MANUSCRIPT TITLE

Tracing the vertebrate selenoproteome reveals expansions in ray-finned fishes and convergent depletions in tetrapods

AUTHORS

Max Ticó[1], Marco Mariotti[1*]

[1] Department of Genetics, Microbiology and Statistics, Universitat de Barcelona, Barcelona, Catalonia, 08024, Spain

* To whom correspondence should be addressed. Email: marco.mariotti@ub.edu

ABSTRACT

Selenoproteins incorporate the rare selenium-containing amino acid selenocysteine (Sec) and play crucial roles for redox homeostasis, stress response, and hormone regulation. Sec is inserted by co-translational recoding of the UGA codon, normally a stop. As consequence, selenoproteins are often misannotated in public databases and require specialized bioinformatic methods and resources.

Here, we present a refined characterization of the composition and evolution of the vertebrate selenoproteome. Based on analyses of 19 gene families across hundreds of genomes, we show that extant selenoproteomes were shaped by extensive gene duplications (53), losses (41), and Sec-to-cysteine (Cys) conversions (19). Tetrapods including mammals encode 24-25 selenoproteins, with variations in 6 families. Notably, the same genes underwent convergent evolutionary events in multiple tetrapods, namely Sec-to-Cys substitutions (SELENOU1, GPX6) and gene losses (SELENOV). In contrast, ray-finned fish exhibit larger and more dynamic selenoproteomes, reinforcing the hypothesis that the selective advantage of Sec is stronger in aquatic environments. We detected selenoprotein duplications spread across the actinopterygian clade involving 13 families, mainly involved in antioxidant defense. The richest

selenoproteomes were found in Salmoninae and Cyprinoidei fish with 55 and 45 selenoproteins, respectively, owing to whole genome duplications. Among our findings, the SELENOP family stands out in lampreys, carrying up to an unprecedented 162 UGAs putatively recoded to Sec.

Our study presents the most comprehensive evolutionary map of vertebrate selenoproteins to date and delineates the specific selenoproteome of each lineage, establishing a foundational framework for selenium biology research in the era of biodiversity genomics.

INTRODUCTION

Selenium is an essential trace element for human health. The biological effects of selenium are mostly mediated by its incorporation in the non-canonical amino acid selenocysteine (Sec) (Fairweather-Tait et al. 2011; Labunskyy et al. 2014; Mariotti and Gladyshev 2022). Sec is the essential component of selenoproteins, a unique class of proteins with crucial functions in both normal physiology and cancer. Sec is often referred to as the 21st amino acid due to its distinctive biosynthetic pathway and incorporation mechanism, which differ fundamentally from those of the 20 canonical amino acids. Importantly, Sec is inserted in correspondence to the UGA codon, which is normally a stop codon. The UGA is co-translationally recoded to Sec in selenoprotein mRNAs in response to local sequence elements, the major of which is called SECIS element mechanism (Howard and Copeland 2019; Copeland and Howard 2021). Moreover, unlike standard amino acids which are synthesized then charged onto their cognate transfer RNAs (tRNAs), Sec is synthesized directly on its own dedicated tRNA, tRNASec, through a multi-step enzymatic process. The selenium required for this reaction is provided in the form of selenophosphate, synthesized by protein SEPHS2, an

essential enzyme in Sec metabolism which, intriguingly, is itself a selenoprotein (Manta et al. 2022).

Since their discovery in the 1970s, Sec and selenoproteins have been extensively studied across diverse fields of biology (Flohé 2009; Davy and Castellano 2018). Indeed, despite representing a small fraction of protein-coding genes, selenoproteins perform many critical functions in redox homeostasis, stress response, hormone maturation, and protein quality control, thus playing key roles in both physiology and cancer biology. This has prompted extensive research into their enzymatic functions, selenium metabolism, and the unique mechanism of Sec biosynthesis. Selenium itself, a trace element with a narrow margin between deficiency and toxicity, has also drawn interest in nutritional and physiological contexts (Fairweather-Tait et al. 2011). As a result, selenoproteins are investigated in a broad range of species, from standard model organisms (e.g., mouse, zebrafish) to agriculturally relevant animals (e.g., cattle, chicken, salmon). Naturally, understanding the full set of selenoproteins encoded by an organism is essential for the meaningful design and interpretation of experiments to elucidate selenium biology.

The human and mouse selenoproteomes were first described in 2003 as comprising 25 and 24 Sec-encoding genes, respectively (Kryukov et al. 2003). Subsequent studies expanded the known selenoproteomes of selected vertebrates and metazoans (Castellano et al. 2004; Castellano et al. 2005; Novoselov et al. 2006; Shchedrina et al. 2007; Jiang et al. 2010; Jiang et al. 2012). In our seminal paper in 2012, we provided the first description of selenoproteome composition and evolution across diverse vertebrates (Mariotti et al. 2012). Although highly influential, that study was constrained by the limited availability of sequenced genomes at the time (44).

In the last 10 years, major advances in sequencing and the rise of biodiversity genomics provided an unprecedented wealth of genomic data (Lewin et al. 2018; Mc Cartney et al. 2024). However, due to their deviation from the standard genetic code, selenoproteins

have not kept pace with the expansion of automated genomics resources. At present, 90% and 50% of selenoprotein genes have substantial annotation errors in Ensembl and NCBI genomes, respectively (Ticó et al. 2024). This also led to unreliable selenoprotein data in evolutionary resources dedicated to orthology (e.g. Ensembl Compara (Vilella et al. 2009), EggNOG (Powell et al. 2014)).

In this manuscript, we present a refined comprehensive characterization of the selenoproteome of vertebrates, obtained by meticulously tracing the evolution of 19 protein families across hundreds of genomes through sequence analysis and phylogenetic reconstruction. Our work provides an essential evolutionary framework to understand the selenoprotein content of any vertebrate genome, expanding on previous efforts by an order of magnitude.

RESULTS

We downloaded 315 genome assemblies from Ensembl representative of all major lineages of vertebrates, complemented by 25 NCBI genomes from early vertebrates and outgroups (Methods). Next, we used the software Selenoprofiles to identify homologs belonging to any of the known metazoan selenoprotein families in those genomes, resulting in 18,579 gene predictions. Phylogenetic reconstruction was then performed for each family, and gene and species trees were manually inspected to infer gene evolutionary events. These comprise gene duplications, gene losses, and the selenoprotein-specific event of the conversion of a Sec residue to Cys. In selenoprotein evolution, Cys to Sec events are also possible, but have been observed only along evolutionary distances much larger than encompassed by vertebrates (Mariotti et al. 2015; Zhang et al. 2022). Each branch of gene trees was manually annotated as a separate orthologous group, assigning gene names as consistent as possible with existing resources and literature. Each putative event was critically assessed, following criteria of

101 maximum parsimony, and seeking consistency with the species tree and support by
102 independent data (e.g. RNA sequences, other genome assemblies from NCBI) (Methods).

103 Thanks to the vast diversity of organisms in our dataset, the great majority of events were
104 shared among multiple species, leading to phylogenetically coherent patterns of gene
105 occurrence that allowed reliable evolutionary inferences. Yet, two lineages stood out for
106 their divergence from this general trend. First, bird genomes exhibited many apparent
107 gene losses scattered across lineages, in a pattern that would imply dozens of
108 independent events (Supplementary Figure S1). This included the apparent absence of
109 SEPHS2, a gene that nevertheless is required for biosynthesis of all other selenoproteins
110 (Mariotti et al. 2015). We ascribed this pattern to a known bias of bird assemblies: they
111 systematically lack certain GC-rich regions that present technical challenges for
112 sequencing (Botero-Castro et al. 2017). Indeed, we found RNA evidence of expression
113 for these putatively lost genes, indicating that they are indeed present in the real
114 biological genomes of birds. Ultimately, we dismissed all selenoprotein losses in Sauria
115 (which includes birds) as artefactual.

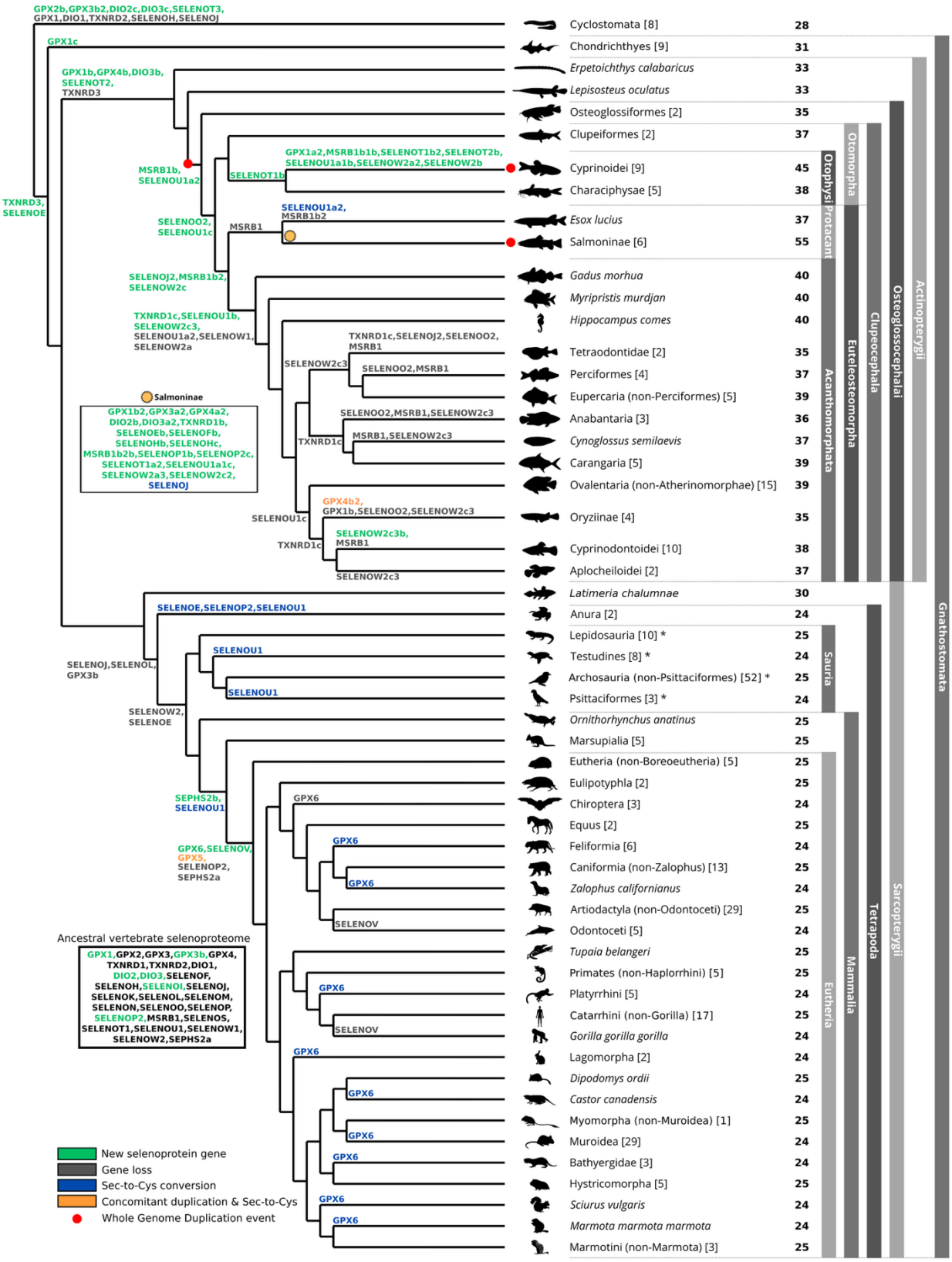
116 Second, the fish lineage of Cyprinoidei (e.g. carps, zebrafish *Danio rerio*) exhibited a
117 massive amount of apparent gene duplications (Supplementary Figure S1), which we
118 could not precisely disentangle by phylogenetic analysis. We realized that the sublineage
119 of Cyprininae (e.g. carps) reportedly underwent extensive recent whole genome
120 duplications and polyploidy, which greatly complicated this task. Therefore, we decided
121 to limit our analysis to the evolutionary events that occurred at the root of Cyprinoidei,
122 i.e. shared between zebrafish and Cyprininae, and dismiss the rest.

123 **Overview of the vertebrate selenoproteome**

124 We identified selenoproteins belonging to 19 protein families. The most represented
125 families were glutathione peroxidases with 1,529 Sec-encoding genes, iodothyronine

deiodinases with 999, and thioredoxin reductases with 721. In general, ray-finned fishes (Actinopterygii) exhibited richer selenoproteomes (33-55 selenoprotein genes) than tetrapods (24-25 genes).

While our analyses replicated well earlier reports performed at a smaller scale, we detected many novel events, particularly gene duplications. The majority affected Actinopterygii and its sublineages, whose selenoproteome appeared not only larger, but also more dynamic. In total, we detected 52 gene duplications (40 in Actinopterygii), and 41 gene losses (25 in Actinopterygii). The most selenoprotein-rich organisms were Salmoninae (e.g. trouts, salmons) with 55 selenoprotein genes, 17 of which emerged by duplication specifically in this lineage, and Cyprinoidei with 45 selenoproteins, 7 of which lineage-specific. On the other hand, Sec-to-Cys conversions were more prevalent in Tetrapoda: we detected 19 of these events overall, 16 of which were in tetrapods. Hereafter, we describe the selenoproteome evolution for each family separately.



139

140 ***Figure 1. Evolution of the vertebrate selenoproteome.*** *The evolutionary events shaping*

141 *selenoprotein genes are shown along the species tree. Leaf nodes represent either*

individual species (italicized) or collapsed lineages; in the latter case, the number in brackets indicates how many species from that lineage are included in our main dataset. Asterisks ('') denote lineages with apparent gene losses that we interpret as artefacts caused by known issues in genome assembly quality. Bold numbers in the rightmost column indicate the number of selenoprotein genes detected in each species or lineage. Grey rectangles highlight taxonomic clades of interest; "Protacant" refers to Protacanthopterygii. Animal silhouettes were sourced from https://www.phylopic.org.*

Glutathione peroxidases (GPX)

Glutathione peroxidases (GPX) are key antioxidant enzymes that protect cells from oxidative damage by reducing hydrogen peroxide and lipid peroxides using glutathione as an electron donor (Toppo et al. 2008; Mariotti et al. 2012). Mammals possess eight GPX homologs with specialized tissue distribution and functions, including five selenoproteins: GPX1 (ubiquitous), GPX2 (gastrointestinal), GPX3 (extracellular, kidney), GPX4 (membrane-associated, ubiquitous), GPX6 (olfactory system); and three Cys paralogs: GPX5, GPX7, and GPX8. Our phylogenetic analysis (Figure 2, Supplementary Figure S2) confirms a well-supported evolutionary classification of GPX proteins, identifying three major groups: GPX1/GPX2, GPX3/GPX5/GPX6, and GPX4/GPX7/GPX8. These findings align with previous studies on GPX evolution (Toppo et al. 2008; Mariotti et al. 2012; Xue et al. 2022).

GPX1, GPX2, GPX3, GPX4 are present in all vertebrates, with the only exception of Cyclostomata that lost GPX1. On the other hand, GPX6 originated at the root of placentals and exhibited a very dynamic evolution. Multiple independent GPX6 Sec-to-Cys conversions had been reported (Mariotti et al. 2012), which we confirmed and expanded in our analyses (Figures 1, 2b, 2d, 3; Supplementary Figure S2). We observed Cys conversions in Platyrrhini primates (New World monkeys, e.g. marmoset, titi), Muroidea rodents (e.g. mouse, rat, hamster), Lagomorpha (e.g. rabbit, hare, pika), Feliformia (e.g.

168 cats, hyenas), Bathyergidae rodents (e.g. mole-rats), and *Marmota marmota marmota*
169 (Figure 3). Furthermore, GPX6 was lost altogether in Chiroptera (bats).

170 We identified several lineage-specific GPX gene duplications across vertebrates. In
171 Cyclostomata (jawless fish), we observed two lineage-specific gene duplications
172 resulting that yielded GPX2b and GPX3b2. In Chondrichthyes (cartilaginous fish), a
173 duplication of GPX1 gave rise to a distinct Sec-paralog, GPX1c. We also confirmed three
174 two previously reported GPX gene duplications at the root of Actinopterygii, leading to
175 the emergence of GPX1b, and GPX4b. These genes were found in all investigated species
176 within this lineage, except for Oryziinae (e.g. medaka), which lost GPX1b. Curiously,
177 Oryziinae also exhibited a Cys-containing paralog emerged by duplication of GPX4b,
178 named GPX4b2. A third gene, GPX3b, was also reported in Actinopterygii; our analysis
179 indicates that it originated at the root of Vertebrata, and was subsequently lost in
180 Tetrapoda.

181 We also observed additional selenoprotein gene duplications in specific branches of
182 Actinopterygii: three in Salmoninae yielding GPX1b2, GPX3a2, and GPX4a2; and one in
183 Cyprinoidei yielding GPX1a2. These duplications were previously reported in an analysis
184 of 8 fish genomes (Xue et al. 2022). Our study demonstrates that they are lineage-
185 specific events, so that we have assigned new names to the duplicated genes to reflect
186 their precise evolutionary trajectories. The correspondence with (Xue et al. 2022) are
187 provided in Supplementary Table T1.

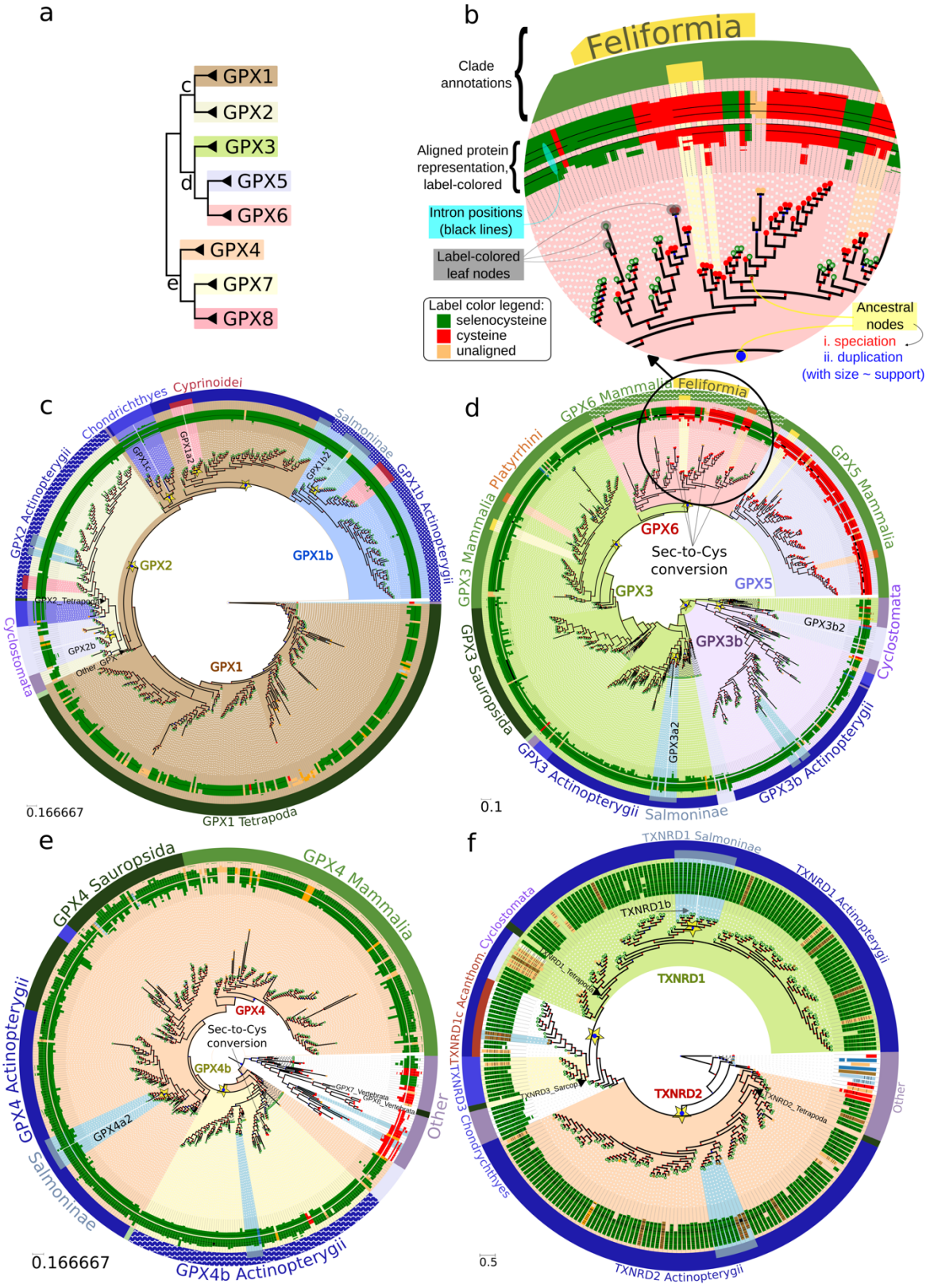


Figure 2. Phylogeny of the GPX and TXNRD families. a) *Topology of the GPX family, split into three clades shown in panels* ***c–e****.* ***b)*** *Zoomed section of the GPX6 gene tree, used*

here to illustrate this form of representation, abundantly used throughout this paper. Each leaf correspond to a putative gene whose protein sequence was included inthe tree. Leaves are color-coded by their label, assigned by Selenoprofiles by examining which codon aligns to the Sec position in the profile (e.g. green for Sec-UGA, red for Cys, orange for unaligned). Moving outward, you will see an equally colored representation of the corresponding aligned protein, which shows its non-gap positions and includes black lines to map where introns are located in the coding sequence. Most intronless genes are pseudogenes, which constitute the prevalent false positives of Selenoprofiles. The outmost circle provides a manual annotation of gene clades, with colors that indicate the taxonomy of the source of those genes (e.g. blue for Actinopterygii, green for mammals). Internal nodes of the gene tree represent either speciation events (red) or duplication events (blue), determined via the species overlap algorithm. The latter are displayed with size proportional to the number of species supporting the duplication. Duplication events deemed genuine via manual analysis are marked with a star. The GPX gene tree visualization is split into groups: ***c)*** *GPX1/GPX2,* ***d)*** *GPX3/GPX5/GPX6,* ***e)*** *GPX4/GPX7/GPX8.* ***f)*** *Gene tree visualization of the TXNRD family. Note that selected clades of the gene tree that did not exhibit evolutionary events were collapsed (e.g. see GPX7 and GPX8 in panel e). Expanded gene trees are available as Supplementary Figures S2 (GPX) and S5 (TXNRD).*

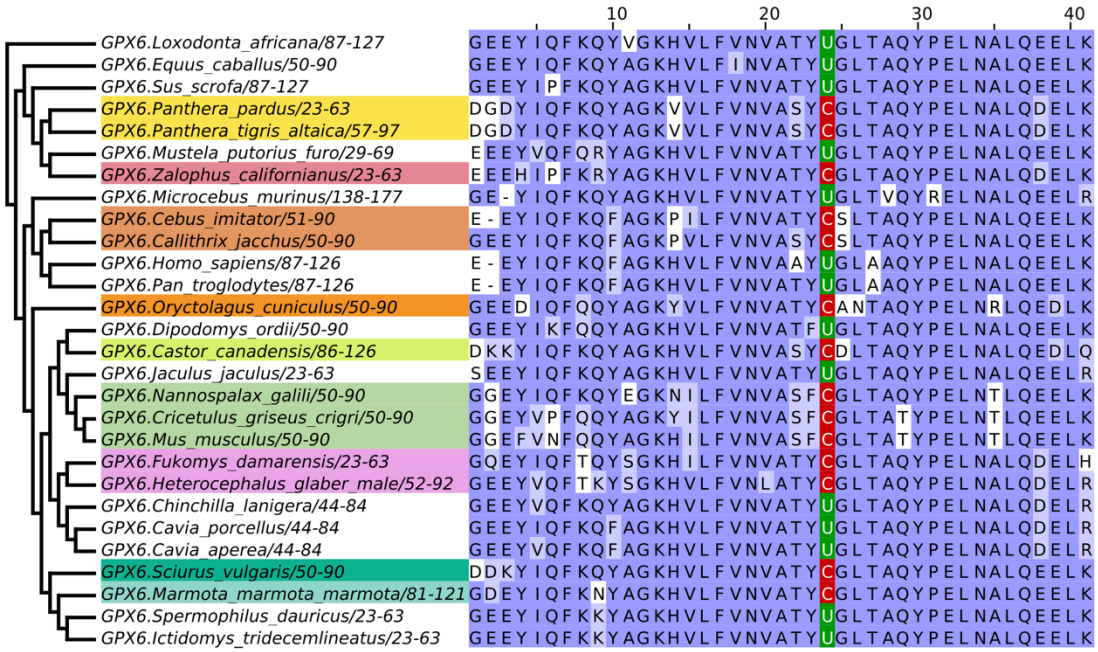


Figure 3. GPX6 multiple sequence alignment. *Selenocysteine residues are shown in green while Cysteine residues are shown in red. Species containing a GPX6-Cysteine homolog are highlighted by lineage.*

<u>Iodothyronine deiodinases (DIO)</u>

Iodothyronine deiodinases (DIO) regulate thyroid hormone metabolism by catalyzing the activation or inactivation of thyroxine (T4) and triiodothyronine (T3) (Darras and Van Herck 2012). The vertebrate DIO family includes three main genes: DIO1, which is involved in peripheral thyroid hormone conversion; DIO2, which plays a critical role in local T3 activation; and DIO3, which inactivates excess thyroid hormones, preventing hyperthyroidism. Through our analyses (Figures 1, 4, Supplementary Figures S1, S3), we detected these three genes in all vertebrate lineages, with the exception of Cyclostomata that lacked DIO1. This gene is present in all investigated Chordates, indicating that DIO1 predate the vertebrate radiation and was then lost in Cyclostomata. All Actinopterygii additionally carried a DIO3 duplicate (DIO3b). We also confirmed two more duplications, one of the ancestral DIO3 gene (also called DIO3a) (Lorgen et al. 2015), to

generate DIO3a2, and the other of DIO2 to yield DIO2b (Figures 1, 4a, 4b, Supplementary Figures S3, S4). Both were previously reported in Salmonids (Lorgen et al. 2015), and we traced them further back to the root of Salmoninae. In Cyclostomata, we identified two lineage-specific duplications resulting in DIO2c and DIO3c.

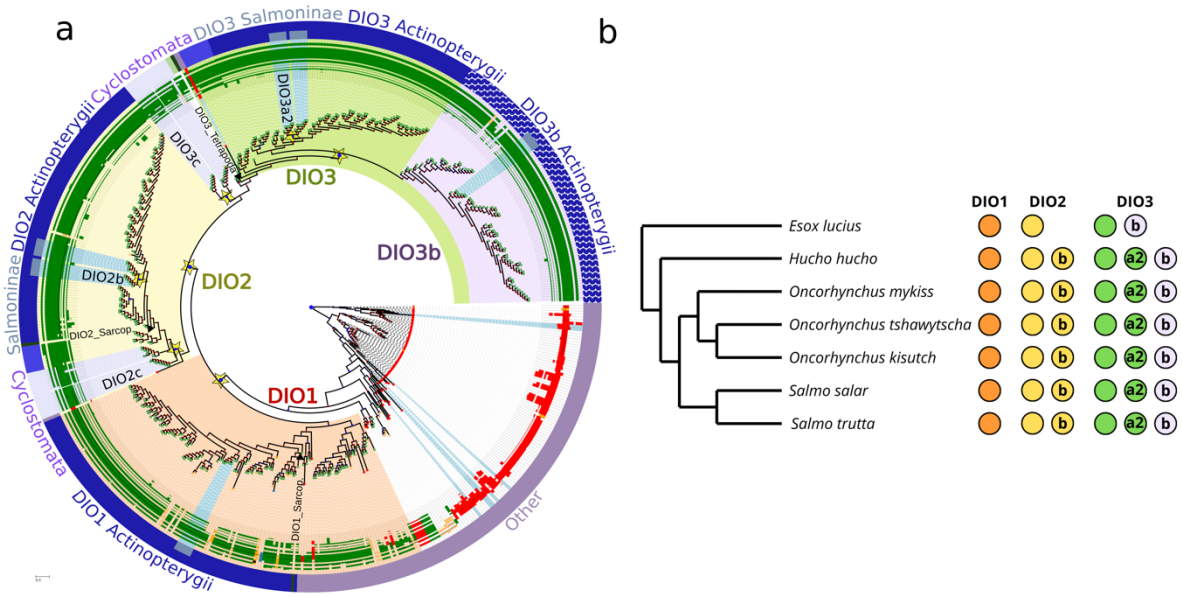


***Figure 4. Phylogeny of the DIO family**. **a)** Gene tree visualization of the DIO family. See full feature description in Figure 2b. **b)** Protacanthopterygii-specific DIO evolution. Selenoprotein genes are shown as circles. Genes are color-labelled according to panel A. Expanded gene and species trees are available as Supplementary Figures S3 and S4.*

<u>Thioredoxin reductases (TXNRD)</u>

Thioredoxin reductases (TXNRDs) play a central role in maintaining cellular redox balance by catalyzing the reduction of thioredoxins, which are essential for regulating oxidative stress responses, cell proliferation, and redox-regulated signaling cascades (Arnér 2009). Tetrapods have three TXNRD genes: TXNRD1 (cytosolic), TXNRD2 (mitochondrial), and TXNRD3/TGR (testis-specific, containing a peculiar N-terminal Grx domain), each of which participates in distinct physiological processes.

242 Reportedly, fishes lack the TXNRD3/TGR gene, but their TXNRD1 gene encodes for a Grx-
243 containing protein as its major isoform, which duplicated to give rise to TXNRD3/TGR in
244 tetrapods (Dou et al., 2022; Mariotti et al., 2012). Yet, our present analysis (Figures 2f, 5,
245 Supplementary Figure S5) did not recapitulate TXNRD evolution as previously proposed.
246 Because we detected TXNRD3 in Chondrichthyes, a lineage predating Actinopterygii, we
247 argue that this gene emerged earlier than tetrapods, and was subsequently lost in ray-
248 finned fishes. We did not find TXNRD3 in Cyclostomata, and thus we map the
249 TXNRD1/TXNRD3 duplication to the root of Gnathostomata. Moreover, because the
250 Cyclostomata gene clearly clusters with TXNRD1, we identify this clade as the ancestral
251 one and TXNRD3 as the gene derived from its duplication.

252 It is reported that the human TXNRD1 gene encodes for a minor transcript isoform that
253 includes this domain ("TxnRd1_v3"). This is absent in mouse, while in fish it represents
254 the predominant TXNRD1 variant (Arnér 2009). Therefore, we systemically investigated
255 the occurrence of Grx domains in the genomic regions of TXNRD1 and TXNRD3 across
256 all vertebrates (Methods). Our results show that Grx domains are a widespread feature
257 of TXNRD1 genes (Figure 5). Yet, they have been lost by several tetrapod lineages
258 independently (Glires, Chiroptera, Sauria, Equus, and Pecora), possibly functionally
259 replaced by TXNRD3. We argue that the Grx-domain is ancestral feature of TXNRD1.

260 We further investigated the ancestral state of vertebrates by examining the position of
261 metazoan outgroups to vertebrates in the TXNRD gene tree (Figure 2f), as well as an
262 extended gene tree featuring additional outgroups available in PhylomeDB (QFO2020
263 seed, *Caenorhabditis elegans* phylome ID: 791). Our phylogenetic analyses indicate that
264 the divergence between TXNRD1/TXNRD3 and TXNRD2 predates the origin of
265 metazoans, as representatives from both clades are found in diverse animal lineages,
266 including the nematode *C. elegans*, the annelid *Helobdella robusta*, and the cnidarian
267 *Nematostella vectensis*. Secondary events are evident in several lineages: the tunicate

Ciona intestinalis lost TXNRD2, while arthropods such as *Ixodes scapularis* and *D. melanogaster* lost TXNRD1. Notably, *D. melanogaster* retains two TXNRD2-derived genes, presumably the result of a secondary duplication. Fitting with our analysis, non-vertebrate animals carried Grx-domains solely in TXNRD1 genes. Due to this domain, these invertebrate genes were referred as TGR in literature, despite weak or absent glutathione reductase activity (Fata et al. 2022). Altogether, we reconstruct that ancestral vertebrates carried a cytosolic Grx-containing TXNRD1 gene and a mitochondrial TXNRD2 gene, the former later giving rise to TXNRD3 in jawed vertebrates.

We further detected two TXNRD1 duplications. One occurred in Salmoninae and yielded TXNRD1b, which retained the Grx domain. Another one gave rise to TXNRD1c in Acanthomorphata, which did not conserve such domain, and also features an atypical GNCUG C-terminus (rather than GCUC found across TXNRDs). TXNRD1c was subsequently convergently lost in several acanthomorph lineages (Figure 1), including Tetraodontidae, Atherinomorphae, Anabantaria, and Carangaria. Lastly, all Cyclostomata (except *Eptatretus burgeri*) lost TXNRD2 and thus encode TXNRD1 as their sole gene in this family.

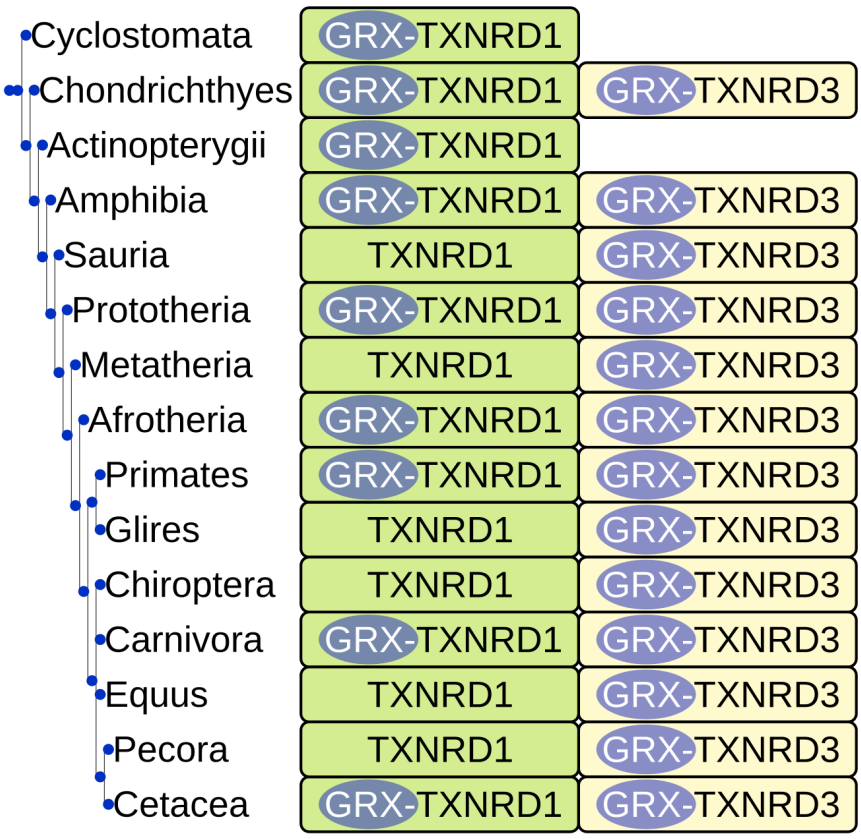


Figure 5. Phylogenetic distribution of the Grx domain in TXNRD1 and TXNRD3 proteins across vertebrate lineages. *Colors are assigned according to the TXNRD gene tree. A non-collapsed version of this figure is available as Supplementary Figures S6.*

<u>SELENOJ</u>

Selenoprotein J (SELENOJ) shows significant similarity to jellyfish J1-crystallins and, together with them, forms a distinct subfamily within the large family of ADP-ribosylation enzymes. While the homology may suggest a structural role which would be unique among selenoproteins, SELENOJ remains functionally uncharacterized. SELENOJ is reportedly restricted to actinopterygian fishes and sea urchins, with Cys homologs reported in cnidarians (Castellano et al. 2005). Our analyses replicated earlier reports, and additionally highlighted the loss of SELENOJ in Cyclostomata. Moreover, we had previously reported a SELENOJ duplication in medaka and stickleback, yielding SELENOJ2. Here, we traced back this duplication to the root of Euteleosteomorpha

298 (euteleosts), which included more than half of fishes in our dataset (Figures 1, 6a, 6b,
299 Supplementary Figures S7, S8). Later, SELENOJ2 was lost independently in
300 Tetraodontidae (e.g. pufferfish). On the other hand, SELENOJ2 underwent a Sec-to-Cys
301 conversion in Salmoninae.

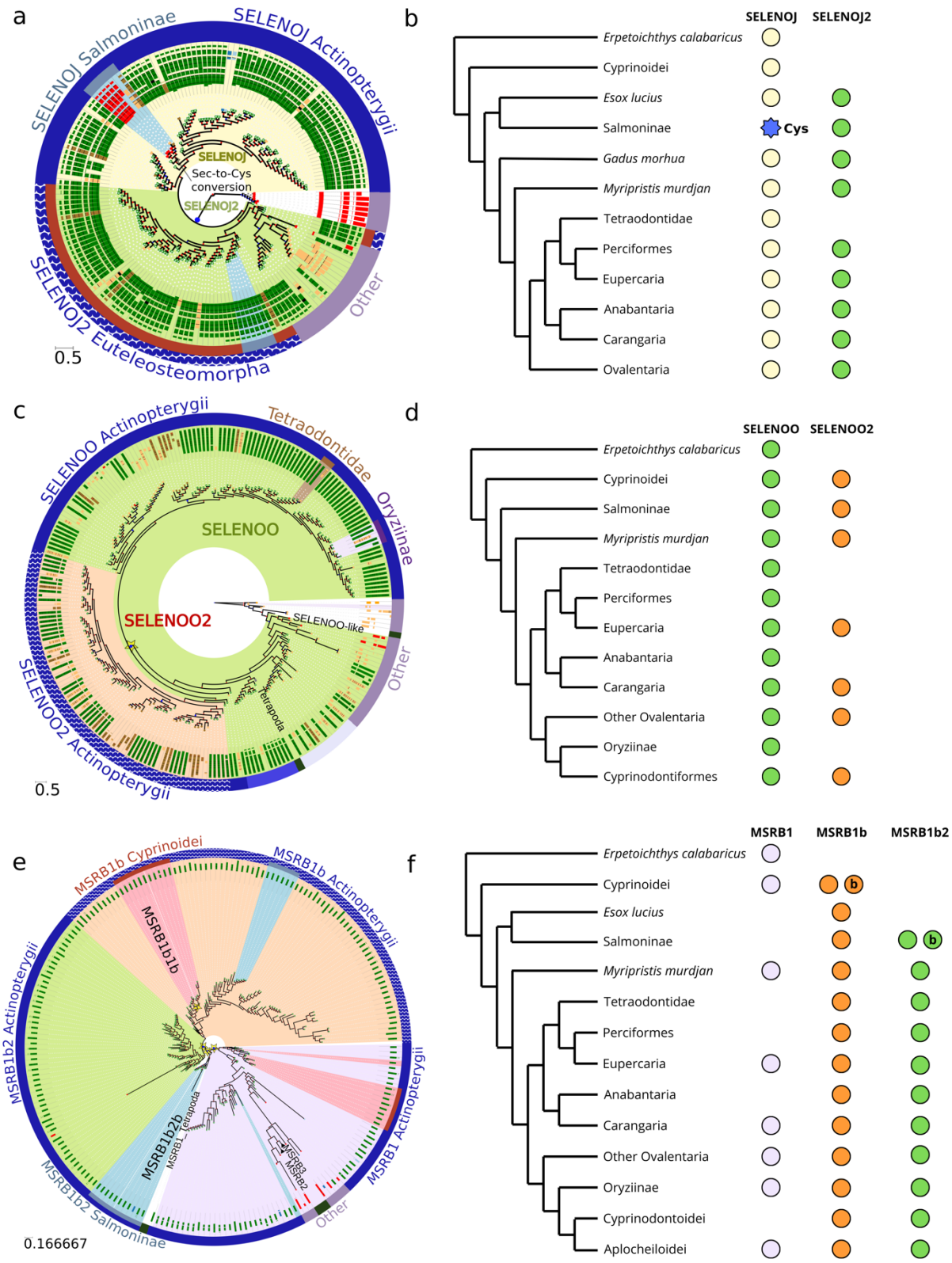
302 SELENOO

303 Selenoprotein O (SELENOO) is the largest known selenoprotein by molecular weight,
304 and belongs to a widespread protein family recently functionally characterized as a
305 mitochondrial pseudokinase that transfers AMP to various protein substrates, implicated
306 in oxidative stress response and metastasis (Sreelatha et al. 2018; Melo et al. 2025).
307 Previous studies identified a SELENOO paralog named SELENOO2 in zebrafish, which
308 was absent in the genomes of fugu, stickleback, medaka, and pufferfish (Mariotti et al.
309 2012). In our analysis, we replicated these results, and yet we observed the presence of
310 SELENOO2 in many other fish genomes, including among others all of Cyprinodontoidei
311 (toothcarps) and Carangaria (e.g. jacks, flatfishes), which implies a complex pattern of
312 gene evolution (Figures 1, 6c, 6d, Supplementary Figures S9, S10). Ultimately, we
313 hypothesize that SELENOO2 emerged at the root of Clupeocephala but was
314 subsequently lost in specific lineages, namely Tetraodontidae, Oryziinae, Anabantaria
315 (e.g. gourami), and Perciformes (e.g. stickleback).

316 Methionine sulfoxide reductase B (MSRB)

317 Methionine sulfoxide reductase B (MSRB, previously known as SELENOR), are enzymes
318 that reduce oxidized methionine residues, with roles for antioxidant defense and redox-
319 based modulation of various pathways including actin assembly (Tarrago et al. 2022).
320 Human possesses one Sec-containing gene, MSRB1, and two Cys paralogs, MSRB2 and
321 MSRB3. Our analyses indicate that the three genes predate the origin of vertebrates.
322 Within fish lineages, MSRB1 underwent a complex history of duplications and losses

323 (Figures 1, 6e, 6f, Supplementary Figures S11, S12). We previously identified a Sec-
324 containing gene, MSRB1b, in five ray-finned fishes (Mariotti et al. 2012). Our present
325 analysis refined its origin to the subclade of Osteoglossocephalai, which comprises all
326 Actinopterygii in our dataset except the MSRB1b-lacking species *Erpetoichthys*
327 *calabaricus* and *Lepisosteus oculatus*. This gene was then duplicated in Cyprinoidei
328 yielding MSRB1b1b, and in Euteleosteomorpha yielding MSRB1b2. The latter gene
329 duplicated again in Salmoninae, yielding MSRB1b2b, while the Northern pike Esox lucius
330 (its sister species within Euteleosteomorpha) lost MSRB1b2. On the other hand, MSRB1
331 has been lost in multiple lineages: Protoacanthopterygii (i.e. the group that includes
332 Salmoninae as well as *Esox lucius*), Tetraodontidae, Perciformes, Anabantaria,
333 Cyprinodontoidei, and the Chinese tongue sole *Cynoglossus semilaevis*. Notably, this
334 reconstruction highlights at least four convergent events of MSRB1 loss in
335 Percomorphacae, as we found MSRB1 only in two subclades, Carangaria (five genomes)
336 and non-Perciformes Eupercaria (five genomes). Notably, other selenoprotein families
337 show similar patterns of convergent losses in this lineage (Figure 1), as discussed later.



338

339 ***Figure 6. Phylogeny of the SELENOJ, SELENOO, and MSRB families. a, c, e)*** *Gene tree*

340 *visualizations of the SELENOJ, SELENOO, and MSRB families, respectively. For tree*

341 *annotation and legend details, see Figure 2b.* ***b, d, f)*** *Actinopterygii species tree*

342 *representations showing the presence of SELENOJ, SELENOO and MSRB paralogs,*
343 *respectively. Expanded gene and species trees are available: Supplementary Figure S7,*
344 *S8 (SELENOJ), S9, S10 (SELENOO), S11, S12 (MSRB).*

345 SELENOT

346 Selenoprotein T (SELENOT) is a thioredoxin-like enzyme anchored to the endoplasmic
347 reticulum (ER) membrane. It is highly conserved throughout evolution, and it has an
348 essential role in cellular homeostasis and redox regulation (Anouar et al. 2018). Previous
349 studies reported three SELENOT genes in actinopterygian fishes (Mariotti et al. 2012),
350 here refined and expanded. We found that most actinopterygian groups possessed two
351 SELENOT genes, namely SELENOT (aka SELENOT1) and SELENOT2; yet, Salmoninae,
352 Otophysi and Cyprinoidei exhibited further lineage-specific duplications. Ultimately, we
353 mapped the origin of SELENOT2 to a duplication of SELENOT1 at the root of
354 Actinopterygii. SELENOT1 duplicated again to yield SELENOT1b in Otophysi, i.e. the group
355 including Cyprinoidei (e.g. zebrafish and carps), and Characiphysae (e.g. tetra, piranha).
356 In Cyprinoidei, further duplications of SELENOT2 and SELENOT1b originated the genes
357 SELENOT2b and SELENOT1b2, respectively (Figures 1, 7a, Supplementary Figure S13).
358 Finally, we observed yet another SELENOT1 duplication in Salmoninae, yielding the gene
359 SELENOT1a2. Interestingly, Cyclostomata also independently expanded this family
360 through one gene duplication, which yielded a divergent paralog that we designated
361 SELENOT3.

362 SELENOU

363 The Selenoprotein U (SELENOU) family includes selenoprotein genes in birds and fishes,
364 the latter reported to carry up to three Sec-containing paralogs (Castellano et al. 2004;
365 Mariotti et al. 2012). Mammals and amphibians are reported to encode only Cys-
366 homologs, with human containing three paralogs (SELENOU1/PRLX2A,

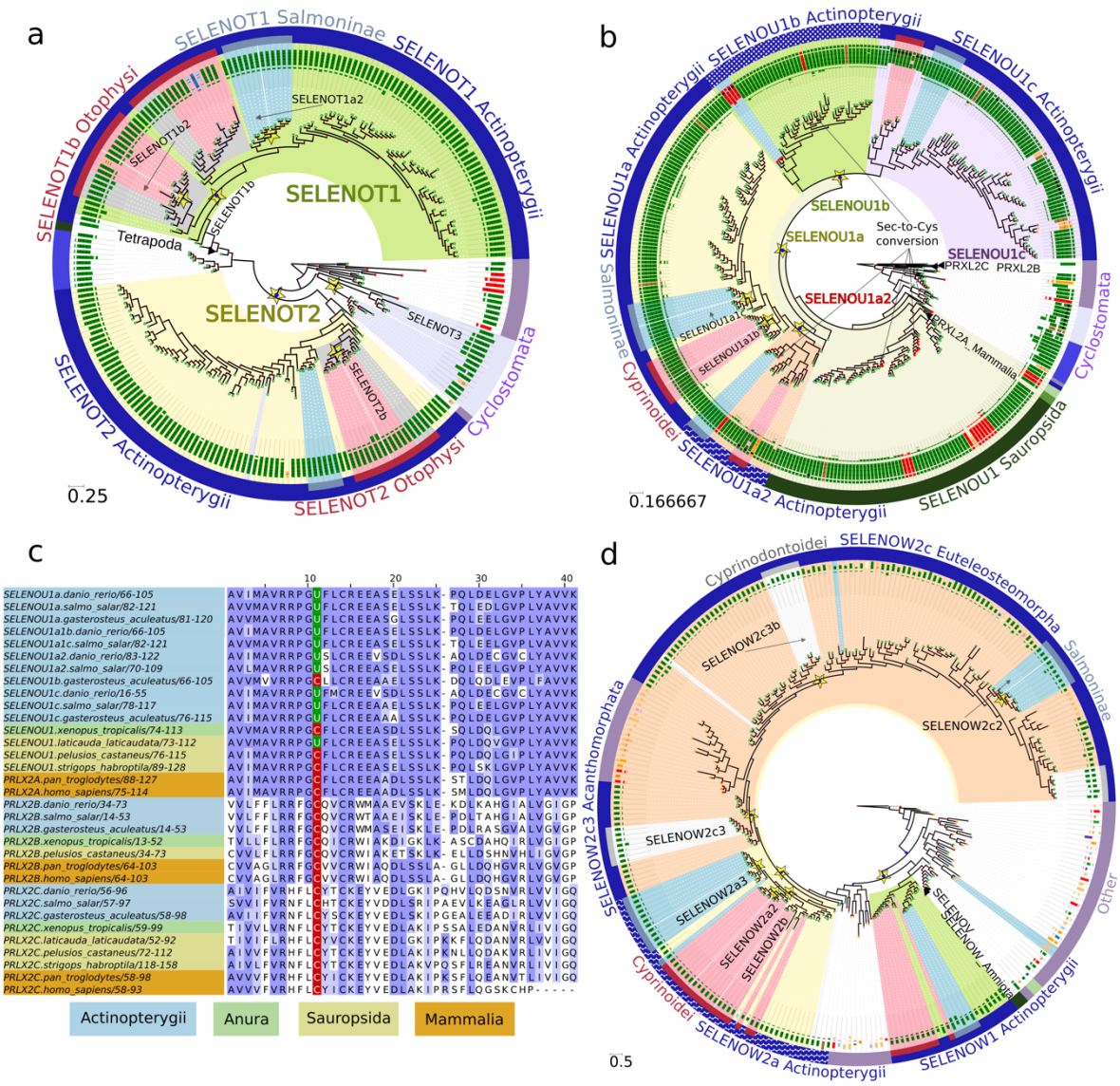
367 SELENOU2/PRLX2C, SELENOU3/PRLX2B), involved in diverse functions such as
368 antioxidant defense, prostaglandin biosynthesis, and glycolysis regulation. SELENOU
369 exhibits a scattered pattern of gene occurrence, implying a complex evolutionary history
370 involving gene duplications, losses, and Sec-to-Cys conversions. Here, we provide our
371 best attempt to explain this with current data (Figures 1, 7b, 7c, Supplementary Figures
372 S1, S14).

373 The split of SELENOU1/2/3 predates metazoa, and all Sec-encoding genes detected in
374 this family belonged to the SELENOU1 branch. Within fish, we mapped a first duplication
375 of the SELENOU1 gene in Osteoglossocephalai, yielding SELENOU1a2. This gene was then
376 lost at the root of the large group of Acanthomorphata, while it converted to Cys
377 homolog in *Esox lucius*. Prior to both of these events, we traced two additional
378 SELENOU1 duplications in Clupeocephala and Acanthomorphata, yielding the
379 SELENOU1c and SELENOU1b genes, respectively. SELENOU1b was then lost at the root
380 of Ovalentaria. Finally, we detected another two independent SELENOU1 duplications in
381 Cyprinoidei and Salmoninae, yielding the gene SELENOU1a1b and SELENOU1a1c,
382 respectively. Within tetrapods, SELENOU1 underwent at least four convergent Sec-to-
383 Cys conversions (Figure 7c), specifically in Theria (i.e. placentals and marsupials), Anura
384 (frogs), Testudines (turtles), and Psittaciformes birds (parrots).

385 <u>SELENOW and SELENOV</u>

386 Selenoprotein W (SELENOW) is a widespread glutathione-dependent antioxidant
387 oxidoreductase implicated in resolution of inflammation, with reported roles in muscle
388 growth and differentiation, as well as in protecting neurons from oxidative stress during
389 neuronal development (Jeong et al. 2002; Misra et al. 2023). Selenoprotein V (SELENOV)
390 is a placental-specific putative oxidoreductase expressed in testis and potentially
391 involved in antioxidant defense and ER stress, whose C-terminus shares substantial
392 sequence homology to SELENOW (Zhang et al. 2020).

393 It was previously reported that SELENOV was lost in *Gorilla gorilla* (Mariotti et al. 2012).
394 Our present analysis identified an additional case of SELENOV loss, specifically in
395 Odontoceti (toothed whales, e.g. dolphins, killer whales) (Figures 1, 7d, Supplementary
396 Figure S15). Regarding the SELENOW branch of this family, mammals have a single gene:
397 SELENOW, also referred to as SELENOW1. Earlier work, replicated and expanded here,
398 pinpointed multiple SELENOW homologs in non-mammalian vertebrates, with a clearly
399 distinct gene clade named SELENOW2. The split of SELENOW1 and SELENOW2
400 predates the vertebrate radiation. Among tetrapods, we detected SELENOW2 only in
401 amphibians (Anura), so that we infer that it was lost at the root of Amniota. In fish, the
402 phylogenetic history of the SELENOW is remarkably complex (FIGURE 7d,
403 Supplementary Figure S15): it constitutes the most difficult vertebrate selenoprotein
404 family to solve conclusively, since no simple scenario can explain the observed pattern
405 of gene presence. In our reconstruction (Figure 1), a duplication of SELENOW2 in
406 Euteleosteomorpha led to the emergence of SELENOW2c. Further along this lineage,
407 specifically in Acanthomorphata, SELENOW1 and SELENOW2a were then lost, so that
408 these genes are absent from the majority of extant fish genomes. On the other hand, we
409 observed several lineage-specific duplications. Two SELENOW2 duplications occurred
410 in Cyprinoidei, yielding genes that we named SELENOW2a2 and SELENOW2b.
411 Salmoninae exhibited another two duplications in the SELENOW2 clade that we
412 attributed to the SELENOW2 and SELENOW2c genes, yielding SELENOW2a3 and
413 SELENOW2c2, respectively. Then, SELENOW2c duplicated in Acanthomorphata
414 yielding SELENOW2c3, and the resulting gene duplication later within this lineage, in
415 Cyprinodontoidei, giving rise to SELENOW2c3b. A series of scattered losses in the
416 SELENOW2c3 clade were also detected but challenging to trace with precision. Our
417 maximum-parsimony reconstruction implies five independent losses of SELENOW2c3
418 within various Acanthomorphata branches (Figure 1).



419

420 ***Figure 7. Phylogeny of the SELENOT, SELENOU and SELENOW families. a, b, d)*** *Gene*
421 *tree visualizations of the SELENOT, SELENOOU and SELENOW families, respectively.*
422 *For tree annotation and legend details, see Figure 2b.* ***c)*** *SELENOU multiple sequence*
423 *alignment. Species are color-labelled according to lineage. Selenocysteine residues are*
424 *marked in green and Cysteine residues in red. Expanded gene trees are available:*
425 *Supplementary Figures S13 (SELENOT), S14 (SELENOU), S15 (SELENOW).*

426 <u>SELENOP</u>

427 Selenoprotein P (SELENOP) is the only mammalian selenoprotein with multiple Sec
428 residues, e.g. 10 in human and mouse, due to its Sec-rich C-terminal tail. SELENOP is a
429 key player in selenium transport and distribution, serving as the primary selenium carrier
430 in the bloodstream, delivering selenium to peripheral tissues via receptor-mediated
431 uptake. We recently characterized the evolution of SELENOP, showing that its Sec
432 content can vary by two orders of magnitude across metazoa (Baclaocos et al. 2019). Its
433 C-terminal tail is challenging to identify using genomic data due to its fast evolution, so
434 that here we mostly focused on gene-level detection. As expected, we identified two
435 phylogenetic clades of vertebrate genes, multi-Sec SELENOP/SELENOP1, and single-Sec
436 SELENOP2/SELENOPB. While the former is reportedly ubiquitous in vertebrate
437 genomes, the latter was lost in Eutheria. Moreover, SELENOP2 was converted to a Cys
438 homolog in Anura. Our present analysis (Figure 8a, Supplementary Figure S19) confirmed
439 these findings, and offered additional insights into early vertebrate evolution.

440 Invertebrate SELENOP homologs are consistently placed at the base of the gene tree,
441 supporting the hypothesis that the SELENOP1/SELENOP2 duplication arose within the
442 vertebrate lineage. Notably, Cyclostomata retain the two paralogs, yet both highly
443 divergent in sequence. One gene occupies a basal position in the tree even compared to
444 vertebrate outgroups, likely due to long-branch attraction artifacts; this is presumably
445 the SELENOP1 ortholog. The other Cyclostomata gene clusters within SELENOP2, and
446 displaying a unique CXXU motif (rather than the more typical UXXC) in its N-terminal
447 thioredoxin-like domain. Though SELENOP2 of other vertebrates carries a single Sec, in
448 Cyclostomata it exhibits an extraordinary expansion of Sec content: it harbors 162 and
449 139 in-frame UGA codons in *Lethenteron reissneri* (Far Eastern brook lamprey) and
450 *Lampetra fluviatilis* (European river lamprey), respectively (Supplementary Document
451 D1). Analogously to other metazoan multi-Sec SELENOP genes (Baclaocos et al. 2019),
452 the vast majority of these putative Sec residues are encoded within an expanded

453 modular C-terminal region, and two SECIS elements are detected downstream. Between
454 the SECISes, we noted the presence of a distinctive region rich in CA dinucleotides
455 (-1500-1800nt). Highly similar matches in TSA data of other Cyclostomata were
456 detected all across the putative transcript (Supplementary Document D1). With this
457 record, lampreys surpasses all other metazoans in the number of Sec-UGAs in SELENOP,
458 including the previous high of 132 reported in bivalves (Baclaocos et al. 2019).

459 Other Salmoninae-specific duplications: SELENOE, SELENOF, SELENOH

460 Our analysis revealed a substantial selenoproteome expansion specific to the lineage of
461 Salmoninae, which may suggest an increased functional reliance of this clade on
462 selenium. Besides the aforementioned expansions in the GPX, DIO, TXNRD, MSRB,
463 SELENOT, SELENOU and SELENOW families (Figure 1), we detected gene duplications
464 of SELENOE and SELENOF, yielding SELENOEb and SELENOFb, respectively, and another
465 two for SELENOH, yielding SELENOHb and SELENOHc (Figure 8, Supplementary Figures
466 S16, S17, S18). Briefly, SELENOF is a ER-resident selenoprotein involved in redox quality
467 control (Yim et al. 2018), recently pinpointed as tumor suppressor (Zigrossi et al. 2022;
468 Flowers et al. 2023). Our results show that SELENOF is ubiquitous to vertebrates. On the
469 other hand, SELENOE (formerly Fep15), a functionally uncharacterized selenoprotein
470 distantly related to SELENOF (Novoselov et al. 2006), was found as selenoprotein only
471 in fishes: Anura carry a Cys-homolog, while the rest of tetrapods (Amniota) lost the gene
472 altogether. Notably, we mapped the origin of SELENOE to the root of Gnathostomata,
473 since we did not find it in Cyclostomata nor in vertebrate outgroups. Finally, SELENOH
474 is a widespread nucleolar oxidoreductase that regulates redox homeostasis and
475 suppresses DNA damage through the p53 pathway, and it is implicated in cancer
476 proliferation (Wu et al. 2014; Cox et al. 2016; Bertz et al. 2018). Apart from the
477 Salmoninae-specific duplications, no other gene evolution event was detected for

478 SELENOE, SELENOF, and SELENOH, with the only exception of the loss of SELENOH in
479 Cyclostomata.

480 Notably, by inspecting the positions of introns relative to the alignment of protein
481 sequences (Figures 2, 4, 7), we noticed that they were conserved in all these Salmoninae-
482 specific gene duplications. This indicates a mechanism of DNA-based gene duplication
483 (as opposed to RNA-based transposition). The previously reported whole genome
484 duplication (WGD) at the root of salmonids (Lien et al. 2016) is therefore a most plausible
485 explanation for the origin of these genes. Importantly, this implies that we expect other
486 (non-Salmoninae) Salmonids to be also affected.

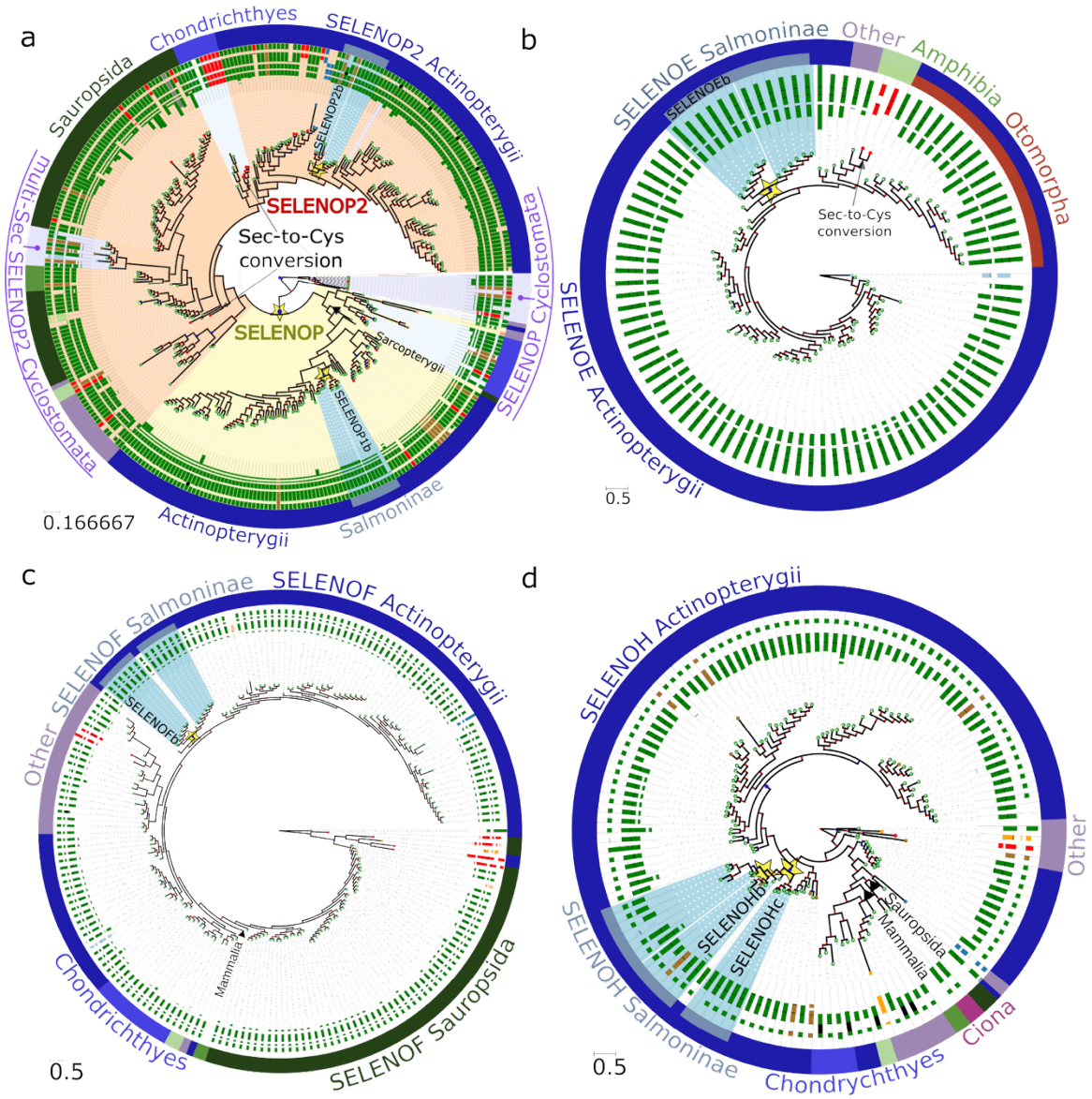


Figure 8. Phylogeny of SELENOP, SELENOE, SELENOOF, and SELENOH . a, b, c, d) *Gene tree visualizations of the* ***SELENOP, SELENOE, SELENOOF, and SELENOH*** *families, respectively. For tree annotation and legend details, see Figure 2b. Expanded gene trees are available: Supplementary Figure S19 (SELENOP), S16 (SELENOE), S17 (SELENOF), S18 (SELENOH).*

<u>Other families: SEPHS2, SELENOL, SELENOK, SELENOM, SELENON, SELENOS</u>

Hereafter, we briefly summarize the remaining vertebrate selenoproteome families, wherein our present analysis recapitulated earlier results without novel findings. As

496 already introduced, SEPHS2 (aka SPS2) is a selenoprotein whose function is critical for
497 Sec biosynthesis. As expected (Mariotti et al. 2012; Mariotti et al. 2015), we identified
498 SEPHS2 in all vertebrate lineages (barring its artefactual absence in many bird genome
499 assemblies, mentioned above). The SEPHS2 gene is intronless only in genomes of
500 Eutheria (placentals). As previously shown (Mariotti et al. 2012), the ancestral intron-
501 containing SEPHS2 gene (referred to as SEPHS2a) duplicated by retrotransposition at the
502 root of Mammalia, yielding an intron-less copy referred to as SEPHS2b. Both copies are
503 present in extant genomes of Marsupialia, while SEPHS2a was then lost in Eutheria. Our
504 results confirmed these findings (Supplementary Figures S1, S20). Thus, the sole
505 placental SEPHS2 gene is actually SEPHS2b. Of note, SEPHS2 duplicated at the root of
506 vertebrates, and independently also in various other metazoan lineages, to yield a non-
507 Sec/non-Cys paralog known as SEPHS1/SPS1 (Mariotti et al. 2015). The vertebrate
508 SEPHS1 gene carries threonine in place of Sec. This kind of conversion is highly unusual
509 for selenoproteins, since Sec is typically catalytic and it is highly dissimilar from canonical
510 amino acids other than Cys. Indeed, SEPHS1 appears to lack the enzymatic activity of
511 SEPHS2 (selenophosphate synthesis from selenide), and instead has a function in gene
512 expression-mediated control of redox homeostasis (Manta et al. 2022).

513 SELENOL is a peculiar selenoprotein featuring two nearby Sec residues that can form a
514 diselenide bond. It features a peroxiredoxin-like motive suggestive of a redox-related
515 function (Shchedrina et al. 2007). As expected (Mariotti et al. 2012), we detected
516 SELENOL in nearly all vertebrates except tetrapods, confirming a gene loss at the root of
517 this lineage (Supplementary Figure S1).

518 The remaining selenoprotein families were detected in a single Sec-containing copy in
519 all investigated genomes, barring pseudogenes and spurious absence attributed to
520 imperfect assembly quality. Curiously, all these single-copy selenoproteins localize to
521 the ER. They include: SELENOK, involved in calcium homeostasis and resolution of ER

522 stress (Odunsi et al. 2025); SELENOM and SELENOS, involved in unfolded protein
523 responses (Shi et al. 2024); and SELENON, involved in calcium and redox homeostasis
524 (Chernorudskiy et al. 2020) (Supplementary Figure S1).

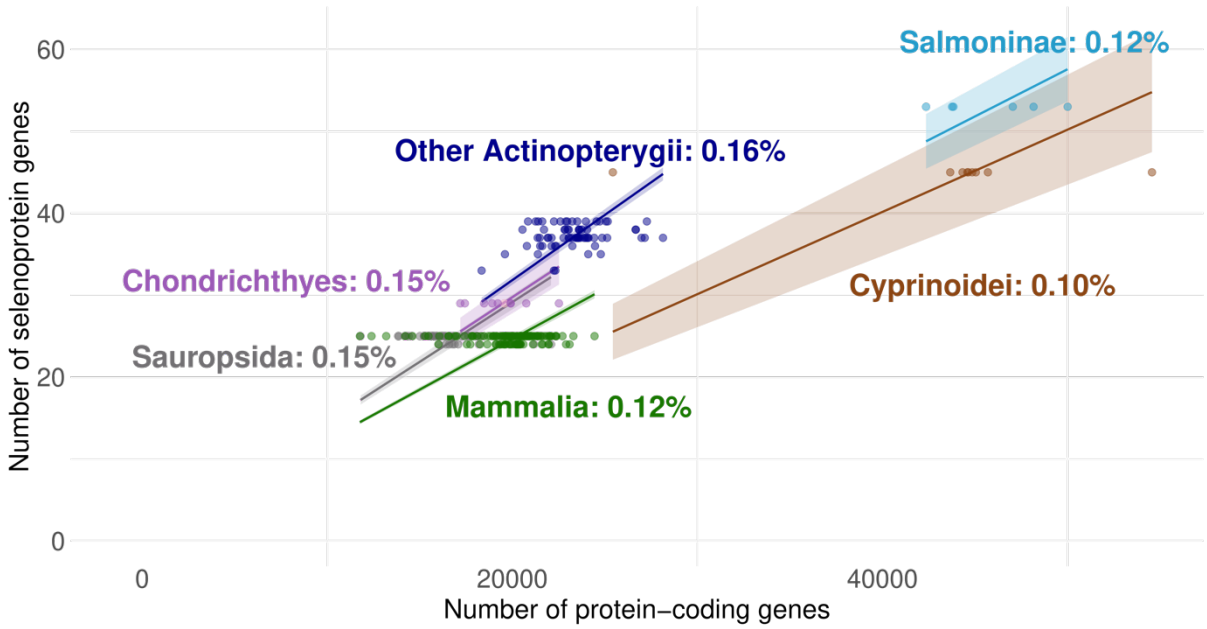
525 DISCUSSION

526 The recent acceleration of genome sequencing and the rise of global biodiversity
527 initiatives has generated an unprecedented abundance of genomic data across
528 vertebrate lineages. This surge presents a unique opportunity to revisit fundamental
529 questions about gene evolution, functional diversity, and lineage-specific adaptations.
530 Here, we leveraged this genomic wealth to provide an updated view of the composition
531 and evolution of the vertebrate selenoproteome, obtained through analyses of hundreds
532 of genomes and meticulous phylogenetic reconstruction which greatly expanded upon
533 previous endeavors. We reinforced and refined existing findings, and uncovered new
534 genes and evolutionary events.

535 Our reconstruction of the ancestral vertebrate selenoproteome consists of 28
536 selenoprotein genes representing all the 19 known vertebrate families (Figure 1). Our
537 conclusions mostly overlap with our earlier analysis (Mariotti et al. 2012), with the
538 exceptions of SELENOE (now mapped to Gnathostomata rather than Vertebrata), and
539 GPX3b (Vertebrata rather than Actinopterygii). We also concluded that these
540 selenoproteins emerged specifically at the root of Vertebrata: GPX1, GPX3b, DIO2, DIO3,
541 SELENOI, SELENOP2. Among these, SELENOI stands out in that the whole family
542 appeared at the root of vertebrates: there is no homolog that includes the Sec position,
543 regardless of the aligned amino acid, outside of vertebrates.

544 On the other hand, our present work substantially expands the selenoproteome in fish
545 lineages. The conservation of intron structures and the phylogenetic correspondence
546 with documented whole-genome duplications suggests that most of these gene

duplications resulted from such events (Figure 1). This raises the question of whether such selenoproteome expansions are adaptive, or are neutral consequences of having a larger proteome. To gain insights in this matter, we compared the size of our predicted selenoproteome with that of the annotated proteome for each species (Figure 9). As expected, the selenoproteome-to-proteome ratio is indeed lower in mammals than fish. Aquatic environments have long been hypothesized to favor selenoprotein usage, a trend first suggested by Lobanov et al. from limited genomic data (Lobanov et al. 2007). Subsequent analyses further corroborated this model by showing relaxed purifying selection on selenoprotein genes in terrestrial vertebrates (Sarangi et al. 2018), alongside reduced Sec content in SELENOP, an established proxy for selenium demand (Sarangi et al. 2018; Baclaocos et al. 2019). Our present analysis with a larger dataset mostly support this pattern. However, we note that the difference in selenoproteome-to-proteome ratios between aquatic and terrestrial species is primarily driven by mammalian loss and actinopterygian expansion. Some lineages diverge from this trend: aquatic Salmoninae and Cyprinoidei show low ratios, despite having the biggest selenoproteomes across vertebrates; and land-dwelling Sauropsida exhibit relatively high ratios (Figure 9).



564 ***Figure 9. Genomic correlation between the number of predicted selenoproteins and***
565 ***annotated proteome size.*** *Each point represents a species and is color-labelled*
566 *according to lineage.*

567 The underlying cause of the selective advantage of Sec in aquatic environments is
568 arguably yet unclear. It has been proposed that land-dwelling organisms must face (i)
569 reduced selenium bioavailability and (ii) increased oxygen exposure (Lobanov et al.
570 2007; Sarangi et al. 2018). While the latter seems intuitively relevant, it contradicts one
571 of the prevalent proposed biochemical rationale for usage of selenium over cysteine
572 (Reich and Hondal 2016; Maroney and Hondal 2018): its resistance to (over)oxidation - a
573 trait that would suggest selenoproteins should be more beneficial in terrestrial
574 environments. Moreover, our data shows that selenoproteins belonging to antioxidant
575 families are more prevalent in fish compared to tetrapods. While further research is
576 required to fully resolve this question, we argue that the current evidence better
577 supports selenium bioavailability as the most likely evolutionary driver in aquatic vs land
578 environments.

579 These observations may depict our own selenoproteins as an unstable remnant of an
580 aquatic past. Yet, it is worth noting that (i) land-living tetrapods exhibit substantial
581 purifying selection, preventing Sec to Cys conversions in most genes (Castellano et al.
582 2009); and (ii) the aquatic selenoproteome is far from static, having expanded under
583 putative adaptive forces at the root of Actinopterygii and in multiple of its clades. We
584 found that these duplications occurred primarily in families linked to antioxidant defense,
585 either with well-established roles (GPX, MSRB) or with putative/indirect functions
586 (SELENOT, SELENOU, SELENOW). In contrast, no variation was observed in ER-resident
587 proteins with functions in calcium homeostasis (SELENOK, SELENON) or protein quality
588 control / unfolded protein response (SELENOM, SELENOS). While this trend was not
589 absolute - e.g. ER-resident SELENOE and SELENOF duplicated in Salmoninae, and the

590 frequently duplicated SELENOT is also ER-resident - it suggests that selenoproteome
591 expansions in Actinopterygii may have been driven primarily by selection for enhanced
592 antioxidant capacity.

593 Among tetrapods, our most striking finding is the recurrence of convergent evolutionary
594 events, including Sec-to-Cys conversions in SELENOU1 and GPX6, as well as
595 independent losses of SELENOV, all of which are supported by robust phylogenetic
596 evidence. It is important to note that, because our reconstruction is based on maximum
597 parsimony, our estimate of such events represents a conservative lower bound.
598 Interestingly, both SELENOV and GPX6 are the "youngest" mammalian selenoproteins
599 and they are specifically expressed in the testis, a tissue recognized for its role in the
600 evolution of new genes (Kaessmann 2010).

601 Our study has some limitations. First, we restricted our search to genes homologous to
602 known selenoprotein families, because existing methods for identifying novel
603 selenoproteins suffer from low specificity (Santesmasses et al. 2020). As a result, we
604 may have overlooked lineage-specific selenoproteins, particularly in fish lineages where
605 selenoproteome expansions are evident. Second, we could only detect events within
606 lineages represented in our dataset. Given the dynamic nature of the Actinopterygii
607 selenoproteome, additional genome sequences from this group will likely reveal further
608 duplications. Third, our analysis assumes the species tree as ground truth, which may
609 not be entirely accurate. For instance, the convergent loss of the same genes in
610 Tetraodontidae, Perciformes, Anabantaria, and *Cynoglossus semilaevis* (Figure 1) could
611 be explained as a single ancestral loss if non-Perciform Eupercaria and Carangaria were
612 instead outgroups of these clades - yet, we found no supporting evidence for this
613 alternative phylogeny in the literature (Hughes et al. 2018). Fourth, further data or
614 alternative methodological approaches could refine the evolutionary history of
615 selenoprotein families with the most complex gene dynamics, such as SELENOW and

616 SELENOU, and provide deeper insight into lineages with highly complex genome
617 evolution, such as Cyprininae.

618 Despite these limitations, we anticipate that our characterization of vertebrate
619 selenoproteins will largely hold up over time and provide a meaningful framework for
620 future evolutionary and functional research.

621 MATERIAL AND METHODS

622 **Genome sequences and species tree**

623 We downloaded all genome assemblies and annotations available in Ensembl version
624 v.108 (Cunningham et al. 2022) comprising 315 genomes, each corresponding to a
625 different species or strain. The phylogenetic tree of species was downloaded from
626 Ensembl Compara (Vilella et al. 2009). To refine our reconstruction of the ancestral
627 vertebrate selenoproteome, we also analyzed NCBI assemblies from species with
628 informative taxonomy. Specifically, we included various early-branching vertebrates:
629 nine cartilaginous fish (Chondrichthyes) and eight jawless vertebrates (Cyclostomata);
630 several chordates, i.e. close outgroups of vertebrates: two ascidians (Tunicata), four
631 amphioxiformes (Branchiostoma); and two remote outgroups: sea urchin
632 *Strongylocentrotus purpuratus* (Echinodermata, deuterostome) and centipede
633 *Strigamia maritima* (Arthropoda, protostome). These 25 genomes, together with the 315
634 Ensembl assemblies, constituted our main dataset. For follow up analyses to confirm or
635 dismiss gene evolution events, we then inspected gene predictions on an extended set
636 of 1,010 genome assemblies downloaded from NCBI, using the NCBI taxonomy tree as
637 the rough backbone for evolutionary interpretation.

638 **Selenoprotein gene prediction**

639 Selenoprotein gene prediction was performed using Selenoprofiles (Mariotti and Guigó
640 2010; Santesmasses et al. 2018; Ticó et al. 2024) version v4.5.0, available at
641 https://github.com/marco-mariotti/selenoprofiles. Selenoprofiles is a computational
642 pipeline to identify members of the known selenoprotein families and related proteins.
643 It uses several homology-based gene prediction programs, and employs a set of
644 manually curated multiple sequence alignments of selenoprotein families to scan
645 genomes for homologues. The profiles we used are available at
646 https://github.com/marco-mariotti/selenoprotein_profiles (v1.3.0). We scanned
647 genome assemblies for all known selenoprotein families in metazoa. For most analyses,
648 we focused on the predictions labelled as "selenocysteine", i.e., the genes encoding for
649 selenoproteins, with no apparent pseudogene features (e.g. frameshifts).

650 **Phylogenetic reconstruction of selenoprotein families**

651 The protein sequences of all Selenoprofiles predictions across Ensembl genomes were
652 collected using the Selenoprofiles *join* utility. These were used as input to phylogenetic
653 reconstruction workflows, available in ETE3 v3.1.3 build framework (Huerta-Cepas et al.
654 2016). Phylogenetic trees were obtained with the "eggnog41" workflow (Powell et al.
655 2014), which internally uses PhyML v20160115.patched for phylogenetic reconstruction
656 (Guindon et al. 2010). We visualized gene trees with the show_tree program in the
657 biotree_tools package v0.0.6 (https://github.com/marco-mariotti/biotree_tools), based
658 on the ETE3 python module (Huerta-Cepas et al. 2016), specifically adapted for custom
659 visualization in this manuscript. The resulting vectorial images were further annotated
660 manually using the Inkscape software. Sequence and tree manipulations throughout the
661 project were carried out using the programs *alignment_tools* and *tree_tools* from the
662 biotree_tools package. Protein sequence alignments were computed using ClustalO
663 v.1.2.4 and processed with Pyaln v0.1.4 (https://github.com/marco-mariotti/pyaln).

664 **Analysis of gene evolution events**

665 We manually inspected gene and species trees to determine gene evolution events:
666 losses, duplications, and Sec-to-Cys conversions. To detect potential duplications, gene
667 trees were processed using the species overlap algorithm: the number of common
668 species between the left and right branch of every ancestral node was used as a
669 quantitative indicator of a gene duplication in the last common ancestor of those species.
670 Putative duplication nodes were colored in the gene tree visualization, and their size was
671 made proportional to this indicator, facilitating our manual review.

672 A visualization of the species tree together with all gene predictions in each genome was
673 produced using the *tree drawer* utility available in Selenoprofiles, customized as follows.
674 First, each gene prediction corresponding to a multi-member family were assigned to a
675 mammalian-centric homology group, hereafter referred to as "subfamily" (e.g. GPX1,
676 GPX2, GPX3, etc), using the *orthology* utility of Selenoprofiles. Then, we encoded the
677 expected selenoproteome for every lineage based on literature; e.g. bats, since they are
678 mammals with no further lineage-specific events reported, were expected to carry one
679 Sec-encoding gene for each of GPX1, GPX2, GPX3, GPX4 and GPX6. In our custom *tree*
680 *drawer* visualization (Supplementary Figure S1), we marked each gene prediction based
681 on the deviation from such expectations; e.g., bats were labelled as missing a Sec-GPX6.
682 This highlighted the presence of extra genes (putative duplications) and missing genes
683 (putative losses). Ultimately, the joint analysis of gene and species trees led us to map
684 each event into the species tree, following criteria of maximum parsimony and
685 consistency with the known phylogenetic relationship among species.

686 Lastly, each event was critically evaluated to exclude potentially artifacts. We required
687 support by at least two sister species for most events. Single-species events were
688 deemed reliable only if supporting independent data (e.g. RNA sequences) existed. Due

689 to imperfect quality of genome assemblies, we sought confirmation of each individual
690 gene loss using independent sequence data from the lineage of interest, as follows. First,
691 we inspected Selenoprofiles predictions in an extended set of NCBI genomes. If the loss
692 was not dismissed, we then selected the protein sequence of the putatively lost gene
693 from a closely related organism, and searched it using TBLASTN at the NCBI web server
694 (Camacho et al. 2009) in all available RNA data, i.e. transcriptome shotgun assemblies
695 (TSAs), Expressed Sequence Tags (ESTs), and Reference RNA sequences (refseq_rna).

696 **Grx-domain analysis in TxnRd genes**

697 To investigate the distribution of glutaredoxin (Grx) domains in thioredoxin reductase
698 (TXNRD) genes, we first collected a set of representative TXNRD protein sequences and
699 scanned them for conserved domains using InterProScan (Jones et al. 2014). We
700 extracted the sequences containing the Grx domain and thus manually curated a high-
701 confidence alignment of Grx domains of TXNRDs. This was used as sequence profile to
702 search for homologous matches in Ensembl vertebrate genomes using Selenoprofiles.
703 Next, we cross-analyzed this data with those TXNRD gene predictions in the same
704 genomes classified as either TXNRD1 or TXNRD3 by the Selenoprofiles lineage utility. We
705 used a custom script based on pyranges v1 to assign Grx domain predictions to TXNRD1
706 and TXNRD3 genes, if located within their extended genomic coordinates (±5,000
707 nucleotides). Unassigned Grx domains were discarded. Finally, we visualized domain
708 architecture and evolutionary patterns using the ETE3 toolkit (Huerta-Cepas et al. 2016).

709 **Gene nomenclature**

710 We strived to assign gene names as consistent as possible with existing resources and
711 literature. To do so, we consulted the relevant publications on vertebrate selenoprotein
712 evolution as well as nomenclature (Regina et al. 2016), as well as current annotations at
713 Uniprot, Ensembl, and NCBI. The reference recommendation for gene names is to derive

714 them from the corresponding mammalian orthologs (Regina et al. 2016). The genes
715 emerged from non-mammalian duplications —many of them previously unreported—
716 were named by appending alphabetical or numerical suffixes to the source gene, e.g.
717 GPX4b, SELENOW2. To enhance readability, suffixes are written in lowercase
718 throughout this work. To avoid ambiguity, when creating new gene names we
719 systematically skipped the suffixes "a" and "1" (in favor of "b" or "2"), because these are
720 inconsistently used in the literature to refer to non-mammalian orthologs of single-copy
721 mammalian genes; e.g. SELENOW is sometimes referred to as SELENOW1 in species that
722 also possess the paralog SELENOW2.

723 DATA AVAILABILITY

724 The latest Selenoprofiles code is available at https://github.com/marco-
725 mariotti/selenoprofiles4 and documented at https://selenoprofiles4.readthedocs.io/.
726 Selenoprotein sequences used in this study split by family can be found in
727 Supplementary Data D2.

728 SUPPLEMENTARY DATA

729 ***Figure S1. Selenoproteome evolution across vertebrate species.*** *Each row represents a*
730 *species and each column corresponds to a selenoprotein family. Rectangles represent*
731 *individual "selenocysteine" Selenoprofiles predictions, color-labeled according to their*
732 *gene annotation in Ensembl: pink for correctly annotated genes, blue for misannotated*
733 *genes, grey for genes absent from the Ensembl annotation. Two graphical features are*
734 *used to indicate the departure from the expected selenoproteome described in 2012 by*
735 *Mariotti et al (coded in the Expectation_table.csv file at https://github.com/marco-*
736 *mariotti/selenoprotein_profiles v.1.2.0): yellow represent genes expected in that lineage*
737 *but undetected by Selenoprofiles; and red crosses mark extra genes compared to*
738 *expectations, therefore filtered by the Selenoprofiles lineage tool.*

Figure S2. Gene tree of the Glutathione peroxidase (GPX) family. *Subfamilies are color-labeled as follows: GPX4 in orange, GPX4b in yellow, GPX3b in purple, GPX3 in green, GPX6 in pink, GPX5 in light purple, GPX2 in beige, GPX1 in brown, and GPX1b in blue. Nodes highlighted in light blue indicate sequences from Salmoninae species. For a detailed explanation of the tree layout and annotation features, refer to Figure 2.*

Figure S3. Gene tree of the Iodothyronine deiodinase (DIO) family. *Subfamilies are color-labeled as follows: DIO1 in orange, DIO2 in yellow, DIO3 in green and DIO3b in purple. See also Figures 2b (layout explanation) and 4 (compressed tree).*

Figure S4. Iodothyronine deiodinase (DIO) evolution across Salmoninae species. *Each row represents a species and each column corresponds to a selenoprotein family. Nodes represent individual Selenoprofiles predictions and are color-coded according to their annotation in Ensembl: pink for well annotated Sec residue, blue for a misannotated Sec residue and gray for missing annotations.*

Figure S5. Gene tree of the Thioredoxin reductase (TXNRD) family. *Subfamilies are color-labeled as follows: TXNRD1 in green, TXNRD2 in orange, TXNRD3 in yellow. Nodes highlighted in light blue indicate sequences from Salmoninae species. See also Figures 2b (layout explanation) and 2f (compressed tree).*

Figure S6. Vertebrate species tree representing the presence/absence of Grx domain in TxnRd sequences. *Colors are assigned according to the TXNRD gene tree.*

Figure S7. Gene tree of the Selenoprotein J (SELENOJ) family. *Subfamilies are color-labeled as follows: SELENOJ in green and SELENOJ2 in yellow. Nodes highlighted in light blue indicate sequences from Salmoninae species. See also Figures 2b (layout explanation) and 6 (compressed tree).*

Figure S8. Selenoprotein J (SELENOJ) evolution across Actinopterygii species. *Each row represents a species and each column corresponds to a selenoprotein family. Nodes*

764 *represent individual Selenoprofiles predictions and are color-coded according to their*
765 *amino acid content at the UGA position: green for selenocysteine-containing*
766 *sequences, red for cysteine-containing sequences, and grey for other classifications*
767 *inferred by Selenoprofiles.*

768 ***Figure S9. Gene tree of the Selenoprotein O (SELENOO) family.*** *Subfamilies are color-*
769 *labeled as follows: SELENOO in green and SELENOO2 in orange. Nodes highlighted in*
770 *brown and lightpurple indicate sequences from Tetraodontidae and Oryziinae species,*
771 *respectively. See also Figures 2b (layout explanation) and 6 (compressed tree).*

772 ***Figure S10. Selenoprotein O (SELENOO) evolution across Actinopterygii species.*** *Each*
773 *row represents a species and each column corresponds to a selenoprotein family.*
774 *Nodes represent individual Selenoprofiles predictions and are color-coded according*
775 *to their amino acid content at the UGA position: green for selenocysteine-containing*
776 *sequences, red for cysteine-containing sequences, and grey for other classifications*
777 *inferred by Selenoprofiles.*

778 ***Figure S11. Gene tree of the Methionine sulfoxide reductase B (MSRB) family.*** *See also*
779 *Figures 2b (layout explanation) and 6 (compressed tree).*

780 ***Figure S12. Methionine sulfoxide reductase B (MSRB) evolution across Actinopterygii***
781 ***species.*** *Each row represents a species and each column corresponds to a*
782 *selenoprotein family. Nodes represent individual Selenoprofiles predictions and are*
783 *color-coded according to their amino acid content at the UGA position: green for*
784 *selenocysteine-containing sequences, red for cysteine-containing sequences, and grey*
785 *for other classifications inferred by Selenoprofiles.*

786 ***Figure S13. Gene tree of the Selenoprotein T (SELENOT) family.*** *Subfamilies are color-*
787 *labeled as follows: SELENOT1 in green and SELENOT2 in yellow. Nodes highlighted in*
788 *light blue indicate sequences from Salmoninae species, while lightpink and grey*

indicate sequences from Cyprinoidei and Otophysi species, respectively. See also Figures 2b (layout explanation) and 7 (compressed tree).

Figure S14. Gene tree of the Selenoprotein U (SELENOU) family. *Subfamilies are color-labeled as follows: SELENOU1a in yellow, SELENOU1a2 in orange, SELENOU1b in green, SELENO1c in purple and Cysteine-homolog PRLX2A in beige. Nodes highlighted in light blue indicate sequences from Salmoninae species, while lightpink indicate sequences from Cyprinoidei. See also Figures 2b (layout explanation) and 7 (compressed tree).*

Figure S15. Gene tree of the Selenoprotein V/W (SELENOV/W) family. *Subfamilies are color-labeled as follows: SELENOW1 from Actinopterygii species in green, SELENOW1 from tetrapoda species in beige, SELENOV in purple, SELENOW2a in yellow and SELENOW2c in orange. Nodes highlighted in light blue indicate sequences from Salmoninae species, while lightpink and lightgrey indicate sequences from Cyprinoidei and Cyprinodontoidei species, respectively. See also Figures 2b (layout explanation) and 7 (compressed tree).*

Figure S16. Gene tree of the Selenoprotein E (SELENOE) family. *See also Figures 2b (layout explanation) and 8 (compressed tree).*

Figure S17. Gene tree of the Selenoprotein F (SELENOF) family. *Nodes highlighted in light blue indicate sequences from Salmoninae species See also Figures 2b (layout explanation) and 8 (compressed tree).*

Figure S18. Gene tree of the Selenoprotein H (SELENOH) family. *Nodes highlighted in light blue indicate sequences from Salmoninae species. See also Figures 2b (layout explanation) and 8 (compressed tree).*

Figure S19. Gene tree of the Selenoprotein P (SELENOP) family. *Nodes highlighted in light blue indicate sequences from Salmoninae species. See also Figures 2b (layout explanation) and 8 (compressed tree).*

814 ***Figure S20. Gene tree of the SEPHS family.*** *See also Figure 2b (layout explanation).*

815 ***Table T1. Comparison of Glutathione Peroxidase (GPX) gene nomenclature between***
816 ***Xue et al. and this study.*** *The table lists the gene names assigned to GPX family*
817 *members in Xue et al. and the corresponding nomenclature used in this study (Ticó et*
818 *al.). The third column provides relevant notes on differences in naming, such as updated*
819 *classification, subfamily assignment, or gene duplication interpretations.*

820 ***Document D1. Sequence analysis of SELENOP2 in lampreys****. The document contains the*
821 *sequences of SELENOP2 genes in two Cyclostomata, obtained via Selenoprofiles in*
822 *genome data, then corroborated by RNA sequences. SECIS elements were predicted*
823 *using SECISearch3.*

824 ***Document D2. Protein sequences of all selenoproteins predicted and analyzed in this***
825 ***study.*** *For details on file structure and content, please refer to the accompanying*
826 *README file.*

827 FUNDING

828 MM is funded by grants RYC2019-027746-I, PID2020-115122GA-I00, PID2023-
829 147164NB-I00 funded by MICIU/AEI /10.13039/501100011033 and by "ESF Investing in
830 your future", FEDER, UE.

831 BIBLIOGRAPHY

832 Anouar Y, Lihrmann I, Falluel-Morel A, Boukhzar L. 2018. Selenoprotein T is a key player
833 in ER proteostasis, endocrine homeostasis and neuroprotection. *Free Radic Biol*
834 *Med* 127:145–152.

835 Arnér ESJ. 2009. Focus on mammalian thioredoxin reductases--important
836 selenoproteins with versatile functions. *Biochim Biophys Acta* [Internet] 1790:495–
837 526. Available from: https://pubmed.ncbi.nlm.nih.gov/19364476/

838 Baclaocos J, Santesmasses D, Mariotti M, Bierła K, Vetick MB, Lynch S, McAllen R, Mackrill
839 JJ, Loughran G, Guigó R, et al. 2019. Processive Recoding and Metazoan Evolution of
840 Selenoprotein P: Up to 132 UGAs in Molluscs. *J Mol Biol* [Internet] 431:4381–4407.
841 Available from: https://pubmed.ncbi.nlm.nih.gov/31442478/

842 Bertz M, Kühn K, Koeberle SC, Müller MF, Hoelzer D, Thies K, Deubel S, Thierbach R, Kipp
843 AP. 2018. Selenoprotein H controls cell cycle progression and proliferation of human
844 colorectal cancer cells. *Free Radic Biol Med* [Internet] 127:98–107. Available from:
845 https://pubmed.ncbi.nlm.nih.gov/29330096/

846 Botero-Castro F, Figuet E, Tilak MK, Nabholz B, Galtier N. 2017. Avian Genomes Revisited:
847 Hidden Genes Uncovered and the Rates versus Traits Paradox in Birds. *Mol Biol Evol*
848 [Internet] 34:3123–3131. Available from:
849 https://pubmed.ncbi.nlm.nih.gov/28962031/

850 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009.
851 BLAST+: architecture and applications. *BMC Bioinformatics* [Internet] 10. Available
852 from: https://pubmed.ncbi.nlm.nih.gov/20003500/

853 Castellano S, Andrés AM, Bosch E, Bayes M, Guigó R, Clark AG. 2009. Low
854 exchangeability of selenocysteine, the 21st amino acid, in vertebrate proteins. *Mol*
855 *Biol Evol* [Internet] 26:2031–2040. Available from:
856 https://pubmed.ncbi.nlm.nih.gov/19487332/

857 Castellano S, Lobanov A V., Chapple C, Novoselov S V., Albrecht M, Hua D, Lescure A,
858 Lengauer T, Krol A, Gladyshev VH, et al. 2005. Diversity and functional plasticity of
859 eukaryotic selenoproteins: Identification and characterization of the SelJ family. *Proc*
860 *Natl Acad Sci U S A* [Internet] 102:16188–16193. Available from:
861 https://www.pnas.org/doi/abs/10.1073/pnas.0505146102

862 Castellano S, Novoselov S V., Kryukov G V., Lescure A, Blanco E, Krol A, Gladyshev VN,
863 Guigó R. 2004. Reconsidering the evolution of eukaryotic selenoproteins: a novel
864 nonmammalian family with scattered phylogenetic distribution. *EMBO Rep*
865 [Internet] 5:71–77. Available from: https://pubmed.ncbi.nlm.nih.gov/14710190/

866 Chernorudskiy A, Varone E, Colombo SF, Fumagalli S, Cagnotto A, Cattaneo A, Briens M,
867 Baltzinger M, Kuhn L, Bachi A, et al. 2020. Selenoprotein N is an endoplasmic
868 reticulum calcium sensor that links luminal calcium levels to a redox activity. *Proc*
869 *Natl Acad Sci U S A* [Internet] 117:21288–21298. Available from:
870 https://pubmed.ncbi.nlm.nih.gov/32817544/

871 Copeland PR, Howard MT. 2021. Ribosome Fate during Decoding of UGA-Sec Codons.
872 *Int J Mol Sci* [Internet] 22. Available from:
873 https://pubmed.ncbi.nlm.nih.gov/34948001/

874 Cox AG, Tsomides A, Kim AJ, Saunders D, Hwang KL, Evason KJ, Heidel J, Brown KK, Yuan
875 M, Lien EC, et al. 2016. Selenoprotein H is an essential regulator of redox
876 homeostasis that cooperates with p53 in development and tumorigenesis. *Proc Natl*
877 *Acad Sci U S A* [Internet] 113:E5562–E5571. Available from:
878 https://pubmed.ncbi.nlm.nih.gov/27588899/

879 Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Austine-
880 Orimoloye O, Azov AG, Barnes I, Bennett R, et al. 2022. Ensembl 2022. *Nucleic Acids*
881 *Res* [Internet] 50:D988–D995. Available from:
882 https://pubmed.ncbi.nlm.nih.gov/34791404/

883 Darras VM, Van Herck SLJ. 2012. Iodothyronine deiodinase structure and function: from
884 ascidians to humans. *J Endocrinol* [Internet] 215:189–206. Available from:
885 https://pubmed.ncbi.nlm.nih.gov/22825922/

886 Davy T, Castellano S. 2018. The genomics of selenium: Its past, present and future.
887 *Biochim Biophys Acta Gen Subj* [Internet] 1862:2427–2432. Available from:
888 https://pubmed.ncbi.nlm.nih.gov/29859258/

889 Fairweather-Tait SJ, Bao Y, Broadley MR, Collings R, Ford D, Hesketh JE, Hurst R. 2011.
890 Selenium in human health and disease. *https://home.liebertpub.com/ars* [Internet]
891 14:1337–1383. Available from:
892 https://www.liebertpub.com/doi/10.1089/ars.2010.3275

893 Fata F, Gencheva R, Cheng Q, Lullo R, Ardini M, Silvestri I, Gabriele F, Ippoliti R, Bulman
894 CA, Sakanari JA, et al. 2022. Biochemical and structural characterizations of
895 thioredoxin reductase selenoproteins of the parasitic filarial nematodes Brugia
896 malayi and Onchocerca volvulus. *Redox Biol* [Internet] 51. Available from:
897 https://pubmed.ncbi.nlm.nih.gov/35276442/

898 Flohé L. 2009. The labour pains of biochemical selenology: the history of selenoprotein
899 biosynthesis. *Biochim Biophys Acta* 1790:1389–1403.

900 Flowers B, Bochnacka O, Poles A, Diamond AM, Kastrati I. 2023. Distinct Roles of
901 SELENOF in Different Human Cancers. *Biomolecules* [Internet] 13. Available from:
902 https://pubmed.ncbi.nlm.nih.gov/36979420/

903 Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New
904 Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing
905 the Performance of PhyML 3.0. *Syst Biol* [Internet] 59:307–321. Available from:
906 https://dx.doi.org/10.1093/sysbio/syq010

907 Howard MT, Copeland PR. 2019. New Directions for Understanding the Codon
908 Redefinition Required for Selenocysteine Incorporation. *Biol Trace Elem Res*
909 [Internet] 192:18–25. Available from: https://pubmed.ncbi.nlm.nih.gov/31342342/

910 Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and Visualization
911 of Phylogenomic Data. *Mol Biol Evol* [Internet] 33:1635. Available from:
912 https://pmc.ncbi.nlm.nih.gov/articles/PMC4868116/

913 Hughes LC, Ortí G, Huang Y, Sun Y, Baldwin CC, Thompson AW, Arcila D, Betancur R, Li
914 C, Becker L, et al. 2018. Comprehensive phylogeny of ray-finned fishes
915 (Actinopterygii) based on transcriptomic and genomic data. *Proc Natl Acad Sci U S*
916 *A* [Internet] 115:6249–6254. Available from:
917 https://www.pnas.org/doi/abs/10.1073/pnas.1719358115

918 Jeong D won, Kim TS, Chung YW, Lee BJ, Kim IY. 2002. Selenoprotein W is a glutathione-
919 dependent antioxidant in vivo. *FEBS Lett* [Internet] 517:225–228. Available from:
920 https://pubmed.ncbi.nlm.nih.gov/12062442/

921 Jiang L, Liu Q, Ni J. 2010. In silico identification of the sea squirt selenoproteome. *BMC*
922 *Genomics* [Internet] 11:289. Available from: http://www.biomedcentral.com/1471-
923 2164/11/289

924 Jiang L, Ni J, Liu Q. 2012. Evolution of selenoproteins in the metazoan. *BMC Genomics*
925 [Internet] 13:446. Available from:
926 http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-13-446

927 Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell
928 A, Nuka G, et al. 2014. InterProScan 5: genome-scale protein function classification.
929 *Bioinformatics* [Internet] 30:1236–1240. Available from:
930 https://dx.doi.org/10.1093/bioinformatics/btu031

931 Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome*
932 *Res* [Internet] 20:1313–1326. Available from:
933 https://pubmed.ncbi.nlm.nih.gov/20651121/

934 Kryukov G V., Castellano S, Novoselov S V., Lobanov A V., Zehtab O, Guigó R, Gladyshev
935 VN. 2003. Characterization of mammalian selenoproteomes. *Science* [Internet]
936 300:1439–1443. Available from: https://pubmed.ncbi.nlm.nih.gov/12775843/

937 Labunskyy VM, Hatfield DL, Gladyshev VN. 2014. Selenoproteins: Molecular Pathways
938 and Physiological Roles. *Physiol Rev* [Internet] 94:739. Available from:
939 https://pmc.ncbi.nlm.nih.gov/articles/PMC4101630/

940 Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, Durbin R,
941 Edwards S V., Forest F, Gilbert MTP, et al. 2018. Earth BioGenome Project:
942 Sequencing life for the future of life. *Proc Natl Acad Sci U S A* [Internet] 115:4325–
943 4333. Available from: https://pubmed.ncbi.nlm.nih.gov/29686065/

944 Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, Hvidsten TR, Leong JS, Minkley
945 DR, Zimin A, et al. 2016. The Atlantic salmon genome provides insights into
946 rediploidization. *Nature 2016 533:7602* [Internet] 533:200–205. Available from:
947 https://www.nature.com/articles/nature17164

948 Lobanov A V., Fomenko DE, Zhang Y, Sengupta A, Hatfield DL, Gladyshev VN. 2007.
949 Evolutionary dynamics of eukaryotic selenoproteomes: Large selenoproteomes
950 may associate with aquatic life and small with terrestrial life. *Genome Biol* [Internet]
951 8:1–16. Available from:
952 https://genomebiology.biomedcentral.com/articles/10.1186/gb-2007-8-9-r198

953 Lorgen M, Casadei E, Król E, Douglas A, Birnie MJ, Ebbesson LOE, Nilsen TO, Jordan WC,
954 Jørgensen EH, Dardente H, et al. 2015. Functional divergence of type 2 deiodinase
955 paralogs in the Atlantic salmon. *Curr Biol* [Internet] 25:936–941. Available from:
956 https://pubmed.ncbi.nlm.nih.gov/25802152/

957 Manta B, Makarova NE, Mariotti M. 2022. The selenophosphate synthetase family: A
958 review. *Free Radic Biol Med* 192:63–76.

959 Mariotti M, Gladyshev VN. 2022. Selenocysteine-containing proteins. *Redox Chemistry*
960 *and Biology of Thiols*:405–421.

961 Mariotti M, Guigó R. 2010. Selenoprofiles: profile-based scanning of eukaryotic genome
962 sequences for selenoprotein genes. *Bioinformatics* [Internet] 26:2656–2663.
963 Available from: https://dx.doi.org/10.1093/bioinformatics/btq516

964 Mariotti M, Ridge PG, Zhang Y, Lobanov A V., Pringle TH, Guigo R, Hatfield DL, Gladyshev
965 VN. 2012. Composition and evolution of the vertebrate and mammalian
966 selenoproteomes. *PLoS One* [Internet] 7. Available from:
967 https://pubmed.ncbi.nlm.nih.gov/22479358/

968 Mariotti M, Santesmasses D, Capella-Gutierrez S, Mateo A, Arnan C, Johnson R, D'Aniello
969 S, Yim SH, Gladyshev VN, Serras F, et al. 2015. Evolution of selenophosphate
970 synthetases: emergence and relocation of function through independent
971 duplications and recurrent subfunctionalization. *Genome Res* [Internet] 25:1256–
972 1267. Available from: http://genome.cshlp.org/content/25/9/1256.full

973 Maroney MJ, Hondal RJ. 2018. Selenium versus sulfur: Reversibility of chemical reactions
974 and resistance to permanent oxidation in proteins and nucleic acids. *Free Radic Biol*
975 *Med* [Internet] 127:228–237. Available from:
976 https://pubmed.ncbi.nlm.nih.gov/29588180/

977 Mc Cartney AM, Formenti G, Mouton A, De Panis D, Marins LS, Leitão HG, Diedericks G,
978 Kirangwa J, Morselli M, Salces-Ortiz J, et al. 2024. The European Reference Genome
979 Atlas: piloting a decentralised approach to equitable biodiversity genomics. *npj*
980 *Biodiversity 2024 3:1* [Internet] 3:1–17. Available from:
981 https://www.nature.com/articles/s44185-024-00054-6

982 Melo LMN, Sabatier M, Ramesh V, Szylo KJ, Fraser CS, Pon A, Mitchell EC, Servage KA,
983 Allies G, Westedt I V., et al. 2025. Selenoprotein O Promotes Melanoma Metastasis

984 and Regulates Mitochondrial Complex II Activity. *Cancer Res* [Internet] 85:942–955.
985 Available from: https://pubmed.ncbi.nlm.nih.gov/39700395/

986 Misra S, Lee TJ, Sebastian A, McGuigan J, Liao C, Koo I, Patterson AD, Rossi RM, Hall MA,
987 Albert I, et al. 2023. Loss of selenoprotein W in murine macrophages alters the
988 hierarchy of selenoprotein expression, redox tone, and mitochondrial functions
989 during inflammation. *Redox Biol* [Internet] 59. Available from:
990 https://pubmed.ncbi.nlm.nih.gov/36516721/

991 Novoselov S V., Hua D, Lobanov A V., Gladyshev VN. 2006. Identification and
992 characterization of Fep15, a new selenocysteine-containing member of the Sep15
993 protein family. *Biochem J* [Internet] 394:575–579. Available from:
994 https://pubmed.ncbi.nlm.nih.gov/16236027/

995 Odunsi A, Kapitonova MA, Woodward G, Rahmani E, Ghelichkhani F, Liu J, Rozovsky S.
996 2025. Selenoprotein K at the intersection of cellular pathways. *Arch Biochem*
997 *Biophys* [Internet] 764. Available from: https://pubmed.ncbi.nlm.nih.gov/39571956/

998 Powell S, Forslund K, Szklarczyk D, Trachana K, Roth A, Huerta-Cepas J, Gabaldón T,
999 Rattei T, Creevey C, Kuhn M, et al. 2014. eggNOG v4.0: nested orthology inference
1000 across 3686 organisms. *Nucleic Acids Res* [Internet] 42:D231–D239. Available from:
1001 https://dx.doi.org/10.1093/nar/gkt1253

1002 Regina BF, Gladyshev VN, Arnér ES, Berry MJ, Bruford EA, Burk RF, Carlson BA, Castellano
1003 S, Chavatte L, Conrad M, et al. 2016. Selenoprotein gene nomenclature. *Journal of*
1004 *Biological Chemistry* [Internet] 291:24036–24040. Available from:
1005 https://www.jbc.org/action/showFullText?pii=S0021925820346731

1006 Reich HJ, Hondal RJ. 2016. Why Nature Chose Selenium. *ACS Chem Biol* [Internet] 11:821–
1007 841. Available from: https://pubmed.ncbi.nlm.nih.gov/26949981/

1008 Santesmasses D, Mariotti M, Gladyshev VN. 2020. Bioinformatics of Selenoproteins.
1009 *Antioxid Redox Signal* [Internet] 33:525–536. Available from:
1010 https://pubmed.ncbi.nlm.nih.gov/32031018/

1011 Santesmasses D, Mariotti M, Guigó R. 2018. Selenoprofiles: A Computational Pipeline for
1012 Annotation of Selenoproteins. *Methods in Molecular Biology* [Internet] 1661:17–28.
1013 Available from: https://link.springer.com/protocol/10.1007/978-1-4939-7258-6_2

1014 Sarangi GK, Romagne F, Castellano S. 2018. Distinct Patterns of Selection in Selenium-
1015 Dependent Genes between Land and Aquatic Vertebrates. *Mol Biol Evol* [Internet]
1016 35:1744–1756. Available from: https://dx.doi.org/10.1093/molbev/msy070

1017 Shchedrina VA, Novoselov S V., Malinouski MY, Gladyshev VN. 2007. Identification and
1018 characterization of a selenoprotein family containing a diselenide bond in a redox
1019 motif. *Proc Natl Acad Sci U S A* [Internet] 104:13919–13924. Available from:
1020 https://www.pnas.org/doi/abs/10.1073/pnas.0703448104

1021 Shi Z, Han Z, Chen J, Zhou JC. 2024. Endoplasmic reticulum-resident selenoproteins and
1022 their roles in glucose and lipid metabolic disorders. *Biochim Biophys Acta Mol Basis*
1023 *Dis* [Internet] 1870. Available from: https://pubmed.ncbi.nlm.nih.gov/38763408/

1024 Sreelatha A, Yee SS, Lopez VA, Park BC, Kinch LN, Pilch S, Servage KA, Zhang J, Jiou J,
1025 Karasiewicz-Urbańska M, et al. 2018. Protein AMPylation by an Evolutionarily
1026 Conserved Pseudokinase. *Cell* [Internet] 175:809-821.e19. Available from:
1027 https://pubmed.ncbi.nlm.nih.gov/30270044/

1028 Tarrago L, Kaya A, Kim HY, Manta B, Lee BC, Gladyshev VN. 2022. The selenoprotein
1029 methionine sulfoxide reductase B1 (MSRB1). *Free Radic Biol Med* [Internet] 191:228–
1030 240. Available from: https://pubmed.ncbi.nlm.nih.gov/36084791/

1031 Ticó M, Sullivan E, Guigó R, Mariotti M. 2024. Overcoming the widespread flaws in the
1032 annotation of vertebrate selenoprotein genes in public databases. *bioRxiv*

1033 [Internet]:2024.10.30.620813. Available from:
1034 https://doi.org/10.1101/2024.10.30.620813

1035 Toppo S, Vanin S, Bosello V, Tosatto SCE. 2008. Evolutionary and structural insights into
1036 the multifaceted glutathione peroxidase (Gpx) superfamily. *Antioxid Redox Signal*
1037 [Internet] 10:1501–1513. Available from:
1038 https://pubmed.ncbi.nlm.nih.gov/18498225/

1039 Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara
1040 GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates.
1041 *Genome Res* [Internet] 19:327–335. Available from:
1042 http://genome.cshlp.org/content/19/2/327.full

1043 Wu RTY, Cao L, Chen BPC, Cheng WH. 2014. Selenoprotein H suppresses cellular
1044 senescence through genome maintenance and redox regulation. *J Biol Chem*
1045 [Internet] 289:34378–34388. Available from:
1046 https://pubmed.ncbi.nlm.nih.gov/25336634/

1047 Xue Y, Chen L, Li B, Xiao J, Wang H, Dong C, Li X, Xu P. 2022. Genome-wide mining of
1048 gpx gene family provides new insights into cadmium stress responses in common
1049 carp (Cyprinus carpio). *Gene* 821:146291.

1050 Yim SH, Everley RA, Schildberg FA, Lee SG, Orsi A, Barbati ZR, Karatepe K, Fomenko DE,
1051 Tsuji PA, Luo HR, et al. 2018. Role of Selenof as a Gatekeeper of Secreted Disulfide-
1052 Rich Glycoproteins. *Cell Rep* [Internet] 23:1387–1398. Available from:
1053 https://pubmed.ncbi.nlm.nih.gov/29719252/

1054 Zhang X, Xiong W, Chen LL, Huang JQ, Lei XG. 2020. Selenoprotein V protects against
1055 endoplasmic reticulum stress and oxidative injury induced by pro-oxidants. *Free*
1056 *Radic Biol Med* [Internet] 160:670–679. Available from:
1057 https://pubmed.ncbi.nlm.nih.gov/32846216/

Zhang Y, Jin J, Huang B, Ying H, He J, Jiang L. 2022. Selenium Metabolism and Selenoproteins in Prokaryotes: A Bioinformatics Perspective. *Biomolecules 2022, Vol. 12, Page 917* [Internet] 12:917. Available from: https://www.mdpi.com/2218-273X/12/7/917/htm

Zigrossi A, Hong LK, Ekyalongo RC, Cruz-Alvarez C, Gornick E, Diamond AM, Kastrati I. 2022. SELENOF is a new tumor suppressor in breast cancer. *Oncogene 2021 41:9* [Internet] 41:1263–1268. Available from: https://www.nature.com/articles/s41388-021-02158-w