid and homozygous in reads from a DNA sample (SRR447492) from the same individual used for the Mmul_8.0 (GCA_000772875.3) assembly, and in an additional 15 samples of Indian-origin; however, 15 samples from Chinese-origin macaques showed a mix of homozy-

Table 3. Summary of SCR curation in model organisms

Gene	Human	Rhesus macaque	Cow	Mouse	Rat	Chicken	Frog[a]		Zebrafish[b]	
							X.tr	X.la		
					Conservation of C-terminal extension					**Range[c]**
ACP2	+	+	+	+	+	–	–	–	–	11
AGO1	+	+	+	+	+	–	–	–	–	34–38
AQP4	+	+	+	+	+	+	+	+(L&S)	+	24–29
LDHB	+	+	+	+	+	–	–	–	–	7
MAPK10	+	+	+	+	+	+	+	+(L)	+	14
MDH1	+	+	+	+	+	+	+	+(L&S)	+(a&b)	19–21
MPZ	+	+	+	+	+	+	+	+(L&S)	–	64–68
MTCH2	+	+	+	+	+	–	–	–	–	11
OPRK1	+	+	+	+	+	+	+	+(L&S)	+	27–30
OPRL1	+	+	+	+	+	+	+[d]	+(L&S)	+	29–32
SACM1L	+	+	+	+	+	+	+	+(L)	–	22–24
VEGFA	+	–	+	+	+	–	–	–	–	22–23
VDR	+	+	–	–	–	–	–	–	–	67
Total # genes	**13**	**12**	**12**	**12**	**12**	**7**	**7**	**7(12)**	**5(6)**	

[a] X.tr (*Xenopus tropicalis*); X.la (*Xenopus laevis*); L&S represent homeologs of X.la.
[b] a&b represent paralogs of zebrafish.
[c] Range of C-terminal extension (in aa) between orthologs, includes aa specified by the upstream stop codon.
[d] X.tr *oprl1* gene has genome problem (see text).

gous AGA and heterozygous AGA/UGA genotypes (data not shown). The AGA codon would result in a predicted C-terminal extension twice the length found in other mammalian models (44 aa versus 22–23 aa). It is not clear if this sequence change would have an influence on SCR, so we elected to not represent the longer SCR form in rhesus macaque at this time.

The variation in the length of the C-terminal extension between species for a given gene ranged from 0 to 5 aa, mostly attributable to length differences between mammalian and non-mammalian species. Remarkably, the extension in *MAPK10* (14 aa) and *SACM1L* (22 aa) was the same in all model organisms. Conservation of the sequence of the C-terminal extension, however, was quite variable. For instance, the extensions in a few genes (e.g. *ACP2*, *AGO1*, *MAPK10, MTCH2* and *VEGFA*) were well conserved throughout, with the 11 aa of the MTCH2 extension being identical in all mammals. For some genes (e.g. *OPRK1*, *OPRL1* and *SACM1L*), strongest conservation was seen near the C-terminus; and for *LDHB* and *MDH1* (Figure 4E), it was mostly the conservation of the C-terminal tripeptide that functions as a peroxisomal targeting signal (PTS1). Thus, sequence conservation of the entire extension may not be important, depending on the role of the C-terminal extension.

The details of SCR curation in human are shown in Table 4 and exemplify some general observations. The upstream (readthrough) stop codon (uStop) in 12 out of 13 genes represented here is UGA, the leakiest of the three stop codons. *MPZ* is the only gene in this set that has UAG as uStop. The uStop is conserved across all model organisms for all genes. The downstream stop codon (dStop) was observed to be a mix of all three stop codons, and UGA was most prevalent even as a dStop, followed by UAG. The various signals that stimulate SCR in different genes are shown. There are a couple of genes (*AGO1* and *MTCH2)* for which readthrough has been observed (15), but no stimulatory signal was identified. On the other hand, *ACP2* and *SACM1L* have highly conserved stem-loop structures (thought to stimulate readthrough) and C-terminal extensions, but efficient readthrough has been difficult to measure in these genes (23). Genes with a UGACUAG motif have been best characterized (16–19) and constitute the majority of genes (7 out of 13) in this sampling. For *MPZ* with UAG as uStop, a quite different conserved six base pair sequence surrounding the UAG stop codon (AAA_UAG_CGG) was shown to be essential for readthrough (21). For *VEGFA,* a unique cis-acting element (Ax element) that binds a hnRNP A2/B1 trans-acting factor was shown to promote readthrough (15). The variety of stimulatory signals suggest the existence of different mechanisms for SCR in eukaryotes.

DISCUSSION

The dual role of AUG in initiating translation and encoding an internal methionine was realized soon after the deciphering of the genetic code. However, the dual/multiple function of stop codon came to be known later with the discovery of UGA recoding to specify Sec (resulting in the expansion of the genetic code to include Sec as the 21st amino acid) and the recoding of all three stop codons as sense amino acids via SCR. The ambiguous nature of stop codons poses a problem for conventional computational tools to accurately predict selenoprotein genes or detect SCR, resulting in misannotation of selenoprotein CDSes on transcripts and genomic sequences, and failure to find novel readthrough isoforms. Several computational methods have been developed to detect selenoprotein genes: for example, one which combines the prediction of eukaryotic SECIS element using SECISearch3 with another tool (Seblastian) that predicts selenoprotein sequences encoded upstream of SECIS elements (32); and a homology-based tool (Selenoprofiles) designed to identify members of known selenoprotein families in sequenced genomes (54). However, their use,

Table 4. SCR curation in human

					Protein isoform					
		Stop codons[a]			Canonical		Readthrough			
Gene	Gene ID	uStop	dStop	SCR signal	Accession	Length[b]	Accession	Length[b]	C-terminal extension[b]	Expt evidence[c]
ACP2	53	UGA	UAG	2° structure	NP_001601.1	423	NP_001343945.1	434	11	23
AGO1	26523	UGA	UAG	ND	NP_036331.1	857	NP_001304051.1	891	34	15
AQP4	361	UGA	UAA	UGACUAG	NP_001641.1	323	NP_001304313.1	352	29	16
LDHB	3945	UGA	UAG	UGACUAG	NP_002291.1	334	NP_001302466.1	341	7	18, 19
MAPK10	5602	UGA	UGA	UGACUAG	NP_620448.1	464	NP_001304998.1	478	14	16
MDH1	4190	UGA	UGA	UGACUAG	NP_005908.1	334	NP_001303303.1	353	19	19
MPZ	4359	UAG	UGA	AAAUAGCGG	NP_000521.2	248	NP_001302420.1	312	64	12
MTCH2	23788	UGA	UAG	ND	NP_055157.1	303	NP_001304160.1	314	11	15
OPRK1	4986	UGA	UGA	UGACUAG	NP_000903.2	380	NP_001305426.1	409	29	16
OPRL1	4987	UGA	UAG	UGACUAG	NP_872588.1	370	NP_001305782.1	399	29	16
SACM1L	22908	UGA	UGA	2° structure	NP_054735.3	587	NP_001306000.1	609	22	23
VEGFA	7422	UGA	UGA	Ax element	NP_001165097.1	191	NP_001303939.1	213	22	15
VDR	7421	UGA	UGA	UGACUAG	NP_000367.1	427	NP_001351014.1	494	67	17

[a] Upstream (uStop) and downstream (dStop) stop codons, ND, not determined.
[b] The numerical values indicate number of aa residues.
[c] Reference for experimental evidence of readthrough.

in conjunction with standard genome annotation pipelines, is not widespread. Our dataset of manually curated and annotated full-length selenoprotein transcript and protein products can serve as a valuable resource for individual research projects, large scale computational analyses, and genome annotation. It expands on the vertebrate selenoprotein data tracked by other databases, such as SelenoDB 2.0, and repositories of translational recoding events, such as Recode-2. Similarly, no single tool is available that can effectively predict SCR in genomes, so our curated set of RefSeqs representing readthrough in several model organisms can be a useful resource for SCR studies, and to populate Recode-2 database, which currently has no entry for readthrough in mammalian cellular genes.

The manually curated selenoprotein dataset described here is unique in its representation of full-length selenoprotein transcript and protein sequences, additional alternatively spliced transcript variants, and expanded functional annotation including gene summary, relevant publications, biological attributes, and feature annotations. The inclusion of nine model organisms in this curation effort allowed comparison of selenoproteomes across phylogenetically distant vertebrate species, and with the existing data on various selenoproteomes in publications and specialized databases, such as SelenoDB 2.0 and Recode-2. For instance, our analysis revealed the restricted loss of *selenoh* gene in Xenopus. The *X. laevis* selenoproteome reported here is novel. It has the same complement of selenoprotein genes as *X. tropicalis*; however, because of its tetraploid genome, *X. laevis* has additional selenoproteins derived from homeologous copies of some selenoprotein genes, making its selenoproteome larger (with 39 selenoproteins), comparable to that of zebrafish (with 37 selenoproteins). Manual curation was essential for correcting errors in the representation of selenoprotein transcripts, CDSes and proteins. Errors resulted from misannotation of UGA as a stop codon on primary transcripts or predicted models, or the opposite, i.e. misinterpretation of UGA codons as Sec codons, exemplified by the mouse *Smcp* gene mistaken to encode a selenoprotein. Manual curation also included standardizing nomenclature of selenoprotein genes across all model organisms. We encountered problems in representing full-length RefSeqs for rat *Selenon* and *X. tropicalis oprl1* genes because of genome issues, which also affected correct model prediction in these regions. Besides the challenge of accurately predicting the dual meaning of stop codons, poor genome quality adds to the problem of correctly annotating full-length gene products by conventional annotation pipelines. This is no better illustrated than the curious problem we faced with the identification of *SEPHS2* gene in chicken and other avian species. At first glance it appeared as if this very essential gene for Sec synthesis was missing in birds because it could not be detected in many of their genomes. However, the identification of ESTs for *SEPHS2* gene in chicken and other avian species confirmed that this gene was indeed present in birds. It is thus important to realize the limitations of automated genome annotation of not only recoded genes, but other genes as well, and the impact of such limitations on evolutionary analyses. Therefore, manually curated data are an important complement to computational genome annotation by NCBI's eukaryotic genome annotation pipeline.

Different mechanisms exist for proteins to acquire new domains or functional modules, such as alternative splicing, leaky AUG scanning, and SCR. SCR, which operates at the level of translation termination to generate diversity at the C-terminus, is complementary to leaky AUG scanning, a means for protein diversification at the N-terminus via alternative translation initiation. The different N- or C-termini can confer different subcellular localization or function; for example NP_001173.2 and NP_001188306.1 for human *ALDH7A1* gene (GeneID:501) generated by alternative AUG usage are localized in the mitochondria and cytosol, respectively. Similarly, pairs of LDHB and MDH1 isoforms generated by alternative stop codon usage are localized in the cytosol and peroxisomes. The exact function of many C-terminally extended isoforms is not known; however, the conservation of C-terminal extension in vertebrates in many

SCR cases described here suggests that readthrough is not simply due to chance, or an accidental misreading of the stop codon, and likely serves a function. Some readthrough isoforms may have normal physiological roles, for instance, peroxisomal LDHx in lactate/pyruvate shuttling and glyoxylate metabolism (18), while some may have pathological roles, such as L-MPZ in stimulating the production of anti L-MPZ antibodies in patients with neuropathies (12). Inducing stop codon readthrough with aminoglycoside antibiotics has gained prominence for treatment of inherited diseases, such as cystic fibrosis, Duchenne muscular dystrophy, and some cancers, caused by genes containing premature termination codon (PTC) mutations (55). Studies in animal models and cultured cells have shown that aminoglycosides can promote PTC readthrough to partially restore full-length protein synthesis and function. However, as for SCR, the efficiency of aminoglycoside-induced PTC readthrough is context dependent, requiring identification of pathogenic mutation context in each patient for effective treatment. While PTC readthrough mechanisms may not be the same as those for SCR, insights gained from SCR studies on factors influencing efficient readthrough, choice of amino acid inserted at the readthrough site, and the development of reliable assays to measure readthrough efficiency, will be valuable in designing therapies for treatment of PTC disorders.

The RefSeq database is unique, not only in its offering of annotated reference sequence dataset of transcripts, proteins and genomes across a broad taxonomic scope, but also in its active maintenance and provision of updates for a diverse set of species in a timely manner to incorporate new data types and current knowledge (26). The high quality of RefSeq products is achieved through a combinatorial approach leveraging computation, manual curation and collaboration, with a focus on representing only full-length and non-redundant data. Manual curation has a vital role in the annotation process in improving the quality of data, which in turn adds to the quality of genome annotation. This is particularly true for genes with exceptional biology that require manual curation for accurate annotation of their gene products. RefSeq is also notable as one of the major reference sequence databases to represent various recoding events, such as +1 ribosomal frameshifting in antizyme genes (27), –1 ribosomal frameshifting in mammalian *PEG10* gene (e.g. NM_015068.3: NP_055883.2), and stop codon recoding (Sec insertion and SCR, this report). The RefSeq dataset is widely used, including by the recoding community (3,16,19,21,50), for a variety of research purposes. To cope with the growing number of vertebrate sequenced genomes, our goal going forward is to consider automatic annotation of selenoprotein genes. Recent improvements in NCBI's eukaryotic genome annotation pipeline has enabled high quality selenoprotein model predictions, including automatic addition of /transl_except qualifier and annotation of Sec as 'U' (e.g. XM_009293291.2 for zebrafish *gpx2*). A similar approach was recently incorporated into the RefSeq prokaryotic genome annotation pipeline and used to successfully annotate several selenoproteins across many bacterial species (56). Manual curation, however, will continue to be needed to provide high quality annotations of important gene sets based on well-supported transcripts for key eukaryotic genomes, and to support improvements to the genome annotation pipeline to model both typical and atypical biology.

DATA AVAILABILITY

The RefSeq data presented here are publicly available from several resources at NCBI (https://www.ncbi.nlm.nih.gov). Information about selenoprotein and SCR genes, transcripts and proteins can be accessed interactively from the Gene, Nucleotide and Protein databases, respectively; from reciprocal links on the Gene and RefSeq records, and from BLAST databases. These data are also included in the bi-monthly comprehensive RefSeq release from: ftp://ftp.ncbi.nlm.nih.gov/refseq/release/. Shown below are examples of web queries relevant to recoding datasets, and useful for retrieving subsets of related records:

- Query the Gene database with a gene symbol, for example, dio1[sym], to retrieve dio1 orthologs, which are displayed in a tabular format, and individual Gene records in the set can be retrieved via the link on the Name/GeneID column.
- Query the Nucleotide or Protein database with a specific attribute, for example, 'protein contains selenocysteine'[prop] or 'stop codon readthrough'[prop] to retrieve all RefSeq records annotated with the respective attribute, not just the dataset represented here. For example, the 'protein contains selenocysteine'[prop] search currently retrieves 552 RefSeq records for 472 selenoprotein genes from 37 taxa. The sidebar on the right shows per-taxon counts, which can be navigated to display taxon-specific data. Using 'Find related data' and selecting Gene database will show the number of selenoprotein genes. The datasets retrieved by both web queries can be directly downloaded using the available 'Send to' options.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to acknowledge RefSeq curators Catherine Farrell and David Webb, and developers Vamsi Kodali and Alexander Souvorov for helpful consults. We would also like to thank our NAR referees for their valuable insights and observations, which helped with annotation and improving this manuscript.

FUNDING

Intramural Research Program of the National Institutes of Health, National Library of Medicine. Funding for open access charge: Intramural Research Program of the National Institutes of Health, National Library of Medicine.
Conflict of interest statement. None declared.

REFERENCES

1. Hatfield,D.L., Tsuji,P.A., Carlson,B.A. and Gladyshev,V.N. (2014) Selenium and selenocysteine: Roles in cancer, health, and development. *Trends Biochem. Sci.*, **39**, 112–120.
2. Kryukov,G.V., Castellano,S., Novoselov,S.V., Lobanov,A.V., Zehtab,O., Guigo,R. and Gladyshev,V.N. (2003) Characterization of mammalian selenoproteomes. *Science*, **300**, 1439–1443.
3. Mariotti,M., Ridge,P.G., Zhang,Y., Lobanov,A.V., Pringle,T.H., Guigo,R., Hatfield,D.L. and Gladyshev,V.N. (2012) Composition and evolution of the vertebrate and mammalian selenoproteomes. *PLoS One*, **7**, e33066.
4. Moghadaszadeh,B. and Beggs,A.H. (2006) Selenoproteins and their impact on human health through diverse physiological pathways. *Physiology (Bethesda)*., **21**, 307–315.
5. Pitts,M.W. and Hoffmann,P.R. (2018) Endoplasmic reticulum-resident selenoproteins as regulators of calcium signaling and homeostasis. *Cell Calc.*, **70**, 76–86.
6. Ursini,F., Heim,S., Kiess,M., Maiorino,M., Roveri,A., Wissing,J. and Flohe,L. (1999) Dual function of the selenoprotein PHGPx during sperm maturation. *Science*, **285**, 1393–1396.
7. Schomburg,L., Schweizer,U., Holtmann,B., Flohe,L., Sendtner,M. and Kohrle,J. (2003) Gene disruption discloses role of selenoprotein P in selenium delivery to target tissues. *Biochem. J.*, **370**, 397–402.
8. Castets,P., Lescure,A., Guicheney,P. and Allamand,V. (2012) Selenoprotein N in skeletal muscle: from diseases to function. *J. Mol. Med.*, **90**, 1095–1107.
9. Firth,A.E. and Brierley,I. (2012) Non-canonical translation in RNA viruses. *J. Gen. Virol.*, **93**, 1385–1409.
10. Chittum,H.S., Lane,W.S., Carlson,B.A., Roller,P.P., Lung,F.D., Lee,B.J. and Hatfield,D.L. (1998) Rabbit beta-globin is extended beyond its UGA stop codon by multiple suppressions and translational reading gaps. *Biochemistry*, **37**, 10866–10870.
11. Geller,A.I. and Rich,A. (1980) A UGA termination suppression tRNATrp active in rabbit reticulocytes. *Nature*, **283**, 41–46.
12. Yamaguchi,Y., Hayashi,A., Campagnoni,C.W., Kimura,A., Inuzuka,T. and Baba,H. (2012) L-MPZ, a novel isoform of myelin P0, is produced by stop codon readthrough. *J. Biol. Chem.*, **287**, 17765–17776.
13. Jungreis,I., Lin,M.F., Spokony,R., Chan,C.S., Negre,N., Victorsen,A., White,K.P. and Kellis,M. (2011) Evidence of abundant stop codon readthrough in Drosophila and other metazoa. *Genome Res.*, **21**, 2096–2113.
14. Dunn,J.G., Foo,C.K., Belletier,N.G., Gavis,E.R. and Weissman,J.S. (2013) Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. *eLife*, **2**, e01179.
15. Eswarappa,S.M., Potdar,A.A., Koch,W.J., Fan,Y., Vasu,K., Lindner,D., Willard,B., Graham,L.M., DiCorleto,P.E. and Fox,P.L. (2014) Programmed translational readthrough generates antiangiogenic VEGF-Ax. *Cell*, **157**, 1605–1618.
16. Loughran,G., Chou,M.Y., Ivanov,I.P., Jungreis,I., Kellis,M., Kiran,A.M., Baranov,P.V. and Atkins,J.F. (2014) Evidence of efficient stop codon readthrough in four mammalian genes. *Nucleic Acids Res.*, **42**, 8928–8938.
17. Loughran,G., Jungreis,I., Tzani,I., Power,M., Dmitriev,R.I., Ivanov,I.P., Kellis,M. and Atkins,J.F. (2018) Stop codon readthrough generates a C-terminally extended variant of the human vitamin D receptor with reduced calcitriol response. *J. Biol. Chem.*, **293**, 4434–4444.
18. Schueren,F., Lingner,T., George,R., Hofhuis,J., Dickel,C., Gartner,J. and Thoms,S. (2014) Peroxisomal lactate dehydrogenase is generated by translational readthrough in mammals. *eLife*, **3**, e03640.
19. Stiebler,A.C., Freitag,J., Schink,K.O., Stehlik,T., Tillmann,B.A., Ast,J. and Bolker,M. (2014) Ribosomal readthrough at a short UGA stop codon context triggers dual localization of metabolic enzymes in Fungi and animals. *PLos Genet.*, **10**, e1004685.
20. Beier,H. and Grimm,M. (2001) Misreading of termination codons in eukaryotes by natural nonsense suppressor tRNAs. *Nucleic Acids Res.*, **29**, 4767–4782.
21. Yamaguchi,Y. and Baba,H. (2018) Phylogenetically conserved sequences around myelin P0 stop codon are essential for translational readthrough to produce L-MPZ. *Neurochem. Res.*, **43**, 227–237.
22. Firth,A.E., Wills,N.M., Gesteland,R.F. and Atkins,J.F. (2011) Stimulation of stop codon readthrough: frequent presence of an extended 3′ RNA structural element. *Nucleic Acids Res.*, **39**, 6679–6691.
23. Loughran,G., Howard,M.T., Firth,A.E. and Atkins,J.F. (2017) Avoidance of reporter assay distortions from fused dual reporters. *RNA*, **23**, 1285–1289.
24. Xin,H., Zhong,C., Nudleman,E. and Ferrara,N. (2016) Evidence for Pro-angiogenic functions of VEGF-Ax. *Cell*, **167**, 275–284.
25. De Bellis,M., Pisani,F., Mola,M.G., Rosito,S., Simone,L., Buccoliero,C., Trojano,M., Nicchia,G.P., Svelto,M. and Frigeri,A. (2017) Translational readthrough generates new astrocyte AQP4 isoforms that modulate supramolecular clustering, glial endfeet localization, and water transport. *Glia*, **65**, 790–803.
26. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
27. Rajput,B., Murphy,T.D. and Pruitt,K.D. (2015) RefSeq curation and annotation of antizyme and antizyme inhibitor genes in vertebrates. *Nucleic Acids Res.*, **43**, 7270–7279.
28. Karsch-Mizrachi,I., Takagi,T. and Cochrane,G. (2018) The international nucleotide sequence database collaboration. *Nucleic Acids Res.*, **46**, D48–D51.
29. Kervestin,S. and Jacobson,A. (2012) NMD: A multifaceted response to premature translational termination. *Nat. Rev. Mol. Cell Biol.*, **13**, 700–712.
30. Romagne,F., Santesmasses,D., White,L., Sarangi,G.K., Mariotti,M., Hubler,R., Weihmann,A., Parra,G., Gladyshev,V.N., Guigo,R. *et al.* (2014) SelenoDB 2.0: annotation of selenoprotein genes in animals and their genetic diversity in humans. *Nucleic Acids Res.*, **42**, D437–D443.
31. Bekaert,M., Firth,A.E., Zhang,Y., Gladyshev,V.N., Atkins,J.F. and Baranov,P.V. (2010) Recode-2: New design, new search tools, and many more genes. *Nucleic Acids Res.*, **38**, D69–D74.
32. Mariotti,M., Lobanov,A.V., Guigo,R. and Gladyshev,V.N. (2013) SECISearch3 and Seblastian: New tools for prediction of SECIS elements and selenoproteins. *Nucleic Acids Res.*, **41**, e149.
33. Gladyshev,V.N., Arner,E.S., Berry,M.J., Brigelius-Flohe,R., Bruford,E.A., Burk,R.F., Carlson,B.A., Castellano,S., Chavatte,L., Conrad,M. *et al.* (2016) Selenoprotein gene nomenclature. *J. Biol. Chem.*, **291**, 24036–24040.
34. Burt,D.W., Carre,W., Fell,M., Law,A.S., Antin,P.B., Maglott,D.R., Weber,J.A., Schmidt,C.J., Burgess,S.C. and McCarthy,F.M. (2009) The chicken gene nomenclature committee report. *BMC Genomics*, **10**, S5.
35. Eppig,J.T. (2017) Mouse genome informatics (MGI) Resource: Genetic, genomic, and biological knowledgebase for the laboratory mouse. *ILAR J.*, **58**, 17–41.
36. Karimi,K., Fortriede,J.D., Lotay,V.S., Burns,K.A., Wang,D.Z., Fisher,M.E., Pells,T.J., James-Zorn,C., Wang,Y., Ponferrada,V.G. *et al.* (2018) Xenbase: A genomic, epigenomic and transcriptomic model organism database. *Nucleic Acids Res.*, **46**, D861–D868.
37. Ruzicka,L., Bradford,Y.M., Frazer,K., Howe,D.G., Paddock,H., Ramachandran,S., Singer,A., Toro,S., Van Slyke,C.E., Eagle,A.E. *et al.* (2015) ZFIN, The zebrafish model organism database: Updates and new directions. *Genesis*, **53**, 498–509.
38. Shimoyama,M., De Pons,J., Hayman,G.T., Laulederkind,S.J., Liu,W., Nigam,R., Petri,V., Smith,J.R., Tutaj,M., Wang,S.J. *et al.* (2015) The rat genome database 2015: Genomic, phenotypic and environmental variations and disease. *Nucleic Acids Res.*, **43**, D743–D750.
39. Yates,B., Braschi,B., Gray,K.A., Seal,R.L., Tweedie,S. and Bruford,E.A. (2017) Genenames.org: The HGNC and VGNC resources in 2017. *Nucleic Acids Res.*, **45**, D619–D625.
40. Howard,M.T., Aggarwal,G., Anderson,C.B., Khatri,S., Flanigan,K.M. and Atkins,J.F. (2005) Recoding elements located adjacent to a subset of eukaryal selenocysteine-specifying UGA codons. *EMBO J.*, **24**, 1596–1607.
41. Bubenik,J.L., Miniard,A.C. and Driscoll,D.M. (2013) Alternative transcripts and 3′UTR elements govern the incorporation of selenocysteine into selenoprotein S. *PLoS One*, **8**, e62102.
42. Hatfield,D.L. and Gladyshev,V.N. (2002) How selenium has altered our understanding of the genetic code. *Mol. Cell. Biol.*, **22**, 3565–3576.

43. Howard,M.T., Carlson,B.A., Anderson,C.B. and Hatfield,D.L. (2013) Translational redefinition of UGA codons is regulated by selenium availability. *J. Biol. Chem.*, **288**, 19401–19413.
44. Session,A.M., Uno,Y., Kwon,T., Chapman,J.A., Toyoda,A., Takahashi,S., Fukui,A., Hikosaka,A., Suzuki,A., Kondo,M. *et al.* (2016) Genome evolution in the allotetraploid frog Xenopus laevis. *Nature*, **538**, 336–343.
45. Taylor,J.S., Braasch,I., Frickey,T., Meyer,A. and Van de Peer,Y. (2003) Genome duplication, a trait shared by 22000 species of ray-finned fish. *Genome Res.*, **13**, 382–390.
46. Wu,S., Mariotti,M., Santesmasses,D., Hill,K.E., Baclaocos,J., Aparicio-Prat,E., Li,S., Mackrill,J., Wu,Y., Howard,M.T. *et al.* (2016) Human selenoprotein P and S variant mRNAs with different numbers of SECIS elements and inferences from mutant mice of the roles of multiple SECIS elements. *Open Biol.*, **6**, 1–16.
47. Turanov,A.A., Everley,R.A., Hybsier,S., Renko,K., Schomburg,L., Gygi,S.P., Hatfield,D.L. and Gladyshev,V.N. (2015) Regulation of selenocysteine content of human selenoprotein P by dietary selenium and insertion of cysteine in place of selenocysteine. *PLoS One*, **10**, e0140353.
48. Ma,S., Hill,K.E., Caprioli,R.M. and Burk,R.F. (2002) Mass spectrometric characterization of full-length rat selenoprotein P and three isoforms shortened at the C terminus. Evidence that three UGA codons in the mRNA open reading frame have alternative functions of specifying selenocysteine insertion or translation termination. *J. Biol. Chem.*, **277**, 12749–12754.
49. Salvatore,D., Harney,J.W. and Larsen,P.R. (1999) Mutation of the Secys residue 266 in human type 2 selenodeiodinase alters 75Se incorporation without affecting its biochemical properties. *Biochimie*, **81**, 535–538.
50. Sunde,R.A., Sunde,G.R., Sunde,C.M., Sunde,M.L. and Evenson,J.K. (2015) Cloning, sequencing, and expression of selenoprotein transcripts in the turkey (Meleagris gallopavo). *PLoS One*, **10**, e0129801.
51. Xu,X.M., Carlson,B.A., Irons,R., Mix,H., Zhong,N., Gladyshev,V.N. and Hatfield,D.L. (2007) Selenophosphate synthetase 2 is essential for selenoprotein biosynthesis. *Biochem. J.*, **404**, 115–120.
52. Karimpour,I., Cutler,M., Shih,D., Smith,J. and Kleene,K.C. (1992) Sequence of the gene encoding the mitochondrial capsule selenoprotein of mouse sperm: Identification of three in-phase TGA selenocysteine codons. *DNA Cell Biol.*, **11**, 693–699.
53. Cataldo,L., Baig,K., Oko,R., Mastrangelo,M.A. and Kleene,K.C. (1996) Developmental expression, intracellular localization, and selenium content of the cysteine-rich protein associated with the mitochondrial capsules of mouse sperm. *Mol. Reprod. Dev.*, **45**, 320–331.
54. Mariotti,M. and Guigo,R. (2010) Selenoprofiles: profile-based scanning of eukaryotic genome sequences for selenoprotein genes. *Bioinformatics*, **26**, 2656–2663.
55. Bidou,L., Allamand,V., Rousset,J.P. and Namy,O. (2012) Sense from nonsense: therapies for premature stop codon diseases. *Trends Mol. Med.*, **18**, 679–688.
56. Haft,D.H., DiCuccio,M., Badretdin,A., Brover,V., Chetvernin,V., O'Neill,K., Li,W., Chitsaz,F., Derbyshire,M.K., Gonzales,N.R. *et al.* (2018) RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res.*, **46**, D851–D860.