MANUSCRIPT TITLE

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

AUTHORS

Max Ticó[1], Jesus Lozano-Fernandez[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 a 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 (56 selenoproteins), losses (50), and Sec-to-cysteine (Cys) conversions (21). 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 Salmonidae and Cyprinoidei fish with 56 and 44 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.

KEYWORDS

Selenocysteine, selenoprotein, selenium, gene family evolution, stop codon readthrough, recoding

BACKGROUND

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) [1–3]. 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 the SECIS (selenocysteine insertion sequence) [4, 5]. 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 [6].

Since their discovery in the 1970s, Sec and selenoproteins have been extensively studied across diverse fields of biology [7, 8]. 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 [3]. As a result, selenoproteins are investigated in a broad range of species, from standard model organisms (e.g., *Mus musculus, Danio rerio*) to agriculturally relevant animals (e.g., *Bos taurus, Gallus gallus, Salmo salar*). 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 murine selenoproteomes were first described in 2003 as comprising 25 and 24 Sec-encoding genes, respectively [9]. Subsequent studies expanded the known selenoproteomes of selected vertebrates and metazoans [10–15]. In our seminal paper in 2012, we provided the first description of selenoproteome composition and evolution across diverse vertebrates [16]. 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 [17, 18]. 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

76 selenoprotein genes have substantial annotation errors in Ensembl and NCBI genomes,
77 respectively [19]. This also led to unreliable selenoprotein data in evolutionary resources
78 dedicated to orthology (e.g. Ensembl Compara [20], EggNOG [21]).

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

85 RESULTS

86 We downloaded 314 genome assemblies from Ensembl v.108 representative of all major
87 lineages of vertebrates, complemented by 26 NCBI genomes from early vertebrates and
88 outgroups (Methods). Next, we used the software Selenoprofiles (Methods) to identify
89 homologs belonging to any of the known metazoan selenoprotein families in those
90 genomes, resulting in 13,517 gene predictions. Phylogenetic reconstruction was then
91 performed for each family, and gene and species trees were manually inspected to infer
92 gene evolutionary events (Methods). These comprise gene duplications, gene losses, and
93 the selenoprotein-specific event of the conversion of a Sec residue to Cys. In
94 selenoprotein evolution, Cys to Sec events are also possible, but have been observed
95 only along evolutionary distances much larger than encompassed by vertebrates [22,
96 23]. Each branch of gene trees was manually annotated as a separate orthologous group,
97 assigning gene names as consistent as possible with existing resources and literature.
98 Tree inspection and orthology assignments were performed taking into account branch
99 support values obtained by UltraFast Bootstrap (UFB). The great majority of orthologous
100 groups were supported by high UFB support values in gene trees (Methods); all

101 exceptions are explicitly mentioned in their relevant section. Each putative event was
102 critically assessed, following criteria of maximum parsimony, and seeking consistency
103 with the species tree and support by independent data (e.g. RNA sequences, other
104 genome assemblies from NCBI) (Methods).

105 Thanks to the vast diversity of organisms in our dataset, the great majority of events were
106 shared among multiple species, leading to phylogenetically coherent patterns of gene
107 occurrence that allowed reliable evolutionary inferences. Yet, two lineages stood out for
108 their divergence from this general trend. First, bird genomes exhibited many apparent
109 gene losses scattered across lineages, in a pattern that would imply dozens of
110 independent events (Supplementary Figure S1). This included the apparent absence of
111 *SEPHS2*, a gene that nevertheless is required for biosynthesis of all other selenoproteins
112 [22], and of *SELENOW*. We ascribed this pattern to a known bias of bird assemblies: they
113 systematically lack certain GC-rich regions that present technical challenges for
114 sequencing [24]. To further support this interpretation, we examined the syntenic block
115 surrounding *SEPHS2*, which includes the essential genes *G6PD* and *IKBKG* (Figure S2).
116 This genomic region is conserved across Sauria; yet, we observed the apparent absence
117 of all three genes in most (but not all) Aves. In contrast, outgroup lineages such as
118 Testudines retain the complete syntenic block, including the three genes. Strikingly, we
119 found RNA evidence of expression for both genes (*SEPHS2* and *SELENOW*) in multiple
120 species despite their apparent absence from their available genome assemblies,
121 demonstrating that these genes are present in the actual genomes of birds. Ultimately,
122 we dismissed all selenoprotein losses in Sauria (which includes birds) as artifactual.

123 Second, the fish lineage of Cyprinoidei (e.g. carps, *Danio rerio*) exhibited a massive
124 amount of apparent gene duplications (Supplementary Figure S1), which we could not
125 precisely disentangle by phylogenetic analysis. We realized that the sublineage of

126 Cyprininae (e.g. carps) reportedly underwent extensive recent whole genome
127 duplications and polyploidy, which greatly complicated this task. Therefore, we decided
128 to limit our analysis to the evolutionary events that occurred at the root of Cyprinoidei,
129 i.e. shared between *D. rerio* and Cyprininae, and dismiss the rest.

130 Several of the outgroups we used have a characterized selenoproteome: *Ciona*
131 *intestinalis* [15], *Branchiostoma floridae* [14], *Drosophila melanogaster* [25], *Strigamia*
132 *maritima* [26, 27]. This allowed us to critically assess their gene predictions, and compare
133 them with those of vertebrates to phase evolutionary events, as detailed later for each
134 family.

135 **Overview of the vertebrate selenoproteome**

136 We identified selenoproteins belonging to 19 protein families. The most represented
137 families were glutathione peroxidases with 1,529 Sec-encoding genes, iodothyronine
138 deiodinases with 988, and thioredoxin reductases with 718. In general, Actinopterygii
139 (ray-finned fishes) exhibited richer selenoproteomes (35-56 selenoprotein genes per
140 genome) than Tetrapoda (24-25 genes per genome) (Figure 1).

141 While our analyses recapitulated earlier reports performed at a smaller scale, we also
142 detected many novel events, particularly gene duplications. The majority affected
143 Actinopterygii and its sublineages, whose selenoproteome appeared not only larger, but
144 also more dynamic. In total, we detected 56 gene duplications (37 in Actinopterygii), and
145 50 gene losses (29 in Actinopterygii). The most selenoprotein-rich organisms were
146 Salmonidae (aka salmonids, e.g. trouts, salmons) with 56 selenoprotein genes, 17 of
147 which emerged by duplication specifically in this lineage, and Cyprinoidei with 44
148 selenoproteins, 6 of which lineage-specific. On the other hand, Sec-to-Cys conversions
149 were more prevalent in Tetrapoda: we detected 21 of these events overall, 16 of which

150 were in Tetrapoda. Hereafter, we describe the selenoproteome evolution for each family
151 separately.

152 Glutathione peroxidases (GPX)

153 Glutathione peroxidases (GPX) are key antioxidant enzymes that protect cells from
154 oxidative damage by reducing hydrogen peroxide and lipid peroxides using glutathione
155 as an electron donor [16, 28]. Mammals possess eight *GPX* homologs with specialized
156 tissue distribution and functions, including five selenoproteins: *GPX1* (ubiquitous), *GPX2*
157 (gastrointestinal), *GPX3* (extracellular, kidney), *GPX4* (membrane-associated,
158 ubiquitous), *GPX6* (olfactory system); and three Cys paralogs: *GPX5*, *GPX7*, and *GPX8*.
159 Our phylogenetic analysis (Figures 2 and 3, Supplementary Figure S3) confirms a well-
160 supported evolutionary classification of GPX proteins, identifying three major groups:
161 GPX1/GPX2, GPX3/GPX5/GPX6, and GPX4/GPX7/GPX8. These findings align with
162 previous studies on GPX evolution [16, 28, 29].

163 *GPX1, GPX2, GPX3, GPX4* are present in all vertebrates, with a single notable exception.
164 In the GPX1/GPX2 group, we observed that the Hyperoartia class of jawless fish
165 (Cyclostomata superclass) possessed two similar genes which phylogenetic
166 reconstruction placed as outgroups of the GPX1 and GPX2 clusters of other vertebrates
167 (Figure 3a). We then noticed that the other class of Cyclostomata, Myxini, has instead a
168 single orthologous gene (Supplementary Figure S1). We infer that GPX1 and GPX2 split in
169 Gnathostomata (jawed vertebrates) and that Cyclostomata likely reflect the ancestral
170 status, so that we named this gene GPX1/2 and its Hyperoartia-specific paralog GPX1/2b.
171 Although the UFB support of the corresponding node in the reconstructed protein tree is
172 only 25%, we argue that we can confidently map the origin of GPX1/2b to Hyperoartia
173 based on the species-tree analysis.

174 *GPX6* originated at the root of placentals and exhibited a very dynamic evolution.
175 Multiple independent *GPX6* Sec-to-Cys conversions had been reported [16], which we
176 confirmed and expanded in our analyses (Figures 1, 2b, 2c, 3c; Supplementary Figure S3).
177 We observed Cys conversions in multiple taxonomic groups, including Platyrrhini
178 primates (New World monkeys, e.g. *Callithrix jacchus*, *Cebus imitator*), Muroidea
179 rodents (e.g. *Mus musculus*, *Rattus norvegicus*, *Mesocricetus auratus*), Lagomorpha
180 (e.g. *Oryctolagus cuniculus*, *Ochotona princeps*), Feliformia (e.g. *Felis catus*, *Panthera*
181 *pardus*), Bathyergidae rodents (e.g. *Fukomys damarensis*). In addition, we also detected
182 species-specific conversions, namely in *Marmota marmota marmota, Sciurus vulgaris,*
183 *Zalophus californianus and Castor canadensis* (Figure 3c). Furthermore, *GPX6* was lost
184 altogether in Chiroptera (bats).

185 We identified several GPX gene duplications across vertebrates. We confirmed two
186 previously reported GPX gene duplications at the root of Actinopterygii leading to the
187 emergence of *GPX4b* and *GPX1b* (the latter with only 20% UFB, but consistent
188 occurrences along the species tree). These genes were found in all investigated species
189 within this lineage, except for Oryziinae fish (e.g. *Oryzias latipes*), which we infer to have
190 lost *GPX1b*. In two Oryziinae species, we initially identified a putative Cys-containing
191 *GPX4b* paralog, yet we dismissed it due to lack of sequence support from RNA data. A
192 third gene, *GPX3b*, was also reported in Actinopterygii; our analysis indicates that it
193 originated at the root of Vertebrata, and was subsequently converted to Cys in Myxini
194 and lost in Tetrapoda.

195 We also observed additional selenoprotein gene duplications in specific branches of
196 Actinopterygii. In Hyperoartia, we observed three lineage-specific events that yielded
197 *GPX1/2b*, *GPX3b2* and *GPX4c*. Then, a lineage-specific duplication of *GPX1* gave rise to a
198 distinct Sec-paralog, *GPX1c*, in Chondrichthyes (cartilaginous fish). We mapped three
199 additional duplications in Salmonidae yielding *GPX1b2*, *GPX3a2*, and *GPX4a2*. All these

events were supported by gene tree nodes with high UFB support values (>70%), except the aforementioned *GPX1/2b.* Notably, these duplications were previously reported in an analysis of 8 fish genomes [29]. Our study demonstrates that they are lineage-specific events, so that we have assigned new names to the duplicated genes to reflect their precise evolutionary trajectories. The correspondence with [29] is provided in Supplementary Table T1.

Iodothyronine deiodinases (DIO)

Iodothyronine deiodinases (DIO) regulate thyroid hormone metabolism by catalyzing the activation or inactivation of thyroxine (T4) and triiodothyronine (T3) [30]. 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, S4), 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* predates the vertebrate radiation and was then lost in Cyclostomata. All Actinopterygii additionally carried a *DIO3* duplicate (*DIO3b*). The placement of *DIO3b* genes in the reconstructed protein tree was mostly consistent with an origin by *DIO3* duplication at the root of Actinopterygii (UFB support of relevant node: 68%), with the exception of Cyprinoidei *DIO3b* which was placed closer to the tree root, possibly due to long branch attraction.

We also confirmed two more duplications, one of the ancestral *DIO3* gene (also called *DIO3a*) [31], to generate *DIO3a2*, and the other of *DIO2* to yield *DIO2b* (Figures 1, 4, Supplementary Figures S4, S5). Both were mapped to Salmonidae with high support, as previously reported [31]. In Hyperoartia (Cyclostomata), we identified two lineage-specific duplications resulting in *DIO2c* and *DIO3c* (absent from Myxini, the other class of Cyclostomata; Supplementary Figure S1). Both were highly supported, with the caveat

226 that *DIO2c* was placed as an outgroup of all *DIO2* genes, rather than at its expected
227 position as sister of Cyclostomata *DIO2*.

228 Thioredoxin reductases (TXNRD)

229 Thioredoxin reductases (TXNRDs) play a central role in maintaining cellular redox
230 balance by catalyzing the reduction of thioredoxins, which are essential for regulating
231 oxidative stress responses, cell proliferation, and redox-regulated signaling cascades
232 [32]. Tetrapods have three TXNRD genes: *TXNRD1* (cytosolic), *TXNRD2* (mitochondrial),
233 and *TXNRD3* (also known as *TGR*; testis-specific, containing a peculiar N-terminal Grx
234 domain), each of which participates in distinct physiological processes. In phylogenetic
235 reconstruction, *TXNRD1* and *TXNRD3* clearly cluster together, indicating a shared origin.

236 Reportedly, fishes lack the *TXNRD3/TGR* gene, but their *TXNRD1* gene encodes for a Grx-
237 containing protein as its major isoform, which duplicated to give rise to TXNRD3/TGR in
238 tetrapods [16, 33]. Yet, our present analysis (Figure 5, Supplementary Figures S6-S9) did
239 not recapitulate TXNRD evolution as previously proposed. While indeed nearly all fish
240 possess a single gene in the *TXNRD1/TXNRD3* group, we claim here that it is the ortholog
241 of *TXNRD3*, rather than *TXNRD1*. This classification is supported by both the topology of
242 reconstructed gene trees (Figure 5a, Supplementary Figure S6) and by syntenic
243 conservation (Supplementary Figures S7, S8). Furthermore, we detected both genes
244 (*TXNRD1* and *TXNRD3)* in Chondrichthyes. Because this lineage serves as an outgroup to
245 both Actinopterygii and Tetrapoda, we argue that *TXNRD1* emerged earlier than
246 tetrapods. To explain its sparse occurrence across Actinopterygii, we hypothesize that it
247 was subsequently lost independently in six Actinopterygian lineages (Figure 1). Curiously,
248 fish *TXNRD1* features an atypical GNCUG C-terminus (rather than GCUC found across
249 TXNRDs). We did not find *TXNRD1* in Cyclostomata, and thus we map the
250 *TXNRD1/TXNRD3* duplication to the root of Gnathostomata. Finally, because the

251 Cyclostomata gene clearly clusters with *TXNRD3*, we identify the *TXNRD3* clade as the
252 ancestral one and *TXNRD1* as the gene derived from its duplication.

253 It is reported that the human *TXNRD1* gene encodes for a minor transcript isoform that
254 includes the Grx domain ("TxnRd1_v3"), which is a predominant feature of *TXNRD3*
255 genes. This transcript isoform is reportedly absent in *M. musculus TXNRD1* [32].
256 Therefore, we systematically investigated the occurrence of Grx domains in the genomic
257 regions of *TXNRD1* and *TXNRD3* across all vertebrates (Methods). Our results show that
258 Grx domains are a widespread feature of *TXNRD1* genes (Figure 5b, Supplementary
259 Figure S9). However, they have been lost in several tetrapod lineages independently
260 (Rodentia, Aves, Equus, Prototheria, Metatheria and Pecora) and in all Actinopterygii
261 retaining the *TXNRD1* gene, possibly due to functional redundancy with *TXNRD3*.
262 Altogether, we argue that the Grx-domain is an ancestral feature of *TXNRD3* which was
263 initially retained in *TXNRD1*, but that was later lost in multiple lineages through
264 convergent evolution.

265 We further investigated the ancestral state of vertebrates by examining the position of
266 other metazoan groups in the TXNRD gene tree (Figure 5a), as well as an extended gene
267 tree featuring additional outgroups available in PhylomeDB [34] (QFO2020 seed,
268 *Caenorhabditis elegans* phylome ID: 791). Our phylogenetic analyses indicate that the
269 divergence between *TXNRD1/TXNRD3* and *TXNRD2* likely predates the origin of
270 metazoans, as representatives from both clades are found in diverse animal lineages,
271 including the nematode *C. elegans*, the annelid *Helobdella robusta*, and the cnidarian
272 *Nematostella vectensis*. Secondary events are evident in several lineages, and detailed
273 hereafter. The tunicate *Ciona intestinalis* lost *TXNRD2*: we only detected one TXNRD
274 homolog in this species, consistent with an earlier report [15], which clusters with
275 *TXNRD1/TXNRD3* in the reconstructed gene tree (Supplementary Figure S6). Analogous
276 observations indicate that arthropods such as *Ixodes scapularis* and *D. melanogaster*

277 lost *TXNRD1* (Supplementary Figure S6). Notably, *D. melanogaster* retains two *TXNRD2*-
278 derived genes, presumably the result of a secondary duplication. Fitting with our
279 analysis, non-vertebrate animals carried Grx-domains solely in *TXNRD3* genes. Due to
280 the presence of this domain, these invertebrate genes were referred as *TGR* in literature,
281 despite weak or absent glutathione reductase activity [35]. Altogether, we reconstruct
282 that ancestral vertebrates carried a cytosolic Grx-containing *TXNRD3* gene and a
283 mitochondrial *TXNRD2* gene, the former later giving rise to *TXNRD1* in Gnathostomata.

284 We further detected one *TXNRD3* duplication occurred in Salmonidae and yielded
285 *TXNRD3b*, which retained the Grx domain. Lastly, the Hyperoartia clade of Cyclostomata
286 lost *TXNRD2* (despite retaining its neighboring genes) and thus encodes *TXNRD3* as their
287 sole gene in this family. *TXNRD2* was detected in its sister clade, Myxini (Supplementary
288 Figure S1).

289 <u>SELENOJ</u>

290 Selenoprotein J (SELENOJ) shows significant similarity to jellyfish J1-crystallins and,
291 together with them, forms a distinct subfamily within the large family of ADP-
292 ribosylation enzymes. While the homology may suggest a structural role which would
293 be unique among selenoproteins, SELENOJ remains functionally uncharacterized. The
294 *SELENOJ* gene is reportedly restricted to actinopterygian fishes and sea urchins, with Cys
295 homologs reported in cnidarians [11]. Our analyses replicated these earlier findings, and
296 additionally highlighted the presence of *SELENOJ* in Amphioxiformes (Supplementary
297 Figure S10). On the other hand, *SELENOJ* was lost in Cyclostomata and Batoidea (rays e.g
298 *Leucoraja erinaceus*)(Supplementary Figure S1). Moreover, we had previously reported
299 a *SELENOJ* duplication in *O. latipes* and *Gasterosteus aculeatus* [16], yielding *SELENOJ2*.
300 Here, we traced back this duplication to the root of Clupeocephala (76% UFB support),
301 which included more than half of fishes in our dataset (Figures 1, 6a, Supplementary
302 Figures S10, S11). Later, *SELENOJ2* was lost independently in Otophysi (e.g *Danio rerio)*

303 and Tetraodontidae (e.g. *Tetraodon nigroviridis*). On the other hand, *SELENOJ*
304 underwent a Sec-to-Cys conversion in Salmonidae.

305 SELENOO

306 Selenoprotein O (SELENOO) is the largest known selenoprotein by molecular weight,
307 and belongs to a widespread protein family recently functionally characterized as a
308 mitochondrial pseudokinase that transfers AMP to various protein substrates, implicated
309 in oxidative stress response and metastasis [36, 37]. Previous studies identified a
310 *SELENOO* paralog named *SELENOO2* in *D. rerio*, which was absent in the genomes of
311 *Takifugu rubripes*, *G. aculeatus*, *O. latipes*, *and T. nigroviridis* [16]. In our analysis, we
312 replicated these results, and yet we observed the presence of *SELENOO2* in many other
313 fish genomes, including among others all of Cyprinodontoidei (toothcarps) and
314 Carangaria (e.g. jacks, flatfishes), which implies a complex pattern of gene evolution
315 (Figures 1, 6b, Supplementary Figures S12, S13). Ultimately, we hypothesize that
316 *SELENOO2* emerged at the root of Clupeocephala but was subsequently lost in specific
317 lineages, namely Tetraodontidae, Oryziinae, Anabantaria, and almost all Perciformes (e.g.
318 *G. aculeatus*).

319 Methionine sulfoxide reductase B (MSRB)

320 Methionine sulfoxide reductase B (MSRB, previously known as SELENOR), are enzymes
321 that reduce oxidized methionine residues, with roles for antioxidant defense and redox-
322 based modulation of various pathways including actin assembly [38]. Humans possess
323 one Sec-containing gene, *MSRB1*, and two Cys paralogs, *MSRB2* and *MSRB3*. Our
324 analyses (Figure 6c, Supplementary Figures S14 and S15) indicate that the three genes
325 predate the origin of vertebrates, since members of each subfamily were detected in
326 invertebrates. Within fish lineages, *MSRB1* underwent a complex history of duplications
327 and losses. We previously identified a Sec-containing gene, *MSRB1b*, in five

328 actinopterygian species [16]. Based on the protein tree topology (Figure 6c,
329 Supplementary Figure S14; UFB support 58%) and the pattern of gene occurrence across
330 the species tree (Supplementary Figure S15), we mapped its origin to the subclade of
331 Osteoglossocephalai, which comprises all Actinopterygii in our dataset except the
332 *MSRB1b*-lacking species *Erpetoichthys calabaricus* and *Lepisosteus oculatus*. Gene
333 counts per species further indicated secondary *MSRB1b* duplications; yet, the
334 reconstructed protein tree exhibited a scattered topology, without the typical expected
335 clustering pattern (Figure 6c). Based mostly on the species tree, we concluded that
336 *MSRB1b* likely duplicated in Salmonidae yielding *MSRB1b2,* and in Euacanthomorphacea
337 yielding *MSRB1b3;* and that both *MSRB1b* and the duplicated copy *MSRB1b3* have been
338 later lost in multiple lineages: *MSRB1b* in Esociformes, Tetraodontidae and Anabantaria
339 and *MSRB1b3* in Perciformes, *Cynoglossus semilaevis* and Cyprinodontiformes. Notably,
340 this reconstruction highlights at least three convergent events of *MSRB1b3* loss in
341 Percomorphacae, and another three loss events of *MSRB1b* in Euteleosteomorpha.
342 Relevantly, other selenoprotein families show similar patterns of convergent losses in
343 this lineage (Figure 1), as discussed later.

344 <u>SELENOT</u>

345 Selenoprotein T (SELENOT) is a thioredoxin-like enzyme anchored to the endoplasmic
346 reticulum (ER) membrane. It is highly conserved throughout evolution, and it has an
347 essential role in cellular homeostasis and redox regulation [39]. While mammals have a
348 single *SELENOT* gene, previous studies reported three homologs in actinopterygian
349 fishes [16], here refined and expanded. We found that most actinopterygian groups
350 possessed two *SELENOT* genes, namely *SELENOT* (aka *SELENOT1*) and *SELENOT2*; yet,
351 Salmonidae, Otophysi and Cyprinoidei exhibited further lineage-specific duplications.
352 Ultimately, we mapped the origin of *SELENOT2* to a duplication of *SELENOT1* at the root
353 of Actinopterygii (Figure 7a). *SELENOT1* duplicated again to yield *SELENOT1b* in Otophysi,

354 i.e. the group including Cyprinoidei (e.g. *D. rerio* and carps), and Characiphysae (e.g.
355 *Electrophorus electricus, Pygocentrus nattereri).* In Cyprinoidei, further duplications of
356 *SELENOT2* and *SELENOT1b* gave rise to the genes *SELENOT2b* and *SELENOT1b2*,
357 respectively (Figures 1, 7a, Supplementary Figure S16). Finally, we observed yet another
358 *SELENOT1* duplication in Salmonidae, yielding the gene *SELENOT1a2*. Interestingly,
359 Hyperoartia also independently expanded this family through one gene duplication,
360 which yielded a divergent paralog that we designated *SELENOT3*. We deem all *SELENOT*
361 duplications highly reliable due to the consistency between the species tree and the
362 protein tree topology, and to the high UFB values in the latter (Figure 7a).

363 <u>SELENOU</u>

364 The Selenoprotein U (SELENOU) family includes selenoprotein genes in birds and fishes,
365 the latter reported to carry up to three Sec-containing paralogs [10, 16]. Mammals and
366 amphibians are reported to encode only Cys-homologs, with human containing three
367 paralogs (*SELENOU1/PRXL2A*, *SELENOU2/PRXL2C*, *SELENOU3/PRXL2B*), involved in
368 diverse functions such as antioxidant defense, prostaglandin biosynthesis, and glycolysis
369 regulation. The SELENOU gene family exhibits a scattered pattern of occurrence,
370 implying a complex evolutionary history involving gene duplications, losses, and Sec-
371 to-Cys conversions. Here, we provide our best attempt to explain their history with
372 current data (Figures 1, 7b, 7c, Supplementary Figures S1, S17).

373 The split of *SELENOU1/2/3* predates vertebrates, as members of the three subfamilies
374 were detected in invertebrate species (Supplementary Figures S17). All Sec-encoding
375 genes detected in this family belonged to the *SELENOU1* branch. Within fish, we mapped
376 a first well-supported duplication of the *SELENOU1* gene in Osteoglossocephalai,
377 yielding *SELENOU1a2* (Figure 7b). This gene was then lost at the root of the large group
378 of Acanthomorphata, while it converted to Cys homolog in *Esox lucius*. Prior to both of
379 these events, we traced two additional *SELENOU1* duplications in Clupeocephala,

380 yielding the *SELENOU1b* and *SELENOU1c* genes. *SELENOU1b* was then lost
381 independently in three lineages: Otophysi, Esocidae and Atherinomorphae
382 (Supplementary Figure S1). Finally, we detected another two independent well-
383 supported *SELENOU1* duplications in Cyprinoidei and Salmonidae, yielding the gene
384 *SELENOU1a1b* and *SELENOU1a1c*, respectively (Figure 7b). Lastly, within tetrapods,
385 *SELENOU1* underwent at least four convergent Sec-to-Cys conversions (Figure 7c),
386 specifically in Theria (i.e. placentals and marsupials), Anura (frogs), Testudines (turtles),
387 and Psittaciformes birds (parrots).

388 <u>SELENOW, SELENOV and MIEN1</u>

389 Selenoprotein W (SELENOW) is a widespread glutathione-dependent antioxidant
390 oxidoreductase implicated in resolution of inflammation, with reported roles in muscle
391 growth and differentiation, as well as in protecting neurons from oxidative stress during
392 neuronal development [40, 41]. Selenoprotein V (SELENOV) is a placental-specific
393 putative oxidoreductase expressed in testis and potentially involved in antioxidant
394 defense and ER stress, whose C-terminus shares substantial sequence similarity to
395 SELENOW [42]. Our present analyses are consistent with SELENOV emerging in
396 placentals by duplication of *SELENOW* (Figure 8a). It was previously reported that the
397 *SELENOV* gene was lost in *Gorilla gorilla* [16]. By analyzing the species tree annotated
398 with *SELENOV* gene predictions (Supplementary Figure S1), we pinpointed additional
399 potential losses, which we sought to corroborate by searching RNA datasets (Methods).
400 Due to *SELENOV* detection in RNA data, we ultimately dismissed all apparent losses
401 except two: the known one in *Gorilla gorilla*, and a novel one in Odontoceti (toothed
402 whales, including dolphins) (Figure 1).
403 Besides *SELENOV*, mammals have a single other homolog in this gene family: *SELENOW*,
404 also referred to as *SELENOW1*. Our analyses indicate that this gene was independently
405 lost in Cyclostomata, Chondrichthyes and within Actinopterygii, specifically in

406 Acanthomorphata, resulting in its absence from the majority of extant fish genomes
407 (Supplementary Figure S1).

408 Earlier work pinpointed multiple *SELENOW* homologs in non-mammalian vertebrates,
409 including two subfamilies named *SELENOW2* and *MIEN1* (formerly Rdx12) [43]. Indeed,
410 our analysis recovered these two clades: they were defined by a duplication node
411 supported by ~35 species, and by performing BLASTP of each sequence, we noted that
412 the majority of one clade matched sequences annotated as SELENOW2, while the other
413 matched MIEN1 annotations. These clades were confidently placed as sister groups, with
414 the *SELENOW* and *SELENOV* clade as their closest relative (Figure 8a). Within the
415 SELENOW2/MIEN1 group, all genes belonging to invertebrates were placed in the
416 *SELENOW2* branch (Supplementary Figure S18). This suggests that *SELENOW2* predates
417 vertebrates, while *MIEN1* emerged within this lineage. We detected at least one *MIEN1*
418 gene in each vertebrate genome, except Cyclostomata wherein it was consistently
419 absent. Therefore, we propose that *MIEN1* originated by a *SELENOW2* duplication at the
420 root of Gnathostomata. In Amniota, MIEN1 genes carry Cys in place of Sec, implying a
421 Sec-to-Cys conversion in this clade. On the other hand, several Actinopterygian lineages
422 contained multiple Sec-containing *MIEN1* copies, which we ascribed to secondary
423 lineage-specific gene duplications, detailed hereafter. Because of poor consistency
424 between the protein tree topology and species taxonomy, here we heavily relied on the
425 species tree to infer gene histories. Cyprinoidei contained three copies; in the protein
426 tree, *MIEN1* and *MIEN1d* exhibited the expected topology for duplication, while *MIEN1b*
427 was placed closer to the root, possibly due to long branch attraction (Figure 8). We
428 consider most likely that two *MIEN1* duplications at the root of Cyprinoidei generated
429 *MIEN1b* and *MIEN1d*. Salmonidae also exhibited three copies: *MIEN1*, *MIEN1e*, *MIEN1c*.
430 We argue that these emerged independently of those in Cyprinoidei, since virtually all
431 other Clupeocephala contained a single copy (Supplementary Figure S1). *MIEN1e* was

found in Salmonidae and also in their outgroup *E. lucius*, which together form the superorder of Protacanthopterygii. We ascribe the *MIEN1e* origin to a duplication at the root of this lineage; the protein tree topology is consistent with this scenario, albeit with very low UFB support in the corresponding node (6%). The other copy (*MIEN1c*), in contrast, was detected in Salmonidae only. Surprisingly, this clustered with Cyprinoidei *MIEN1* and *MIEN1d*. We consider that the most likely scenario is that *MIEN1c* emerged in Salmonidae after the split from *E. lucius*, then its high divergence confounded its phylogenetic position. Lastly, Poeciliidae contained two copies, *MIEN1* and *MIEN1f*. We ascribed this to a gene duplication at the root of this lineage; the protein tree topology is consistent with this, despite an unexpected clustering with *MIEN1* of Chondrichthyes.

The *SELENOW2* branch of the gene tree includes selenoprotein genes from invertebrates, Cyclostomata, Amphibia, and diverse Actinopterygii. The absence of Amniotes implies a gene loss in this lineage, as previously reported [16]. In Actinopterygii, on the other hand, the history of this gene is more obscure and difficult to solve. *SELENOW2* occurs in a scattered pattern across the species tree, apparently implying many independent losses, which may seem unlikely (Supplementary Figure S1). Therefore, we first set to assess the quality of our tree-based orthology assessment using a complementary syntenic analysis. We observed limited but detectable syntenic conservation around *SELENOW1* across Actinopterygii and Tetrapoda, shown by consistent nearby presence of *EHD2/EHD2A* (Supplementary Figure S19). Similarly, genes in the *SELENOW2* tree branch are consistently located in the vicinity of the *TRPM7* gene in all tested Actinopterygii and Amphibia (Supplementary Figure S20). As per MIEN1, syntenic conservation was clear within Actinopterygii (*SLC4A1A, UBTF, ATXN7I3/ATXN7I3A*) and within Amniota (*IKZF3*, *PGAP3, GRB7, ERBB2*), but not in-between (Supplementary Figure S21). No common nearby genes were observed between *SELENOW1*, *SELENOW2* and *MIEN1*. Altogether, these analyses corroborated

458 the true orthology of these protein clades. Therefore, we proceeded to map *SELENOW2*
459 losses within Actinopterygii by examining the species tree annotated with gene presence
460 (Supplementary Figure S1) and seeking validation of putative losses in RNA data.
461 Strikingly, we identified 13 independent losses (after dismissing two due to *SELENOW2*
462 detection in RNA sequences; Supplementary Table T2), displayed in Figure 1. Although
463 additional *SELENOW2*-like RNA sequences were recovered in certain genomes (e.g.,
464 *Oryzias melastigma, Ictalurus punctatus*), further inspection indicated that these
465 sequences likely represented contamination, and they were therefore excluded.

466 SELENOP

467 Selenoprotein P (SELENOP) is the only mammalian selenoprotein with multiple Sec
468 residues, e.g. 10 in human and *M. musculus*, due to its Sec-rich C-terminal tail. SELENOP
469 is a key player in selenium transport and distribution, serving as the primary selenium
470 carrier in the bloodstream, delivering selenium to peripheral tissues via receptor-
471 mediated uptake. We recently characterized the evolution of the *SELENOP* gene family,
472 showing that its Sec content can vary by two orders of magnitude across Metazoa [44].
473 Its C-terminal tail is challenging to identify using genomic data due to its fast evolution,
474 so that here we mostly focused on gene-level detection. As expected, we identified two
475 phylogenetic clades of vertebrate genes, multi-Sec *SELENOP/SELENOP1*, and single-Sec
476 *SELENOP2/SELENOPB*. While the former is reportedly ubiquitous in vertebrate
477 genomes, the latter was lost in Eutheria [44]. Moreover, *SELENOP2* was converted to a
478 Cys homolog in Anura. Our present analysis (Figure 8b, Supplementary Figure S22)
479 confirmed these findings, and offered additional insights into early vertebrate evolution.

480 Invertebrate *SELENOP* homologs are consistently recovered outside of the clade
481 containing both vertebrate paralogs, supporting the hypothesis that the
482 *SELENOP1/SELENOP2* duplication arose at the root of vertebrates. Most fish genomes

483 contained the two Sec-*SELENOP* genes; a notable exception was the Selachii division
484 within Chondrichthyes (i.e. sharks), where *SELENOP2* was converted to Cys homolog. In
485 Salmonidae, we traced one lineage-specific gene duplication for each paralog, yielding
486 *SELENOP1b* and *SELENOP2b*. Like most fish, Cyclostomata contains two genes, one
487 within the *SELENOP1* branch and the other within *SELENOP2*. Interestingly, the latter
488 displays a unique CXXU motif (rather than the more typical UXXC) in its N-terminal
489 thioredoxin-like domain. Moreover, though *SELENOP2* of other vertebrates encodes a
490 single Sec, in Cyclostomata it exhibits an extraordinary expansion of Sec content: it
491 harbors 162 and 139 in-frame UGA codons in the lampreys *Lethenteron reissneri* and
492 *Lampetra fluviatilis*, respectively. These genes were detected in genome assemblies,
493 and their sequence was confirmed by highly similar matches in transcriptome shotgun
494 assembly (TSA) data of other Cyclostomata, spanning all the putative transcript
495 (Supplementary Document D1). Analogously to other *SELENOP* genes [44], the vast
496 majority of putative Sec residues are encoded within an expanded modular C-terminal
497 region, and two SECIS elements are detected downstream. We argue that all these in-
498 frame UGAs are translated as Sec: (i) homologous SECIS elements from human, *M.*
499 *musculus* and *D. rerio SELENOP* have been shown to support processive insertion at
500 multiple Sec-UGA sites in a variety of *in vitro* and *in vivo* models [4, 44]; and (ii) the
501 lamprey coding sequence contains more than 130 stop codons exclusively of the UGA
502 type (rather than a mixture with UAA or UAG), providing strong evidence of selective
503 pressure to preserve the same peculiar gene architecture that is observed in hundreds
504 of multi-Sec SELENOP genes across Metazoa [44]. With this record, lampreys surpasses
505 all other metazoans in the number of Sec-UGAs in *SELENOP*, including the previous high
506 of 132 reported in bivalves [44]. Between the SECISes, we noted the presence of a
507 distinctive region rich in CA dinucleotides (-1500-1800nt) of unknown significance
508 (Supplementary Document D1).

509 Other Salmonidae-specific duplications: SELENOE, SELENOF, SELENOH

510 Our analysis revealed a substantial selenoproteome expansion specific to the lineage of
511 Salmonidae, which may suggest an increased functional reliance of this clade on
512 selenium. Besides the aforementioned expansions in the GPX, DIO, TXNRD, MSRB,
513 SELENOT, SELENOU and SELENOW protein families (Figure 1), we detected well-
514 supported gene duplications of *SELENOE* and *SELENOF*, yielding *SELENOEb* and
515 *SELENOFb*, respectively, and another two for *SELENOH*, yielding *SELENOHb* and
516 *SELENOHc* (Figure 9, Supplementary Figures S23, S24, S25). Briefly, SELENOF is a ER-
517 resident selenoprotein involved in redox quality control [45], recently pinpointed as
518 tumor suppressor [46, 47]. Our results show that *SELENOF* is ubiquitous to vertebrates.
519 On the other hand, SELENOE (formerly Fep15), a functionally uncharacterized
520 selenoprotein distantly related to SELENOF [12], was found as selenoprotein only in
521 fishes: Anura carry a Cys-homolog, while the rest of tetrapods (Amniota) lost the gene
522 altogether. Notably, we mapped the origin of the *SELENOE* gene to the root of
523 Gnathostomata, since we did not find it in Cyclostomata nor in the non-vertebrate
524 outgroups. Finally, SELENOH is a widespread nucleolar oxidoreductase that regulates
525 redox homeostasis and suppresses DNA damage through the p53 pathway, and it is
526 implicated in cancer proliferation [48–50]. Apart from the Salmonidae-specific
527 duplications, no other gene evolution event was detected for *SELENOE*, *SELENOF*, and
528 *SELENOH*, with the only exception of the loss of *SELENOH* in Cyclostomata, and in
529 independently in *Callorhinchus milii*.

530 It is worth mentioning that, because all the Salmonidae genomes in our dataset belong
531 to its major subgroup, Salmoninae, we had initially parsimoniously mapped all the gene
532 duplications described above to Salmoninae. However, inspection of intron positions
533 relative to protein sequence alignments (Figures 2, 3, 4, 6) revealed that they were
534 conserved across these lineage-specific duplications. This pattern is consistent with

535 DNA-based gene duplication rather than RNA-based retrotransposition. Consequently,
536 the whole-genome duplication (WGD) previously reported at the root of Salmonidae
537 [51] provides the most plausible explanation for the origin of these genes, leading us to
538 map these events to Salmonidae as a whole rather than exclusively to Salmoninae.

539 Other families: SEPHS2, SELENOL, SELENOK, SELENOM, SELENON, SELENOS

540 Hereafter, we briefly summarize the remaining vertebrate selenoproteome families,
541 wherein our present analysis recapitulated earlier results without novel findings. As
542 already introduced, SEPHS2 (aka SPS2) is a selenoprotein whose function is critical for
543 Sec biosynthesis. As expected [16, 22], we identified the *SEPHS2* gene in all vertebrate
544 lineages. (As discussed in the first section of the Results, the gene appears to be absent
545 from many assemblies of birds, but we attribute this to a sequencing bias rather than
546 true gene loss). The *SEPHS2* gene is intronless only in genomes of Eutheria (placentals).
547 As previously shown [16], the ancestral intron-containing *SEPHS2* gene (referred to as
548 *SEPHS2a*) duplicated by retrotransposition at the root of Mammalia, yielding an intron-
549 less copy referred to as *SEPHS2b*. Both copies are present in extant genomes of
550 Marsupialia, while *SEPHS2a* was then lost in Eutheria. Our results confirmed these
551 findings (Supplementary Figures S1, S26). Thus, the sole placental *SEPHS2* gene is
552 actually *SEPHS2b*. Of note, *SEPHS2* duplicated at the root of vertebrates, and
553 independently also in various other metazoan lineages, to yield a non-Sec/non-Cys
554 paralog known as *SEPHS1/SPS1* [22]. The vertebrate *SEPHS1* gene carries threonine in
555 place of Sec. This kind of conversion is highly unusual for selenoproteins, since Sec is
556 typically catalytic and it is highly dissimilar from canonical amino acids other than Cys.
557 Indeed, SEPHS1 appears to lack the enzymatic activity of SEPHS2 (selenophosphate
558 synthesis from selenide), and instead has a function in gene expression-mediated
559 control of redox homeostasis [6, 52].

560 SELENOL is a peculiar selenoprotein featuring two nearby Sec residues that can form a
561 diselenide bond. It features a peroxiredoxin-like motive suggestive of a redox-related
562 function [13]. As expected [16], we detected the *SELENOL* gene in nearly all vertebrates
563 except Tetrapoda, confirming a gene loss at the root of this lineage (Supplementary
564 Figure S1).

565 The remaining selenoprotein families were detected in a single Sec-containing copy in
566 all investigated genomes (with the only exception of *SELENOM*, lost in Hyperoartia;
567 Supplementary Figure S1), barring pseudogenes and spurious absence attributed to
568 imperfect assembly quality. Curiously, all these single-copy selenoproteins localize to
569 the ER. They include: SELENOK, involved in calcium homeostasis and resolution of ER
570 stress [53]; SELENOM and SELENOS, involved in unfolded protein responses [54]; and
571 SELENON, involved in calcium and redox homeostasis [55].

572 <u>Shifts in selenoproteome-to-proteome ratio across vertebrate evolution</u>

573 Our analyses revealed that the vertebrate selenoproteome expanded through many
574 gene duplications in fish lineages. The conservation of intron structures and the
575 phylogenetic correspondence with documented WGD suggests that many of these gene
576 duplications resulted from such events (Figure 1). This raises the question of whether
577 such selenoproteome expansions are adaptive, or are neutral consequences of having a
578 larger proteome. To gain insights in this matter, we compared the size of our predicted
579 selenoproteome with that of the annotated proteome for each species (Figure 10). We
580 fit a zero-intercept linear model separately for each phylogenetic clade of interest
581 (Mammalia, Sauropsida, Chondrichthyes, Salmonidae, Cyprinoidei, and non-
582 Salmonidae non-Cyprinoidei Actinopterigyii), with the hypothesis that the slope (i.e. the
583 selenoproteome-to-proteome ratio) may differ among clades. We found that
584 Actinopterygii have significantly larger selenoproteome-to-proteome ratios than
585 Mammalia (Figure 10). Importantly, aquatic environments have long been hypothesized

586 to favor selenoprotein usage [56], and our present analysis mostly supports this pattern.
587 Yet, some lineages diverge from this trend: aquatic Salmonidae and Cyprinoidei, which
588 have the biggest selenoproteomes across vertebrates, show low ratios (albeit with
589 substantial statistical uncertainty, due to the few available genomes and their spread in
590 apparent proteome size; see their wide confidence intervals in Figure 10). On the other
591 hand, Sauropsida exhibits relatively high ratios despite being land-dwelling. Overall, this
592 data is best explained by an ancestral state of rich selenoproteome-to-proteome ratio,
593 subsequently strongly decreased in Mammalia, and possibly slightly increased in
594 Actinopterygii.

595 To formally assess the impact of habitat on selenoproteome size, we fitted a series of
596 regression models relating proteome size (denoted as P) to selenoproteome size (S), with
597 habitat (H: aquatic vs. terrestrial) included either as an additive term or as an interaction
598 with proteome size. Because species are not statistically independent, we implemented
599 phylogenetic generalized least squares under both Brownian motion (neutrality) and
600 Ornstein–Uhlenbeck (balancing selection) models, thus incorporating the species tree as
601 a covariance structure. We also tested alternative codings of habitat in which, besides
602 fish, Cetacea and Amphibia were classified as aquatic. Additionally, we considered
603 models in which only Cetacea were classified as aquatic, while Amphibia were treated
604 as terrestrial. Across all tested models, the effect of aquatic habitat was highly significant
605 (Table 1), consistently indicating an enrichment of the selenoproteome in aquatic
606 lineages. Models with the interaction term (S ~ P + H*P + 0) provided a better fit than
607 additive models (S ~ P + H + 0), consistent with aquatic species having a proportionally
608 larger selenoproteome relative to proteome size. Likelihood ratio tests further supported
609 the habitat effect, as each model outperformed its corresponding null model lacking H
610 (Table 1).

611 ## DISCUSSION

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

621 Our reconstruction of the ancestral vertebrate selenoproteome consists of 27
622 selenoprotein genes representing all the 19 known vertebrate Sec-encoding gene
623 families (Figure 1). Our conclusions mostly overlap with our earlier analysis [16], with the
624 exceptions of *GPX2* and *SELENOE* (both now mapped to Gnathostomata rather than
625 Vertebrata), and *GPX3b* (Vertebrata rather than Actinopterygii). Besides, we now
626 consider *TXNRD3* as the ancestral gene in the *TXNRD1/TXNRD3* group, with *TXNRD1*
627 emerging in Gnathostomata. We also concluded that these selenoproteins emerged
628 specifically at the root of Vertebrata: *GPX3b, DIO2, DIO3, SELENOI, SELENOP2*. Among
629 these, *SELENOI* stands out in that the whole family appeared at the root of vertebrates:
630 there is no homolog that includes the Sec position, regardless of the aligned amino acid,
631 outside of vertebrates.

632 On the other hand, our present work substantially expands the catalog of selenoproteins
633 in fish. Aquatic environments have long been hypothesized to favor selenoprotein usage,
634 a trend first suggested by Lobanov et al. from limited genomic data [56]. Subsequent
635 analyses further corroborated this model by showing relaxed purifying selection on
636 selenoprotein genes in terrestrial vertebrates [57], alongside reduced Sec content in
637 SELENOP, an established proxy for selenium demand [44, 57]. Our present analysis of

638 the selenoprotein-to-proteome ratio, leveraging the largest dataset to date and
639 phylogeny-aware models, provides further support for this evolutionary pattern.
640 However, we note that the difference between the selenoproteome of aquatic and land-
641 dwelling species appears primarily driven by mammalian loss and actinopterygian
642 expansion.

643 The underlying cause of the selective advantage of Sec in aquatic environments is
644 arguably yet unclear. It has been proposed that land-dwelling organisms must face (i)
645 reduced selenium bioavailability and (ii) increased oxygen exposure [56, 57]. While the
646 latter seems intuitively relevant, it contradicts one of the prevalent proposed
647 biochemical rationale for usage of selenium (Sec) over sulphur (Cys) [58, 59]: its
648 resistance to (over)oxidation - a trait that would suggest selenoproteins should be more
649 beneficial in terrestrial environments. Moreover, our data shows that selenoproteins
650 belonging to antioxidant families are more prevalent in fish compared to tetrapods.
651 While further research is required to fully resolve this question, we argue that the current
652 evidence better supports selenium bioavailability as the most likely evolutionary driver
653 of Sec usage in aquatic vs land environments.

654 These observations may depict our own selenoproteins as an unstable remnant of an
655 aquatic past. Yet, it is worth noting that (i) land-living tetrapods exhibit substantial
656 purifying selection, preventing Sec to Cys conversions in most genes [60]; and (ii) the
657 aquatic selenoproteome is far from static, having expanded under putative adaptive
658 forces at the root of Actinopterygii and in multiple of its clades. We found that these
659 duplications occurred primarily in protein families linked to antioxidant defense, either
660 with well-established roles (GPX, MSRB) or with putative/indirect functions (SELENOT,
661 SELENOU, SELENOW). In contrast, no variation was observed in ER-resident proteins
662 with functions in calcium homeostasis (SELENOK, SELENON) or protein quality control /
663 unfolded protein response (SELENOM, SELENOS). While this trend was not absolute -

664 e.g. ER-resident SELENOE and SELENOF duplicated in Salmonidae, and the frequently
665 duplicated SELENOT is also ER-resident - it suggests that selenoproteome expansions in
666 Actinopterygii may have been driven primarily by selection for enhanced antioxidant
667 capacity. While we did not investigate the exact mechanism of gene duplications, we
668 note that, strikingly, 40 selenoprotein duplications out of the 56 detected across
669 vertebrates were mapped to taxonomic nodes corresponding with reported WGD
670 events (Figure 1). This observation reinforces the importance of WGDs in shaping
671 vertebrate proteomes.

672 Among tetrapods, our most striking finding is the recurrence of convergent evolutionary
673 events, including Sec-to-Cys conversions in SELENOU1 and GPX6, as well as
674 independent losses of SELENOV, all of which are supported by robust phylogenetic
675 evidence. It is important to note that, because our reconstruction is based on maximum
676 parsimony, our estimate of such events represents a conservative lower bound.
677 Interestingly, both SELENOV and GPX6 are the "youngest" mammalian selenoproteins
678 and they are specifically expressed in the testis, a tissue recognized for its role in the
679 evolution of new genes [61].

680 Our study has some limitations. First, we restricted our search to genes homologous to
681 known selenoprotein families, because existing methods for identifying novel
682 selenoproteins suffer from low specificity [62]. As a result, we may have overlooked
683 lineage-specific selenoproteins, particularly in fish lineages where selenoproteome
684 expansions are evident. Second, we could only detect events within lineages
685 represented in our dataset. Given the dynamic nature of the Actinopterygii
686 selenoproteome, additional genome sequences from this group will likely reveal further
687 duplications. Third, our analysis assumes the species tree as ground truth, which may
688 not be entirely accurate. For instance, the convergent loss of the same genes in
689 Tetraodontidae, Perciformes, Anabantaria, and *Cynoglossus semilaevis* (Figure 1) could

690 be explained as a single ancestral loss if the non-Perciform Eupercaria and Carangaria
691 were instead the outgroups of these clades - yet, we found no supporting evidence for
692 this alternative phylogeny in the literature [63]. Fourth, further data or alternative
693 methodological approaches could refine the evolutionary history of selenoprotein
694 families with the most complex gene dynamics, such as SELENOW and SELENOU, and
695 provide deeper insight into lineages with highly complex genome evolution, such as
696 Cyprininae.

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

700 MATERIAL AND METHODS

701 Genome sequences and species tree

702 We downloaded all genome assemblies and annotations available in Ensembl version
703 v.108 [64] comprising 315 genomes, each corresponding to a different species or strain.
704 These included 310 vertebrates, two tunicates, one nematode (*C. elegans*), one insect
705 (*D. melanogaster*), and yeast *Saccharomyces cerevisae* (here excluded from all
706 analyses). The phylogenetic tree of species was downloaded from Ensembl Compara
707 v.e110 [20]. To refine our reconstruction of the ancestral vertebrate selenoproteome, we
708 also analyzed NCBI assemblies from species with informative taxonomy; identifiers are
709 provided in Supplementary Table T3. Specifically, we included various early-branching
710 vertebrates: nine cartilaginous fish (Chondrichthyes) and nine jawless vertebrates
711 (Hyperoartia class, belonging to the Cyclostomata superclass); then four
712 amphioxiformes (Branchiostoma), serving as additional close relatives to vertebrates,
713 besides the two tunicates already in Ensembl; and two remote outgroups:
714 *Strongylocentrotus purpuratus* (Echinodermata, deuterostome) and *Strigamia maritima*

715 (Arthropoda, protostome). These 24 genomes, together with the 314 Ensembl
716 assemblies, constituted our main dataset. Later, we searched two additional
717 Cyclostomata genomes, belonging to the class of Myxini, whose predictions were
718 included in the protein tree reconstruction of TXNRD. For the rest of families, we
719 distinguished between Cyclostomata, Myxini, and Hyperoartia specific events by
720 inspecting the species tree annotated with selenoprotein predictions (Supplementary
721 Figure S1). For follow up analyses to confirm or dismiss gene evolution events, we then
722 inspected gene predictions on an extended set of 1,172 vertebrate genome assemblies
723 downloaded from NCBI, using the NCBI taxonomy tree as the rough backbone for
724 evolutionary interpretation.

725 **Selenoprotein gene prediction**

726 Selenoprotein gene prediction was performed using Selenoprofiles [19, 65, 66] version
727 v4.5.0, available at https://github.com/marco-mariotti/selenoprofiles. Selenoprofiles is
728 a computational pipeline to identify members of the known selenoprotein families and
729 related proteins. It uses several homology-based gene prediction programs, and
730 employs a set of manually curated multiple sequence alignments of selenoprotein
731 families to scan genomes for homologues. The profiles we used are available at
732 https://github.com/marco-mariotti/selenoprotein_profiles (v1.3.0). We scanned
733 genome assemblies for all known selenoprotein families in Metazoa. For most analyses,
734 we focused on the predictions labelled as "selenocysteine", i.e., the genes encoding for
735 selenoproteins, with no apparent pseudogene features (e.g. frameshifts).

736 **Phylogenetic reconstruction of selenoprotein families**

737 The protein sequences of all Selenoprofiles predictions across Ensembl genomes were
738 collected using the Selenoprofiles *join* utility, and alignments were computed using
739 ClustalO v.1.2.4 [67]. These were used as input to phylogenetic reconstruction using IQ-

740 TREE v3.0.1 [68]. For each selenoprotein family, the best-fitting substitution model was
741 automatically selected using ModelFinderPlus. The selected models are listed in
742 Supplementary Table T4. Phylogenetic trees were inferred under the corresponding
743 substitution model, and the branch support values were estimated using 1,000 ultrafast
744 bootstrap replicates [68]. We visualized gene trees with the show_tree program in the
745 biotree_tools package v0.0.6 (https://github.com/marco-mariotti/biotree_tools), based
746 on the ETE3 python module [69], specifically adapted for custom visualization in this
747 manuscript. The resulting vectorial images were further annotated manually using the
748 Inkscape software. Sequence and tree manipulations throughout the project were
749 carried out using the programs *alignment_tools* and *tree_tools* from the biotree_tools
750 package, all based on ETE3 [69]. Sequences in the *SELENOP* gene family were trimmed
751 prior to tree building to remove the highly divergent C-terminal Sec-rich tail, using
752 custom code based on Pyaln v0.1.4 (https://github.com/marco-mariotti/pyaln).

753 **Analysis of gene evolution events**

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

761 A visualization of the species tree together with all gene predictions in each genome was
762 produced using the *tree drawer* utility available in Selenoprofiles, customized as follows.
763 First, each gene prediction corresponding to a multi-member gene family were assigned
764 to a mammalian-centric homology group, hereafter referred to as gene "subfamily" (e.g.

765 *GPX1, GPX2, GPX3*, etc), using the *orthology* utility of Selenoprofiles. Then, we encoded
766 the expected selenoproteome for every lineage based on literature; e.g. bats, since they
767 are mammals with no further lineage-specific events reported, were expected to carry
768 one Sec-encoding gene for each of *GPX1, GPX2, GPX3, GPX4* and *GPX6*. In our custom
769 *tree drawer* visualization (Supplementary Figure S1), we marked each gene prediction
770 based on the deviation from such expectations; e.g., bats were labelled as missing a Sec-
771 *GPX6*. This highlighted the presence of extra genes (putative duplications) and missing
772 genes (putative losses). Ultimately, the joint analysis of gene and species trees led us to
773 manually map each event into the species tree, following criteria of maximum parsimony
774 and consistency with the known phylogenetic relationship among species.

775 Lastly, each event was critically evaluated to exclude potential artifacts. We required
776 support by at least two sister species for most events. Single-species events were
777 deemed reliable only if supporting independent data (e.g. RNA sequences) existed. Due
778 to imperfect quality of genome assemblies, we sought confirmation of each individual
779 gene loss using independent sequence data from the lineage of interest, as follows. At
780 this stage, each putative gene loss consisted in the lack of an apparently functional gene
781 (i.e., without mutations introducing frameshifts or in-frame stops) in the set of
782 Selenoprofiles predictions on genome assemblies in our main dataset. First, we
783 inspected the full set of Selenoprofiles predictions for the relevant gene family in an
784 extended set of NCBI genomes from the relevant clade. If the loss was not dismissed,
785 we then selected the protein sequence of the putatively lost gene from a closely related
786 organism, and searched it using TBLASTN at the NCBI web server [70] in all RNA data
787 available for the clade that included the focal species, i.e. TSAs, Expressed Sequence
788 Tags (ESTs), and Reference RNA sequences (refseq_rna). While absence of
789 transcriptomic data alone is not conclusive evidence of gene loss (e.g. due to tissue-

790 specific expression), we argue that it provides strong support when it is combined with
791 the concomitant absence from multiple genome assemblies in the same clade.

792 **Grx-domain analysis in TXNRD genes**

793 To investigate the distribution of Grx domains encoded in TXNRD genes, we first
794 collected a set of representative TXNRD protein sequences and scanned them for
795 conserved domains using InterProScan [71]. We extracted the sequences containing the
796 Grx domain and thus manually curated a high-confidence alignment of Grx domains of
797 TXNRDs. These were used as sequence profiles to search for homologous matches in
798 Ensembl vertebrate genomes using Selenoprofiles. Next, we cross-analyzed this data
799 with those TXNRD gene predictions in the same genomes classified as either *TXNRD1* or
800 *TXNRD3* by the Selenoprofiles *lineage* utility. We used a custom script based on
801 pyranges v1 [72] to assign Grx domain predictions to *TXNRD1* and *TXNRD3* genes, if
802 located within their extended genomic coordinates (±5,000 nucleotides). Unassigned
803 Grx domains were discarded. Finally, we visualized domain architecture and
804 evolutionary patterns by visualizing data through custom scripts based on the ETE3
805 library [63].

806 **Synteny analysis**

807 Through custom scripts based on the library pyranges v1 [72], we selected all gene
808 annotations that overlapped the genomic loci of the gene of interest (*TXNRD1*, *TXNRD3*,
809 and *SELENOW2/MIEN1*; Supplementary Figures S7, S8, S19, S20) predicted by
810 Selenoprofiles, extended on each side by 300,000 bp, in a set of representative species.
811 For these syntenic analyses, we used Ensembl v.115 annotations, upon realizing their
812 annotated gene names were improved compared to v.108, which we had used for the
813 rest of our work. We then used python code based on the graphics library pyrangeyes
814 v.1.0.7 (https://github.com/pyranges/pyrangeyes) to produce custom visualizations

815 wherein genes with the same name were colored accordingly, allowing to assess their
816 syntenic conservation. To inspect the homologous region in species lacking the gene of
817 interest, we performed an analogous process on a gene that consistently occurred in the
818 its neighborhood in other species (*CHST11* for *TXNRD1*; *TRPM* for *SELENOW2*; *RSPH4A*
819 for *SELENOW1*). Synteny representations were manually collated next to the species tree
820 using the Inkscape software.

821 **Gene nomenclature**

822 We strived to assign gene names as consistent as possible with existing resources and
823 literature. To do so, we consulted the relevant publications on vertebrate selenoprotein
824 evolution and nomenclature [73], as well as current annotations at Uniprot, Ensembl, and
825 NCBI. The reference recommendation for gene names is to derive them from the
826 corresponding mammalian orthologs [73]. The genes emerged from non-mammalian
827 duplications —many of them previously unreported— were named by appending
828 alphabetical or numerical suffixes to the source gene, e.g. *GPX4b*, *SELENOW2*. To
829 enhance readability, suffixes are written in lowercase throughout this work. To avoid
830 ambiguity, when creating new gene names we systematically skipped the suffixes "a"
831 and "1" (in favor of "b" or "2"), because these are inconsistently used in the literature to
832 refer to non-mammalian orthologs of single-copy mammalian genes; e.g. *SELENOW* is
833 sometimes referred to as *SELENOW1* in species that also possess the paralog
834 *SELENOW2*.

835 **Regression models of selenoproteome and proteome sizes**

836 Regression models were implemented in R, using libraries *stats* for linear models and
837 likelihood ratio tests, and *phylolm* for phylogenetic generalized least squares. A first
838 analysis with zero-intercept linear models (Figure 10) was run with 314 annotated species
839 (Ensembl v.108), used to estimate proteome size (number of unique protein-coding

genes), whereas Selenoprofiles was used to estimate selenoproteome size. A second analysis (Table 1) was run with the species subset that had a fully resolved taxonomic tree available in Ensembl Compara (263 genomes), using either linear models or phylogenetic generalized least squares; refer to the text and Table 1 for further details.

DECLARATIONS

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

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

Competing interests

The authors declare that they have no competing interests.

Funding

MM is funded by grants RYC2019-027746-I, PID2020-115122GA-I00, PID2023-147164NB-I00, and JL-F by CNS2022-135805 and PID2022-137753NA-I00, all these funded by MICIU/AEI /10.13039/501100011033 and by "ESF Investing in your future", FEDER, UE. MM and JL-F received support from Comissió Interdepartamental de Recerca i Innovació Tecnològica (2021SGR00279).

863 ## Authors' contributions

864 MT analysed all the data, performed the evolutionary analyses, prepared all the figures
865 and wrote the draft of the manuscript. JL-F contributed to data interpretation and revised
866 the manuscript. MM supervised the project, secured funding, provided expert
867 interpretation of the results and was a major contributor in writing the manuscript.

868 ## Acknowledgements

869 Not applicable.

870 # REFERENCES

871 1. Labunskyy VM, Hatfield DL, Gladyshev VN. Selenoproteins: Molecular Pathways and
872 Physiological Roles. Physiol Rev. 2014;94:739.
873 https://doi.org/10.1152/PHYSREV.00039.2013.

874 2. Mariotti M, Gladyshev VN. Selenocysteine-containing proteins. Redox Chemistry and
875 Biology of Thiols. 2022;:405–21. https://doi.org/10.1016/B978-0-323-90219-9.00012-1.

876 3. Fairweather-Tait SJ, Bao Y, Broadley MR, Collings R, Ford D, Hesketh JE, et al. Selenium
877 in human health and disease. Antioxid Redox Signal. 2011;14:1337–83.
878 https://doi.org/10.1089/ARS.2010.3275.

879 4. Howard MT, Copeland PR. New Directions for Understanding the Codon Redefinition
880 Required for Selenocysteine Incorporation. Biol Trace Elem Res. 2019;192:18–25.
881 https://doi.org/10.1007/S12011-019-01827-Y.

882 5. Copeland PR, Howard MT. Ribosome Fate during Decoding of UGA-Sec Codons. Int J
883 Mol Sci. 2021;22. https://doi.org/10.3390/IJMS222413204.

884 6. Manta B, Makarova NE, Mariotti M. The selenophosphate synthetase family: A review.
885 Free Radic Biol Med. 2022;192:63–76.
886 https://doi.org/10.1016/J.FREERADBIOMED.2022.09.007.

887 7. Flohé L. The labour pains of biochemical selenology: The history of selenoprotein
888 biosynthesis. Biochimica et Biophysica Acta (BBA) - General Subjects. 2009;1790:1389–
889 403. https://doi.org/10.1016/j.bbagen.2009.03.031.

890 8. Davy T, Castellano S. The genomics of selenium: Its past, present and future. Biochim
891 Biophys Acta Gen Subj. 2018;1862:2427–32.
892 https://doi.org/10.1016/j.bbagen.2018.05.020.

893 9. Kryukov G V., Castellano S, Novoselov S V., Lobanov A V., Zehtab O, Guigó R, et al.
894 Characterization of mammalian selenoproteomes. Science. 2003;300:1439–43.
895 https://doi.org/10.1126/SCIENCE.1083516.

896 10. Castellano S, Novoselov S V., Kryukov G V., Lescure A, Blanco E, Krol A, et al.
897 Reconsidering the evolution of eukaryotic selenoproteins: a novel nonmammalian family
898 with scattered phylogenetic distribution. EMBO Rep. 2004;5:71–7.
899 https://doi.org/10.1038/SJ.EMBOR.7400036.

900 11. Castellano S, Lobanov A V., Chapple C, Novoselov S V., Albrecht M, Hua D, et al.
901 Diversity and functional plasticity of eukaryotic selenoproteins: Identification and
902 characterization of the SelJ family. Proc Natl Acad Sci U S A. 2005;102:16188–93.
903 https://doi.org/10.1073/PNAS.0505146102/SUPPL_FILE/05146FIG9.JPG.

904 12. Novoselov S V., Hua D, Lobanov A V., Gladyshev VN. Identification and
905 characterization of Fep15, a new selenocysteine-containing member of the Sep15 protein
906 family. Biochem J. 2006;394 Pt 3:575–9. https://doi.org/10.1042/BJ20051569.

907 13. Shchedrina VA, Novoselov S V., Malinouski MY, Gladyshev VN. Identification and
908 characterization of a selenoprotein family containing a diselenide bond in a redox motif.
909 Proc Natl Acad Sci U S A. 2007;104:13919–24.
910 https://doi.org/10.1073/PNAS.0703448104/SUPPL_FILE/03448TABLE4.PDF.

911 14. Jiang L, Ni J, Liu Q. Evolution of selenoproteins in the metazoan. BMC Genomics.
912 2012;13. https://doi.org/10.1186/1471-2164-13-446.

913 15. Jiang L, Liu Q, Ni J. In silico identification of the sea squirt selenoproteome. BMC
914 Genomics. 2010;11:289. https://doi.org/10.1186/1471-2164-11-289.

915 16. Mariotti M, Ridge PG, Zhang Y, Lobanov A V., Pringle TH, Guigo R, et al. Composition
916 and evolution of the vertebrate and mammalian selenoproteomes. PLoS One. 2012;7.
917 https://doi.org/10.1371/JOURNAL.PONE.0033066.

918 17. Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, et al. Earth
919 BioGenome Project: Sequencing life for the future of life. Proc Natl Acad Sci U S A.
920 2018;115:4325–33. https://doi.org/10.1073/PNAS.1720115115.

921 18. Mc Cartney AM, Formenti G, Mouton A, De Panis D, Marins LS, Leitão HG, et al. The
922 European Reference Genome Atlas: piloting a decentralised approach to equitable
923 biodiversity genomics. npj Biodiversity 2024 3:1. 2024;3:1–17.
924 https://doi.org/10.1038/s44185-024-00054-6.

925 19. Ticó M, Sullivan E, Guigó R, Mariotti M. Overcoming the widespread flaws in the
926 annotation of vertebrate selenoprotein genes in public databases. PLoS Comput Biol.
927 2026;22:e1013885. https://doi.org/10.1371/journal.pcbi.1013885.

928 20. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. EnsemblCompara
929 GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome
930 Res. 2009;19:327–35. https://doi.org/10.1101/GR.073585.107.

931 21. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et
932 al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology
933 resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019;47:D309–
934 14. https://doi.org/10.1093/NAR/GKY1085.

935 22. Mariotti M, Santesmasses D, Capella-Gutierrez S, Mateo A, Arnan C, Johnson R, et al.
936 Evolution of selenophosphate synthetases: emergence and relocation of function
937 through independent duplications and recurrent subfunctionalization. Genome Res.
938 2015;25:1256–67. https://doi.org/10.1101/GR.190538.115.

939 23. Zhang Y, Jin J, Huang B, Ying H, He J, Jiang L. Selenium Metabolism and Selenoproteins
940 in Prokaryotes: A Bioinformatics Perspective. Biomolecules 2022, Vol 12, Page 917.
941 2022;12:917. https://doi.org/10.3390/BIOM12070917.

942 24. Botero-Castro F, Figuet E, Tilak MK, Nabholz B, Galtier N. Avian Genomes Revisited:
943 Hidden Genes Uncovered and the Rates versus Traits Paradox in Birds. Mol Biol Evol.
944 2017;34:3123–31. https://doi.org/10.1093/MOLBEV/MSX236.

945 25. Chapple CE, Guigó R. Relaxation of selective constraints causes independent
946 selenoprotein extinction in insect genomes. PLoS One. 2008;3.
947 https://doi.org/10.1371/journal.pone.0002968.

948 26. Chipman AD, Ferrier DEK, Brena C, Qu J, Hughes DST, Schröder R, et al. The First
949 Myriapod Genome Sequence Reveals Conservative Arthropod Gene Content and
950 Genome Organisation in the Centipede Strigamia maritima. PLoS Biol. 2014;12:e1002005.
951 https://doi.org/10.1371/journal.pbio.1002005.

952 27. Mariotti M. Selenocysteine Extinctions in Insects. 2016;:113–40.
953 https://doi.org/10.1007/978-3-319-24244-6_5.

954 28. Toppo S, Vanin S, Bosello V, Tosatto SCE. Evolutionary and structural insights into the
955 multifaceted glutathione peroxidase (Gpx) superfamily. Antioxid Redox Signal.
956 2008;10:1501–13. https://doi.org/10.1089/ARS.2008.2057.

957 29. Xue Y, Chen L, Li B, Xiao J, Wang H, Dong C, et al. Genome-wide mining of gpx gene
958 family provides new insights into cadmium stress responses in common carp (Cyprinus
959 carpio). Gene. 2022;821:146291. https://doi.org/10.1016/J.GENE.2022.146291.

960 30. Darras VM, Van Herck SLJ. Iodothyronine deiodinase structure and function: from
961 ascidians to humans. J Endocrinol. 2012;215:189–206. https://doi.org/10.1530/JOE-12-
962 0204.

963 31. Lorgen M, Casadei E, Król E, Douglas A, Birnie MJ, Ebbesson LOE, et al. Functional
964 divergence of type 2 deiodinase paralogs in the Atlantic salmon. Curr Biol. 2015;25:936–
965 41. https://doi.org/10.1016/J.CUB.2015.01.074.

966 32. Arnér ESJ. Focus on mammalian thioredoxin reductases--important selenoproteins
967 with versatile functions. Biochim Biophys Acta. 2009;1790:495–526.
968 https://doi.org/10.1016/J.BBAGEN.2009.01.014.

969 33. Dou Q, Turanov AA, Mariotti M, Hwang JY, Wang H, Lee SG, et al. Selenoprotein
970 TXNRD3 supports male fertility via the redox regulation of spermatogenesis. Journal of
971 Biological Chemistry. 2022;298.
972 https://doi.org/10.1016/J.JBC.2022.102183/ATTACHMENT/F4C4EF7D-2A2A-49B7-
973 B3BB-B4E7E586A468/MMC3.PDF.

974 34. Huerta-Cepas J, Bueno A, Dopazo J, Gabaldón T. PhylomeDB: a database for genome-
975 wide collections of gene phylogenies. Nucleic Acids Res. 2008;36 suppl_1:D491–6.
976 https://doi.org/10.1093/NAR/GKM899.

977 35. Fata F, Gencheva R, Cheng Q, Lullo R, Ardini M, Silvestri I, et al. Biochemical and
978 structural characterizations of thioredoxin reductase selenoproteins of the parasitic
979 filarial nematodes Brugia malayi and Onchocerca volvulus. Redox Biol. 2022;51.
980 https://doi.org/10.1016/j.redox.2022.102278.

981 36. Melo LMN, Sabatier M, Ramesh V, Szylo KJ, Fraser CS, Pon A, et al. Selenoprotein O
982 Promotes Melanoma Metastasis and Regulates Mitochondrial Complex II Activity.
983 Cancer Res. 2025;85:942–55. https://doi.org/10.1158/0008-5472.CAN-23-2194.

984 37. Sreelatha A, Yee SS, Lopez VA, Park BC, Kinch LN, Pilch S, et al. Protein AMPylation
985 by an Evolutionarily Conserved Pseudokinase. Cell. 2018;175:809-821.e19.
986 https://doi.org/10.1016/J.CELL.2018.08.046.

987 38. Tarrago L, Kaya A, Kim HY, Manta B, Lee BC, Gladyshev VN. The selenoprotein
988 methionine sulfoxide reductase B1 (MSRB1). Free Radic Biol Med. 2022;191:228–40.
989 https://doi.org/10.1016/J.FREERADBIOMED.2022.08.043.

990 39. Anouar Y, Lihrmann I, Falluel-Morel A, Boukhzar L. Selenoprotein T is a key player in
991 ER proteostasis, endocrine homeostasis and neuroprotection. Free Radic Biol Med.
992 2018;127:145–52. https://doi.org/10.1016/J.FREERADBIOMED.2018.05.076.

993 40. Jeong D won, Kim TS, Chung YW, Lee BJ, Kim IY. Selenoprotein W is a glutathione-
994 dependent antioxidant in vivo. FEBS Lett. 2002;517:225–8.
995 https://doi.org/10.1016/S0014-5793(02)02628-5.

996 41. Misra S, Lee TJ, Sebastian A, McGuigan J, Liao C, Koo I, et al. Loss of selenoprotein W
997 in murine macrophages alters the hierarchy of selenoprotein expression, redox tone, and
998 mitochondrial functions during inflammation. Redox Biol. 2023;59.
999 https://doi.org/10.1016/J.REDOX.2022.102571.

1000 42. Zhang X, Xiong W, Chen LL, Huang JQ, Lei XG. Selenoprotein V protects against
1001 endoplasmic reticulum stress and oxidative injury induced by pro-oxidants. Free Radic
1002 Biol Med. 2020;160:670–9. https://doi.org/10.1016/J.FREERADBIOMED.2020.08.011.

1003 43. Dikiy A, Novoselov S V., Fomenko DE, Sengupta A, Carlson BA, Cerny RL, et al. SelT,
1004 SelW, SelH, and Rdx12: Genomics and Molecular Insights into the Functions of
1005 Selenoproteins of a Novel Thioredoxin-like Family†. Biochemistry. 2007;46:6871–82.
1006 https://doi.org/10.1021/BI602462Q.

1007 44. Baclaocos J, Santesmasses D, Mariotti M, Bierła K, Vetick MB, Lynch S, et al.
1008 Processive Recoding and Metazoan Evolution of Selenoprotein P: Up to 132 UGAs in
1009 Molluscs. J Mol Biol. 2019;431:4381–407. https://doi.org/10.1016/J.JMB.2019.08.007.

1010 45. Yim SH, Everley RA, Schildberg FA, Lee SG, Orsi A, Barbati ZR, et al. Role of Selenof
1011 as a Gatekeeper of Secreted Disulfide-Rich Glycoproteins. Cell Rep. 2018;23:1387–98.
1012 https://doi.org/10.1016/J.CELREP.2018.04.009.

1013 46. Flowers B, Bochnacka O, Poles A, Diamond AM, Kastrati I. Distinct Roles of SELENOF
1014 in Different Human Cancers. Biomolecules. 2023;13.
1015 https://doi.org/10.3390/BIOM13030486.

1016 47. Zigrossi A, Hong LK, Ekyalongo RC, Cruz-Alvarez C, Gornick E, Diamond AM, et al.
1017 SELENOF is a new tumor suppressor in breast cancer. Oncogene 2021 41:9. 2022;41:1263–
1018 8. https://doi.org/10.1038/s41388-021-02158-w.

1019 48. Bertz M, Kühn K, Koeberle SC, Müller MF, Hoelzer D, Thies K, et al. Selenoprotein H
1020 controls cell cycle progression and proliferation of human colorectal cancer cells. Free
1021 Radic Biol Med. 2018;127:98–107.
1022 https://doi.org/10.1016/J.FREERADBIOMED.2018.01.010.

1023 49. Cox AG, Tsomides A, Kim AJ, Saunders D, Hwang KL, Evason KJ, et al. Selenoprotein
1024 H is an essential regulator of redox homeostasis that cooperates with p53 in
1025 development and tumorigenesis. Proc Natl Acad Sci U S A. 2016;113:E5562–71.
1026 https://doi.org/10.1073/PNAS.1600204113.

1027 50. Wu RTY, Cao L, Chen BPC, Cheng WH. Selenoprotein H suppresses cellular
1028 senescence through genome maintenance and redox regulation. J Biol Chem.
1029 2014;289:34378–88. https://doi.org/10.1074/JBC.M114.611970.

1030 51. Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, et al. The Atlantic salmon
1031 genome provides insights into rediploidization. Nature 2016 533:7602. 2016;533:200–5.
1032 https://doi.org/10.1038/nature17164.

1033 52. Bianchi D, Borza R, De Zan E, Huelsz-Prince G, Gregoricchio S, Dekker M, et al.
1034 Zincore, an atypical coregulator, binds zinc finger transcription factors to control gene
1035 expression. Science. 2025;389. https://doi.org/10.1126/science.adv2861.

1036 53. Odunsi A, Kapitonova MA, Woodward G, Rahmani E, Ghelichkhani F, Liu J, et al.
1037 Selenoprotein K at the intersection of cellular pathways. Arch Biochem Biophys.
1038 2025;764. https://doi.org/10.1016/J.ABB.2024.110221.

1039 54. Shi Z, Han Z, Chen J, Zhou JC. Endoplasmic reticulum-resident selenoproteins and
1040 their roles in glucose and lipid metabolic disorders. Biochim Biophys Acta Mol Basis Dis.
1041 2024;1870. https://doi.org/10.1016/J.BBADIS.2024.167246.

1042 55. Chernorudskiy A, Varone E, Colombo SF, Fumagalli S, Cagnotto A, Cattaneo A, et al.
1043 Selenoprotein N is an endoplasmic reticulum calcium sensor that links luminal calcium
1044 levels to a redox activity. Proc Natl Acad Sci U S A. 2020;117:21288–98.
1045 https://doi.org/10.1073/PNAS.2003847117/-/DCSUPPLEMENTAL.

1046 56. Lobanov A V., Fomenko DE, Zhang Y, Sengupta A, Hatfield DL, Gladyshev VN.
1047 Evolutionary dynamics of eukaryotic selenoproteomes: Large selenoproteomes may
1048 associate with aquatic life and small with terrestrial life. Genome Biol. 2007;8:1–16.
1049 https://doi.org/10.1186/GB-2007-8-9-R198/FIGURES/7.

1050 57. Sarangi GK, Romagne F, Castellano S. Distinct Patterns of Selection in Selenium-
1051 Dependent Genes between Land and Aquatic Vertebrates. Mol Biol Evol. 2018;35:1744–
1052 56. https://doi.org/10.1093/MOLBEV/MSY070.

1053 58. Maroney MJ, Hondal RJ. Selenium versus sulfur: Reversibility of chemical reactions
1054 and resistance to permanent oxidation in proteins and nucleic acids. Free Radic Biol Med.
1055 2018;127:228–37. https://doi.org/10.1016/J.FREERADBIOMED.2018.03.035.

1056 59. Reich HJ, Hondal RJ. Why Nature Chose Selenium. ACS Chem Biol. 2016;11:821–41.
1057 https://doi.org/10.1021/ACSCHEMBIO.6B00031.

1058 60. Castellano S, Andrés AM, Bosch E, Bayes M, Guigó R, Clark AG. Low exchangeability
1059 of selenocysteine, the 21st amino acid, in vertebrate proteins. Mol Biol Evol.
1060 2009;26:2031–40. https://doi.org/10.1093/MOLBEV/MSP109.

1061 61. Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome Res.
1062 2010;20:1313–26. https://doi.org/10.1101/GR.101386.109.

1063 62. Santesmasses D, Mariotti M, Gladyshev VN. Bioinformatics of Selenoproteins.
1064 Antioxid Redox Signal. 2020;33:525–36. https://doi.org/10.1089/ARS.2020.8044.

1065 63. Hughes LC, Ortí G, Huang Y, Sun Y, Baldwin CC, Thompson AW, et al. Comprehensive
1066 phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic
1067 data. Proc Natl Acad Sci U S A. 2018;115:6249–54.
1068 https://doi.org/10.1073/PNAS.1719358115/SUPPL_FILE/PNAS.1719358115.SAPP.PDF.

1069 64. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al.
1070 Ensembl 2022. Nucleic Acids Res. 2022;50:D988–95.
1071 https://doi.org/10.1093/NAR/GKAB1049.

1072 65. Mariotti M, Guigó R. Selenoprofiles: profile-based scanning of eukaryotic genome
1073 sequences for selenoprotein genes. Bioinformatics. 2010;26:2656–63.
1074 https://doi.org/10.1093/BIOINFORMATICS/BTQ516.

66. Santesmasses D, Mariotti M, Guigó R. Selenoprofiles: A Computational Pipeline for Annotation of Selenoproteins. Methods in Molecular Biology. 2018;1661:17–28. https://doi.org/10.1007/978-1-4939-7258-6_2.

67. Sievers F, Higgins DG. Clustal Omega for making accurate alignments of many protein sequences. Protein Sci. 2018;27:135–45. https://doi.org/10.1002/pro.3290.

68. Wong TKF, Ly-Trong N, Ren H, Baños H, Roger AJ, Susko E, et al. IQ-TREE 3: Phylogenomic Inference Software using Complex Evolutionary Models. 2025. https://doi.org/10.32942/X2P62N.

69. Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol Biol Evol. 2016;33:1635. https://doi.org/10.1093/MOLBEV/MSW046.

70. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10. https://doi.org/10.1186/1471-2105-10-421.

71. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30:1236–40. https://doi.org/10.1093/BIOINFORMATICS/BTU031.

72. Stovner EB, Ticó M, Muñoz Del Campo E, Pallarès-Albanell J, Chawla K, Saetrom P, et al. Pyranges v1: a Python framework for ultrafast sequence interval operations. bioRxiv. 2025;:2025.12.11.693639. https://doi.org/10.64898/2025.12.11.693639.

73. Regina BF, Gladyshev VN, Arnér ES, Berry MJ, Bruford EA, Burk RF, et al. Selenoprotein gene nomenclature. Journal of Biological Chemistry. 2016;291:24036–40. https://doi.org/10.1074/jbc.M116.756155.

1098 74. Yu D, Ren Y, Uesaka M, Beavan AJS, Muffato M, Shen J, et al. Hagfish genome
1099 elucidates vertebrate whole-genome duplication events and their evolutionary
1100 consequences. Nat Ecol Evol. 2024;8:519–35. https://doi.org/10.1038/s41559-023-
1101 02299-z.

1102 75. Nakatani Y, Shingate P, Ravi V, Pillai NE, Prasad A, McLysaght A, et al. Reconstruction
1103 of proto-vertebrate, proto-cyclostome and proto-gnathostome genomes provides new
1104 insights into early vertebrate evolution. Nature Communications 2021 12:1. 2021;12:4489-
1105 . https://doi.org/10.1038/s41467-021-24573-z.

1106 76. Dehal P, Boore JL. Two Rounds of Whole Genome Duplication in the Ancestral
1107 Vertebrate. PLoS Biol. 2005;3:e314. https://doi.org/10.1371/journal.pbio.0030314.

1108 77. Qi M, Clark J, Moody ERR, Pisani D, Donoghue PCJ. Molecular Dating of the Teleost
1109 Whole Genome Duplication (3R) Is Compatible With the Expectations of Delayed
1110 Rediploidization. Genome Biol Evol. 2024;16. https://doi.org/10.1093/gbe/evae128.

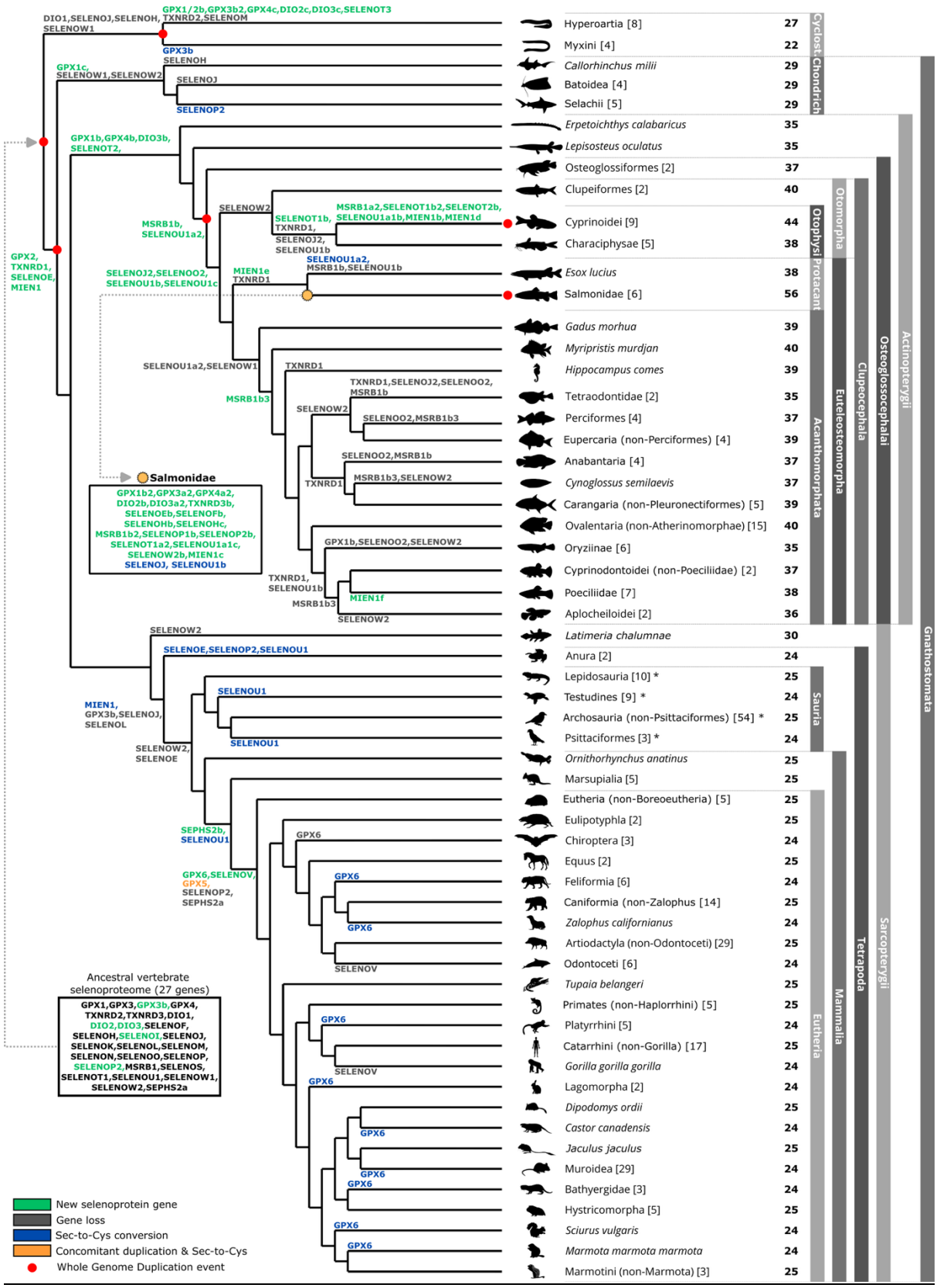
1111 78. Macqueen DJ, Johnston IA. A well-constrained estimate for the timing of the salmonid
1112 whole genome duplication reveals major decoupling from species diversification.
1113 Proceedings of the Royal Society B: Biological Sciences. 2014;281:20132881.
1114 https://doi.org/10.1098/rspb.2013.2881.

1115 79. Xu P, Xu J, Liu G, Chen L, Zhou Z, Peng W, et al. The allotetraploid origin and
1116 asymmetrical genome evolution of the common carp Cyprinus carpio. Nature
1117 Communications 2019 10:1. 2019;10:4625-. https://doi.org/10.1038/s41467-019-12644-1.

1118

1119

1120 ## FIGURES, TABLES AND ADDITIONAL FILES



1121

1122 ***Figure 1. Evolution of the vertebrate selenoproteome.*** *The evolutionary events shaping*

1123 *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 artifacts 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; "Cyclost." Stands for Cyclostomata, and "Protacant" for Protacanthopterygii. Red dots mark whole genome duplications (WGD) reported in literature [74–79]. Animal silhouettes were sourced from https://www.phylopic.org.*

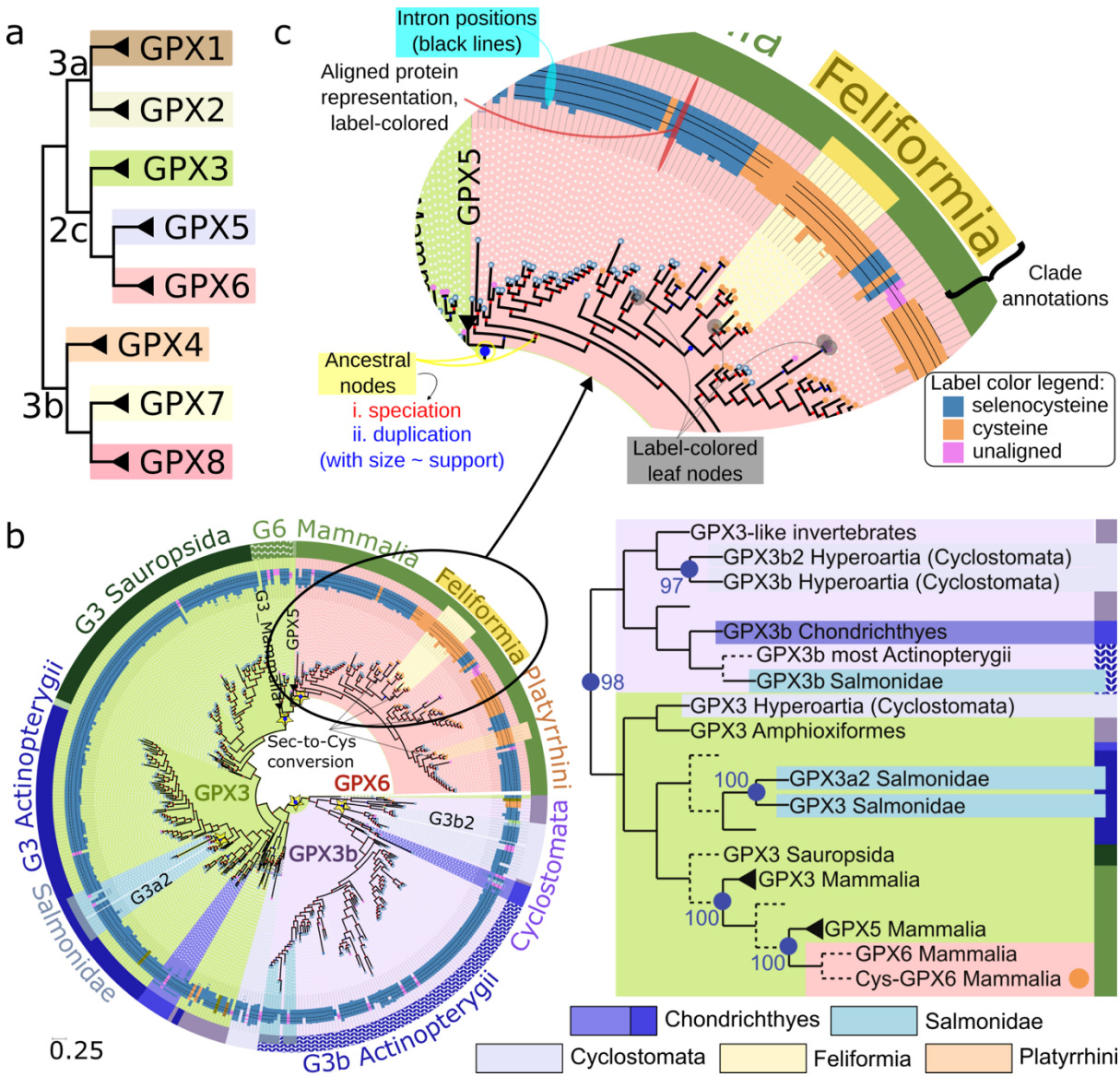
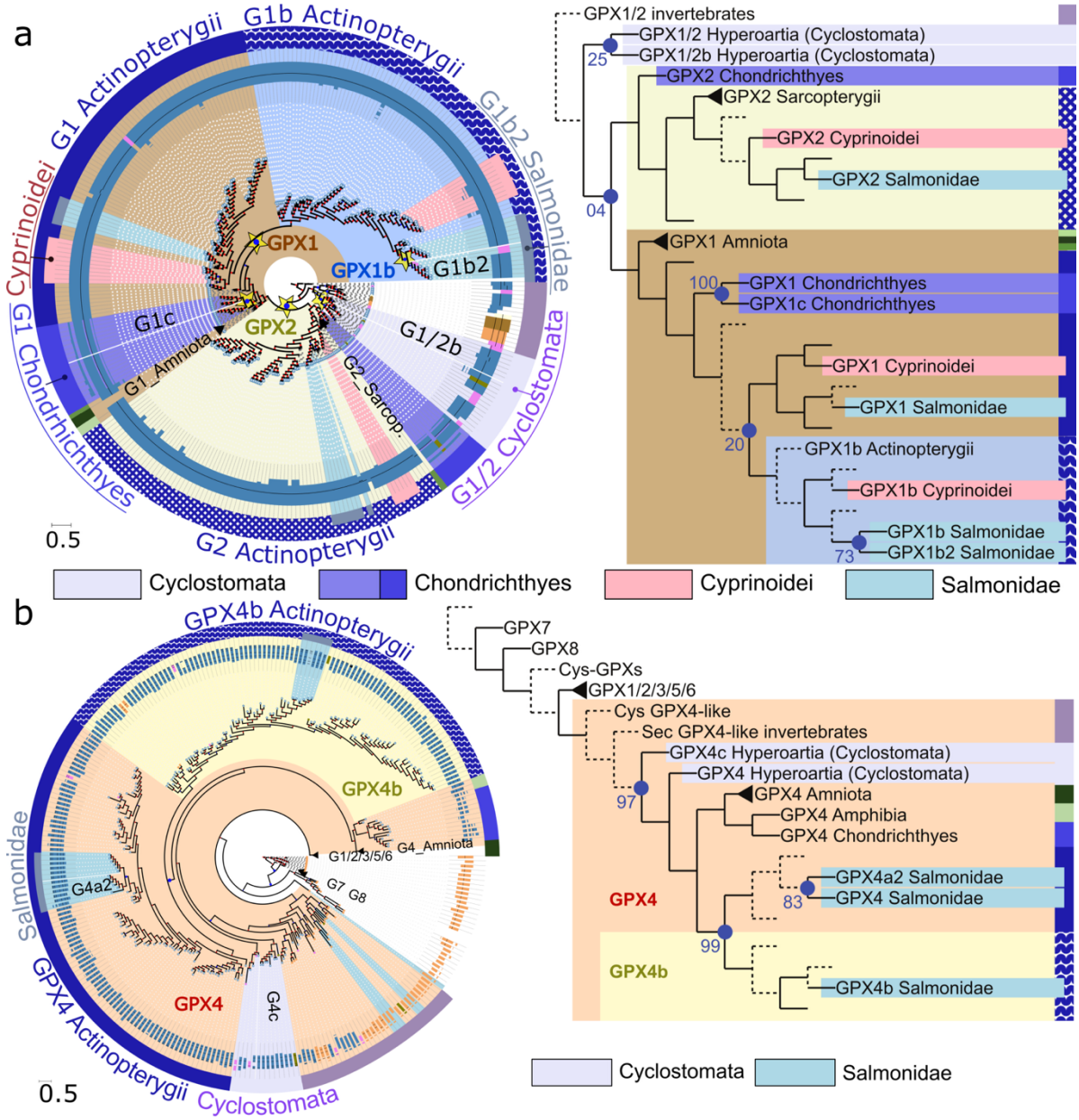
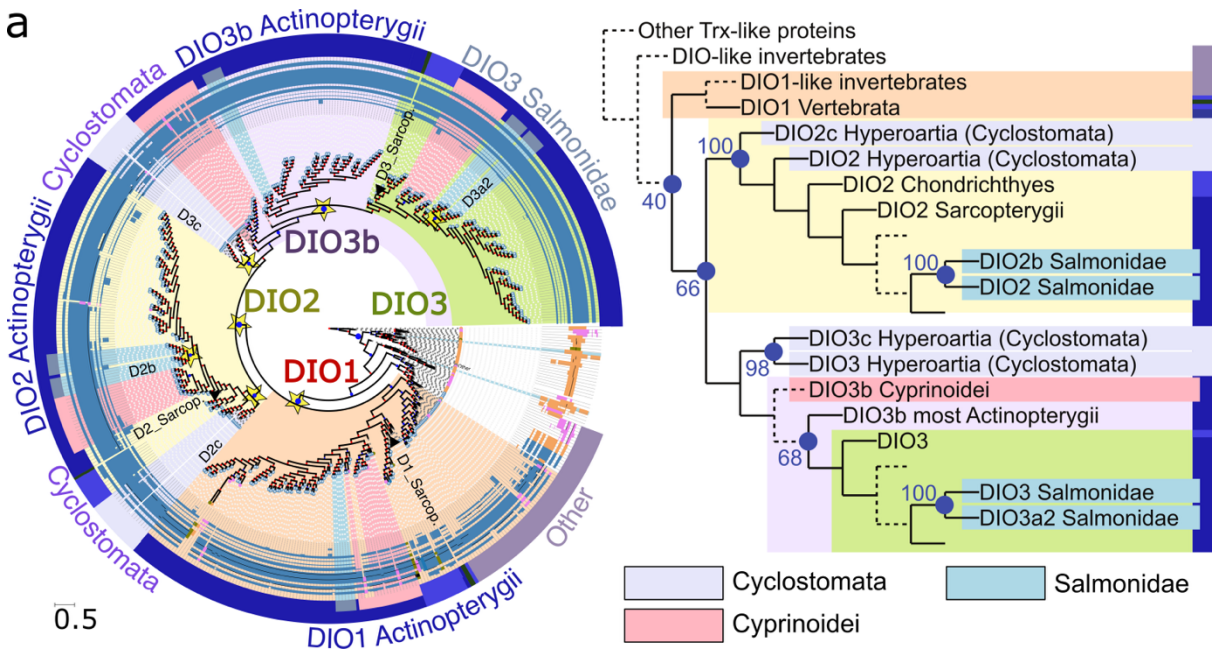


Figure 2. Phylogeny of the GPX family. *The GPX gene tree was computed collating all sequences of this family, but it is split in different representations for visualization purposes.* ***a)*** *Topology of the GPX family, split into three clades shown in panels b, Figure 3a and 3b.* ***b)*** *Phylogeny of GPX3/5/6 clade. To the right of the main tree, a summarized tree representation is shown, displaying the evolutionary events identified in the full tree, with selected clades collapsed for visualization. Dotted lines in the summary tree represent paraphyletic, non-collapsible branches. Duplication events are indicated by blue nodes labeled with their UFB support values. The legend below shows the color scheme used to represent the taxonomic groups of the species highlighted in the tree. Note that certain monophyletic clades of the gene tree were collapsed (e.g.*

see GPX5). The full expanded (non-collapsed) gene tree is available as Supplementary Figure S3. ***c****) Zoomed section of the GPX6 gene tree, used here to illustrate this form of representation, abundantly used throughout this paper. Each leaf corresponds to a putative gene whose protein sequence was included in the 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. blue for Sec-UGA, orange for Cys, pink 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 Mammalia). 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. The nodes that define duplication events deemed genuine via manual analysis are marked with a star. Branch lengths represent the expected number of substitutions per site and were estimated by maximum likelihood under the best-fitting substitution model in IQ-TREE.*


a
G1 Actinopterygii
G1b Actinopterygii
G1b2 Salmonidae
G1/2 Cyclostomata
G2 Actinopterygii
G1 Chondrichthyes
Cyprinoidei
GPX1
GPX1b
G1b2
GPX2
G1c
G1_Amniota
G1/2b
G2_Sarcop.
0.5
GPX1/2 invertebrates
GPX1/2 Hyperoartia (Cyclostomata)
GPX1/2b Hyperoartia (Cyclostomata)
25
GPX2 Chondrichthyes
GPX2 Sarcopterygii
GPX2 Cyprinoidei
GPX2 Salmonidae
04
GPX1 Amniota
100
GPX1 Chondrichthyes
GPX1c Chondrichthyes
GPX1 Cyprinoidei
GPX1 Salmonidae
20
GPX1b Actinopterygii
GPX1b Cyprinoidei
GPX1b Salmonidae
73
GPX1b2 Salmonidae
Cyclostomata
Chondrichthyes
Cyprinoidei
Salmonidae
b
GPX4b Actinopterygii
Salmonidae
GPX4 Actinopterygii
Cyclostomata
GPX4b
G4_Amniota
G1/2/3/5/6
G7
G8
G4a2
GPX4
G4c
0.5
GPX7
GPX8
Cys-GPXs
GPX1/2/3/5/6
Cys GPX4-like
Sec GPX4-like invertebrates
GPX4c Hyperoartia (Cyclostomata)
GPX4 Hyperoartia (Cyclostomata)
97
GPX4 Amniota
GPX4 Amphibia
GPX4 Chondrichthyes
GPX4a2 Salmonidae
83
GPX4 Salmonidae
GPX4
99
GPX4b
GPX4b Salmonidae
Cyclostomata
Salmonidae
c
10
20
30
40
GPX6.Loxodonta_africana/87-127 GEEYIQFKQYVGKHVLFVNVATYUGLTAQYPELNALQEELK
GPX6.Equus_caballus/50-90 GEEYIQFKQYAGKHVLFINVATYUGLTAQYPELNALQEELK
GPX6.Sus_scrofa/87-127 GEEYIPFKQYAGKHVLFVNVATYUGLTAQYPELNALQEELK
GPX6.Panthera_pardus/23-63 DGDYIQFKQYAGKVVLFVNVASYCGLTAQYPELNALQDELK
GPX6.Panthera_tigris_altaica/57-97 DGDYIQFKQYAGKVVLFVNVASYCGLTAQYPELNALQDELK
GPX6.Mustela_putorius_furo/29-69 EEEYVQFQRYAGKHVLFVNVATYUGLTAQYPELNALQEELK
GPX6.Zalophus_californianus/23-63 EEEHIPFKRYAGKHVLFVNVATYCGLTAQYPELNALQDELK
GPX6.Microcebus_murinus/138-177 GE-YIQFKQYAGKHVLFVNVATYUGLTVQYRELNALQEELR
GPX6.Cebus_imitator/51-90 E-EYIQFKQFAGKPILFVNVATYCSLTAQYPELNALQEELK
GPX6.Callithrix_jacchus/50-90 GEEYIQFKQFAGKPVLFVNVASYCSLTAQYPELNALQEELK
GPX6.Homo_sapiens/87-126 E-EYIQFKQFAGKHVLFVNVAAYUGLAAQYPELNALQEELK
GPX6.Pan_troglodytes/87-126 E-EYIQFKQFAGKHVLFVNVATYUGLAAQYPELNALQEELK
GPX6.Oryctolagus_cuniculus/50-90 GEEDIQFQQYAGKYVLFVNVATYCANTAQYPELNRLQEDLK
GPX6.Dipodomys_ordii/50-90 GEEYIKFQQYAGKHVLFVNVATFUGLTAQYPELNALQEELK
GPX6.Castor_canadensis/86-126 DKKYIQFKQYAGKHVLFVNVASYCDLTAQYPELNALQEDLQ
GPX6.Jaculus_jaculus/23-63 SEEYIQFKQYAGKHVLFVNVATYUGLTAQYPELNALQEELR
GPX6.Nannospalax_galili/50-90 GGEYIQFKQYEGKNILFVNVASFCGLTAQYPELNTLQEELK
GPX6.Cricetulus_griseus_crigri/50-90 GGEYVPFQQYAGKYILFVNVASFCGLTATYPELNTLQEELK
GPX6.Mus_musculus/50-90 GGEFVNFQQYAGKHILFVNVASFCGLTATYPELNTLQEELK
GPX6.Fukomys_damarensis/23-63 GQEYIQFTQYSGKHILFVNVATYCGLTAQYPELNALQDELH
GPX6.Heterocephalus_glaber_male/52-92 GEEYVQFTKYSGKHVLFVNLATYCGLTAQYPELNALQDELR
GPX6.Chinchilla_lanigera/44-84 GEEYVQFKQYAGKHVLFVNVATYUGLTAQYPELNALQDELR
GPX6.Cavia_porcellus/44-84 GEEYIQFKQFAGKHVLFVNVATYUGLTAQYPELNALQDELR
GPX6.Cavia_aperea/44-84 GEEYVQFKQFAGKHVLFVNVATYUGLTAQYPELNALQDELR
GPX6.Sciurus_vulgaris/50-90 DDKYIQFKQYAGKHVLFVNVATYCGLTAQYPELNALQEELK
GPX6.Marmota_marmota_marmota/81-121 GDEYIQFKNYAGKHVLFVNVATYCGLTAQYPELNALQEELK
GPX6.Spermophilus_dauricus/23-63 GEEYIQFKKYAGKHVLFVNVATYUGLTAQYPELNALQEELK
GPX6.Ictidomys_tridecemlineatus/23-63 GEEYIQFKKYAGKHVLFVNVATYUGLTAQYPELNALQEELK

1163 ***Figure 3. Phylogeny of the GPX family, continued from figure 2. a, b)*** *Gene tree*
1164 *visualization of the GPX1/2 and GPX4/7/8 clades, respectively. See full feature*
1165 *description in Figure 2c The full gene tree is available as Supplementary Figure S3.* ***f)***
1166 *Multiple sequence alignment of GPX6 proteins including representatives that underwent*
1167 *Cys conversions. Sec residues are shown in blue, while Cys residues at this position are*
1168 *in orange. The identifiers of GPX6-Cys homologs are highlighted.*



1169

1170 ***Figure 4. Phylogeny of the DIO family. a)*** *Gene tree visualization of the DIO family. See*
1171 *full feature description in Figure 2c. Expanded gene and species trees are available as*
1172 *Supplementary Figures S4 and S5.*

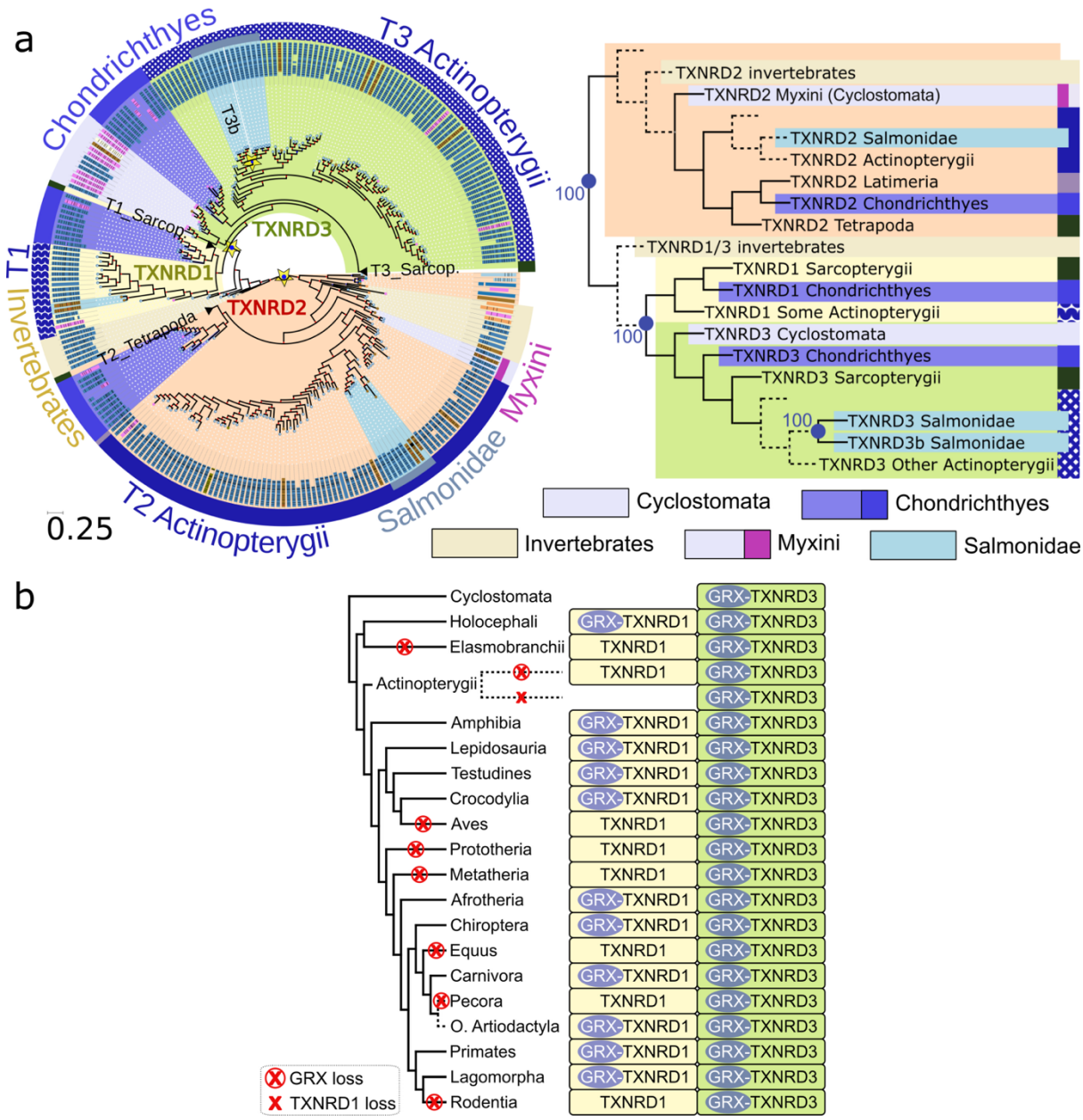
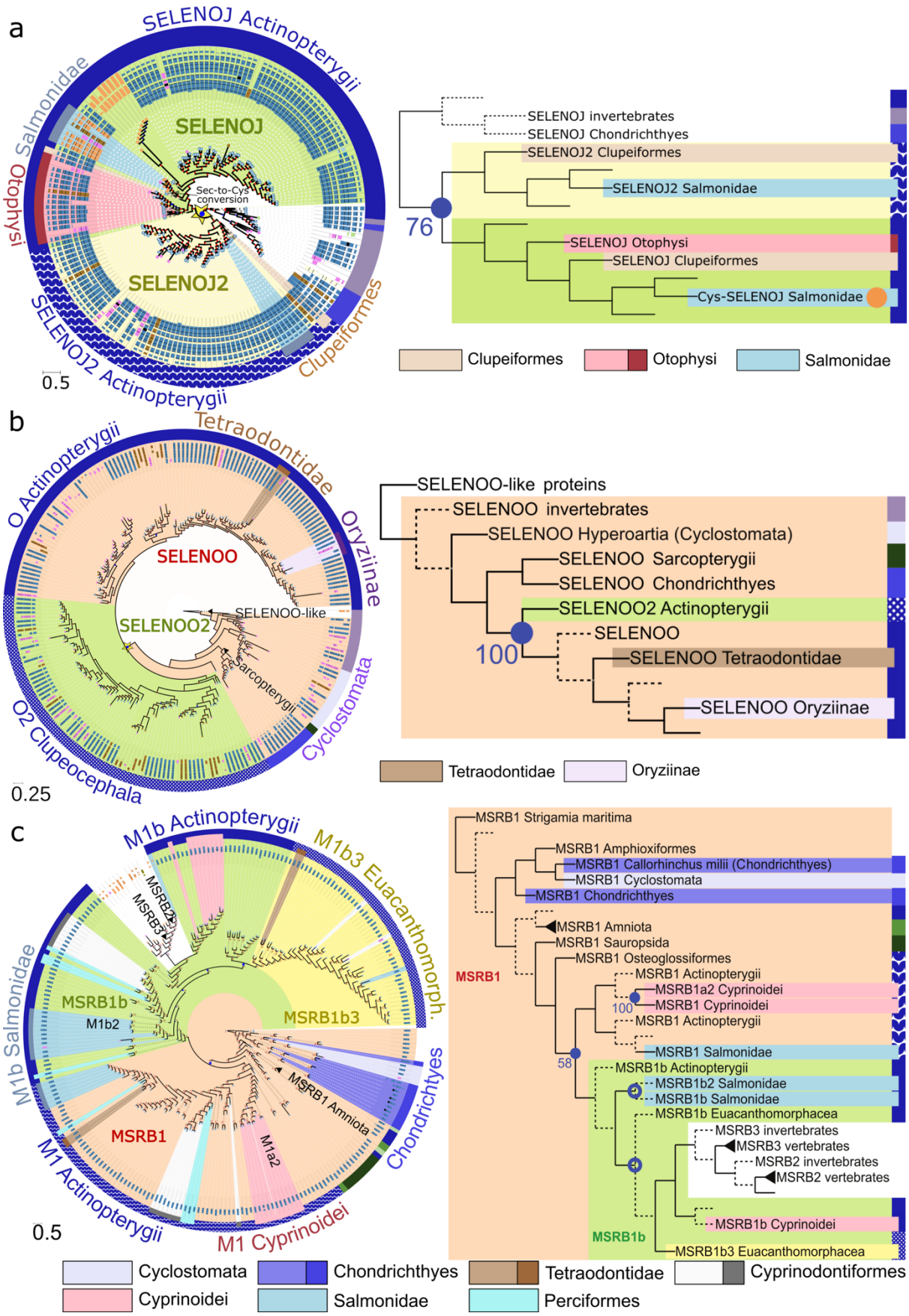
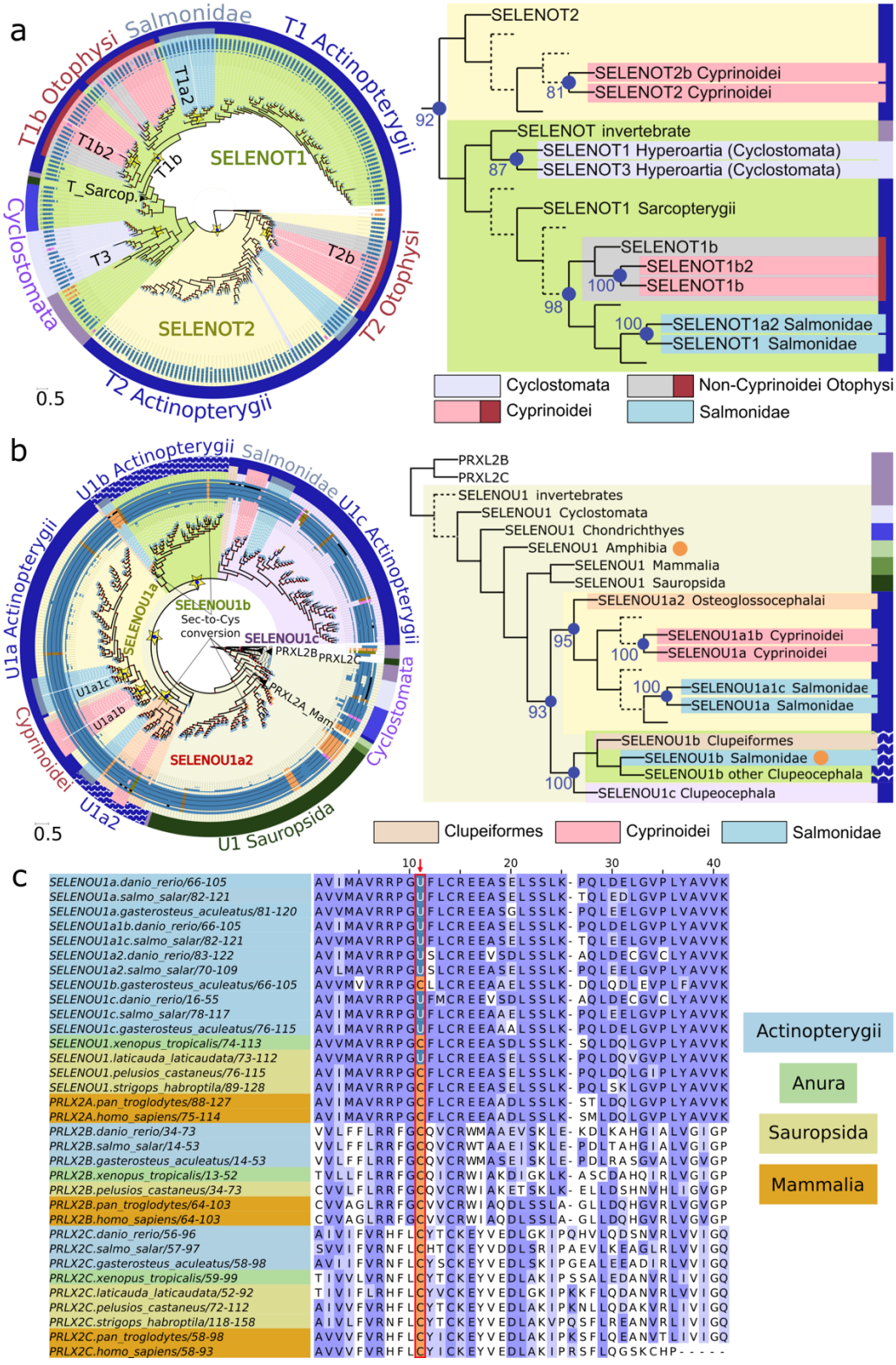


Figure 5. Phylogeny of the TXNRD family. a) *Gene tree visualization of the TXNRD family. See full feature description in Figure 2c. Expanded gene and species trees are available as Supplementary Figures S6, while Supplementary Figures S7 and S8 contain analyses of the synteny of TXNRD1 and TXNRD3 genes.* ***b)*** *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 Figure S9.*


a
SELENOJ Actinopterygii
Salmonidae
Otophysi
SELENOJ
Sec-to-Cys conversion
SELENOJ2
SELENOJ2 Actinopterygii
Clupeiformes
0.5
SELENOJ invertebrates
SELENOJ Chondrichthyes
SELENOJ2 Clupeiformes
SELENOJ2 Salmonidae
76
SELENOJ Otophysi
SELENOJ Clupeiformes
Cys-SELENOJ Salmonidae
Clupeiformes
Otophysi
Salmonidae
b
O Actinopterygii
Tetraodontidae
Oryziinae
SELENOO
SELENOO-like
SELENOO2
Sarcopterygii
Cyclostomata
O2 Clupeocephala
0.25
SELENOO-like proteins
SELENOO invertebrates
SELENOO Hyperoartia (Cyclostomata)
SELENOO Sarcopterygii
SELENOO Chondrichthyes
SELENOO2 Actinopterygii
SELENOO
100
SELENOO Tetraodontidae
SELENOO Oryziinae
Tetraodontidae
Oryziinae
c
M1b Actinopterygii
M1b3 Euacanthomorph.
MSRB2
MSRB3
M1b Salmonidae
MSRB1b
M1b2
MSRB1b3
MSRB1 Amniota
Chondrichtyes
MSRB1
M1a2
M1 Actinopterygii
M1 Cyprinoidei
0.5
MSRB1 Strigamia maritima
MSRB1 Amphioxiformes
MSRB1 Callorhinchus milii (Chondrichthyes)
MSRB1 Cyclostomata
MSRB1 Chondrichthyes
MSRB1 Amniota
MSRB1 Sauropsida
MSRB1 Osteoglossiformes
MSRB1
MSRB1 Actinopterygii
MSRB1a2 Cyprinoidei
100
MSRB1 Cyprinoidei
MSRB1 Actinopterygii
MSRB1 Salmonidae
58
MSRB1b Actinopterygii
MSRB1b2 Salmonidae
MSRB1b Salmonidae
MSRB1b Euacanthomorphacea
MSRB3 invertebrates
MSRB3 vertebrates
MSRB2 invertebrates
MSRB2 vertebrates
MSRB1b Cyprinoidei
MSRB1b
MSRB1b3 Euacanthomorphacea
Cyclostomata
Chondrichthyes
Tetraodontidae
Cyprinodontiformes
Cyprinoidei
Salmonidae
Perciformes

1181

Figure 6. Phylogeny of the SELENOJ, SELENOO, and MSRB families. a, b, c) *Gene tree visualizations of the SELENOJ, SELENOO, and MSRB families, respectively. See full feature description in Figure 2c. Expanded gene and species trees are available: Supplementary Figure S10, S11 (SELENOJ), S12, S13 (SELENOO), S14, S15 (MSRB).*



1187 ***Figure 7. Phylogeny of the SELENOT, SELENOU families. a, b)*** *Gene tree visualizations*
1188 *of the SELENOT and SELENOOU families, respectively. See full feature description in*
1189 *Figure 2c. Expanded gene trees are available as Supplementary Figures S16 (SELENOT)*
1190 *and S17 (SELENOU).* ***c)*** *SELENOU multiple sequence alignment. Species are color-*
1191 *labelled according to lineage. Sec residues are marked in blue and Cys residues in*
1192 *orange.*

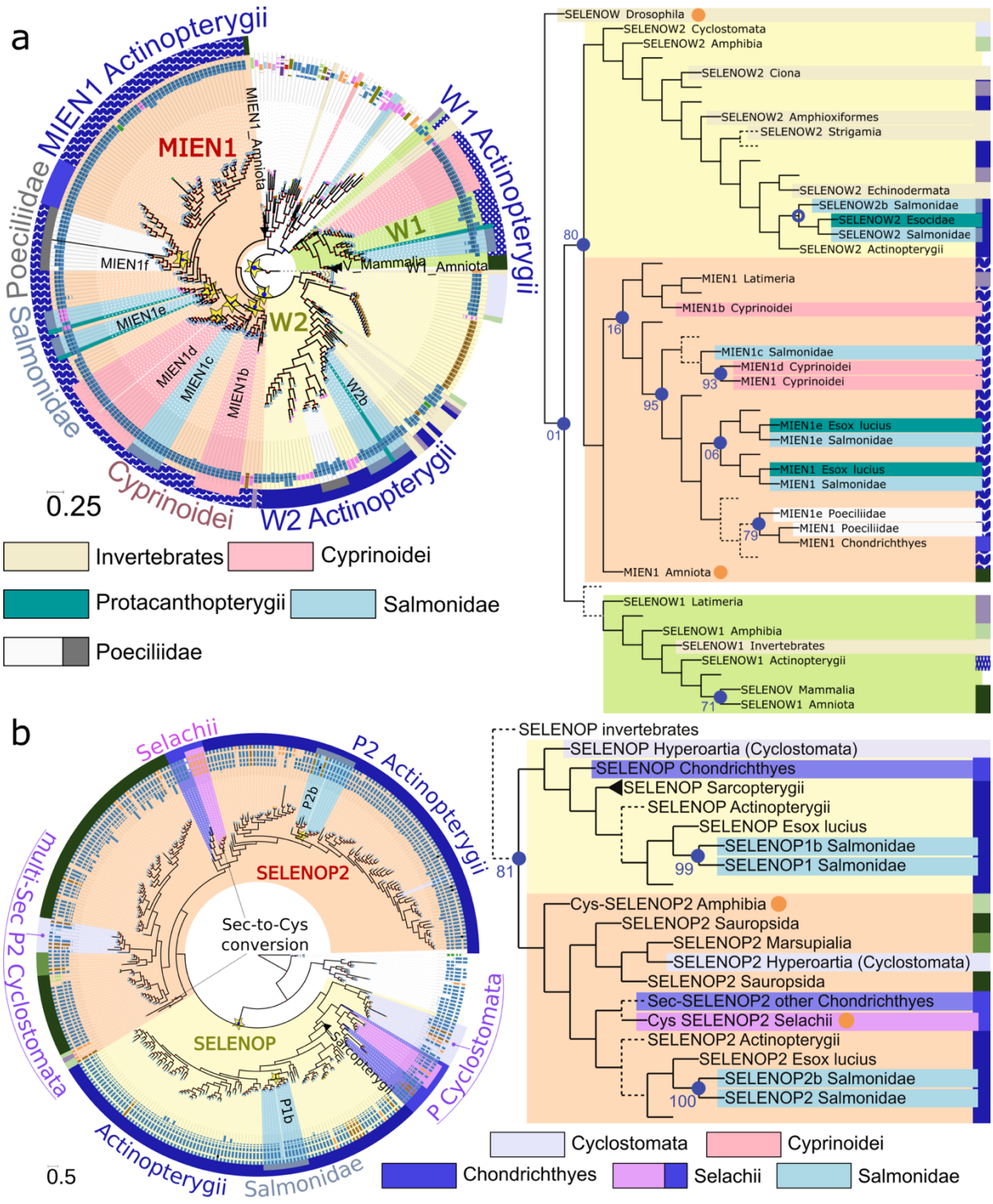


Figure 8. Phylogeny of the SELENOW and SELENOP families. a, b) *Gene tree visualizations of the SELENOW and SELENOOP families, respectively. For tree annotation and legend details, see Figure 2c. Expanded gene trees are available as Supplementary Figures S18 (SELENOW) and S22 (SELENOP).*

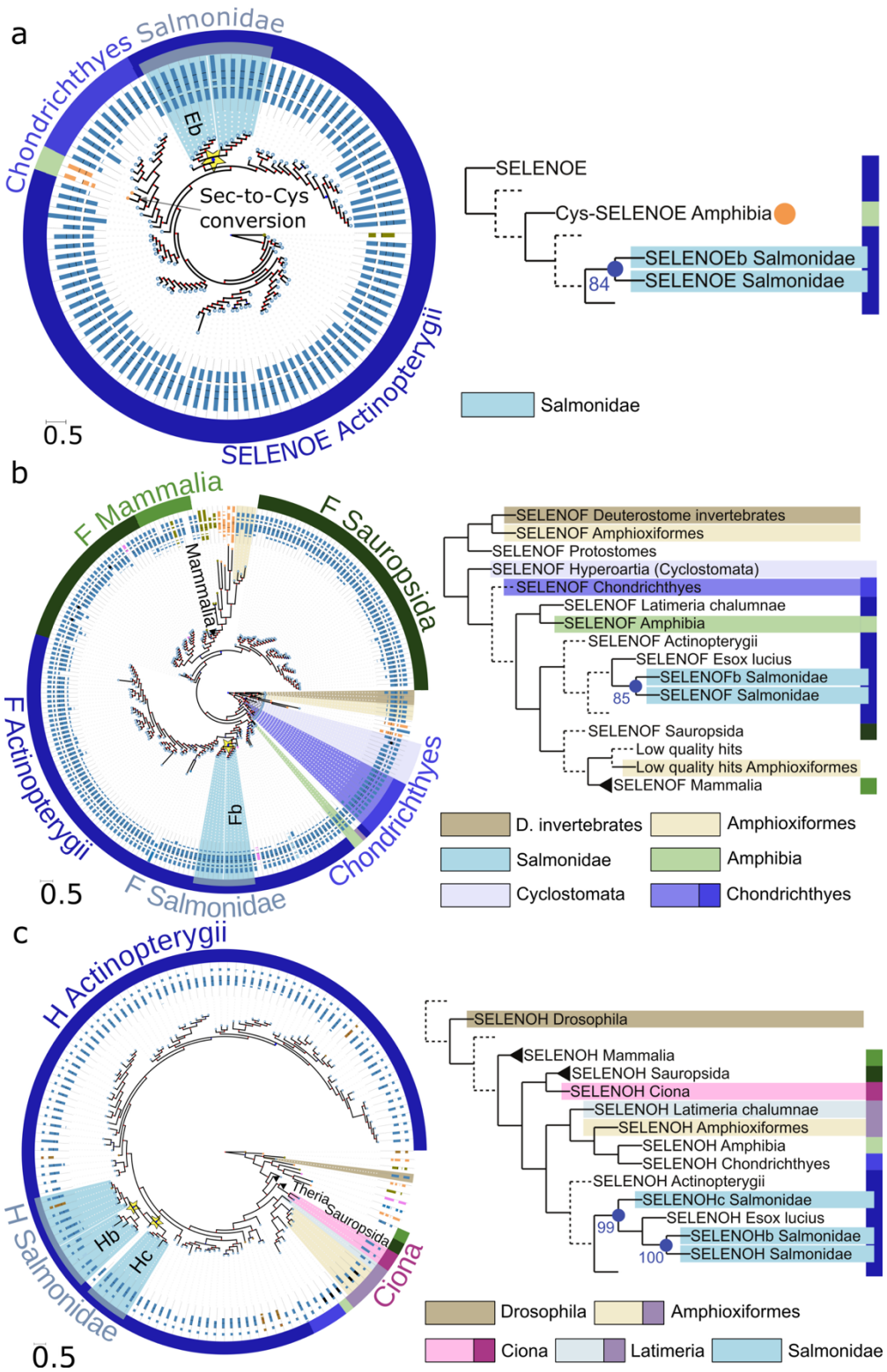


Figure 9. Phylogeny of SELENOE, SELENOF, and SELENOH. a, b, c) *Gene tree visualizations of the SELENOE, SELENOF, and SELENOH families, respectively. For tree annotation and legend details, see Figure 2c. Expanded gene trees are available as Supplementary Figure S23 (SELENOE), S24 (SELENOF), and S25 (SELENOH).*

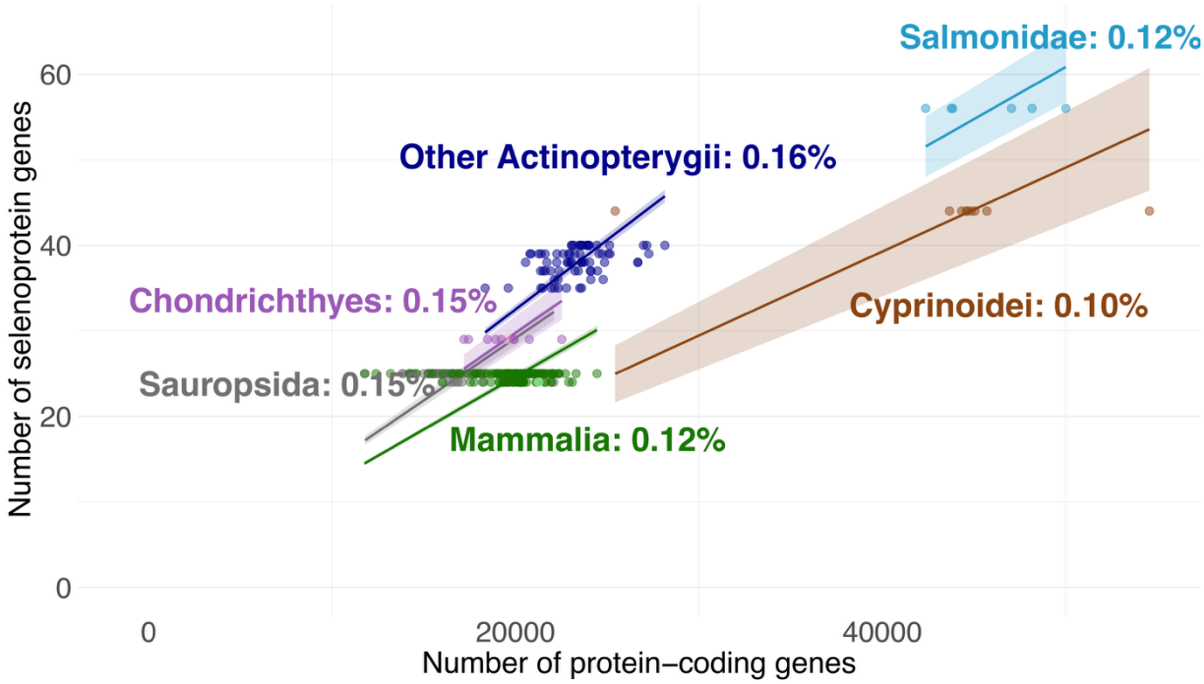


Figure 10. Genomic correlation between annotated proteome size and the number of predicted selenoproteins. *Each point represents a species and is color-labelled according to lineage. Trend lines corresponding to zero-intercept linear models fit separately for each lineage (Methods) are displayed. Shading indicates the 95% confidence interval on the slope. The colored text provides the color key and includes the average selenoproteome-to-proteome ratio as percentage.*

Formula	Model	Tetrapods aquatic?	H coefficient P-value	LRT p-value
S ~ P + H * P + 0	PGLS + BM	No	2.2e-16	2.51e-77
S ~ P + H * P + 0	PGLS + BM	Cetacea + Amphibia	2.11e-12	9.08e-68
S ~ P + H * P + 0	PGLS + BM	Cetacea only	2.2e-16	1.45e-72
S ~ P + H * P + 0	PGLS + OU	No	2.2e-16	2.51e-77
S ~ P + H * P + 0	PGLS + OU	Cetacea + Amphibia	2.11e-12	9.08e-68
S ~ P + H * P + 0	PGLS + OU	Cetacea only	2.2e-16	1.45e-72
S ~ P + H * P + 0	LM	No	2.2e-16	3.38e-184
S ~ P + H * P + 0	LM	Cetacea + Amphibia	2.2e-16	7.44e-88
S ~ P + H * P + 0	LM	Cetacea only	2.2e-16	9.11e-142
S ~ P + H + 0	PGLS + BM	No	8.24e-08	9.8e-69
S ~ P + H + 0	PGLS + BM	Cetacea + Amphibia	8.69e-07	2.45e-63

S ~ P + H + 0	PGLS + BM	Cetacea only	6.15e-07	1.24e-63
S ~ P + H + 0	PGLS + OU	No	8.24e-08	9.8e-69
S ~ P + H + 0	PGLS + OU	Cetacea + Amphibia	8.69e-07	2.45e-63
S ~ P + H + 0	PGLS + OU	Cetacea only	6.15e-07	1.24e-63
S ~ P + H + 0	LM	No	2.2e-16	2.19e-134
S ~ P + H + 0	LM	Cetacea + Amphibia	2.2e-16	1.26e-84
S ~ P + H + 0	LM	Cetacea only	2.2e-16	5.02e-106

Table 1. Analysis of selenoproteome-to-proteome ratio in aquatic vs terrestrial vertebrates. *The table contains results of different regression models fitted with data from 263 vertebrates (Methods). Two formulas were evaluated: an additive model (S ~ P + H + 0) and an interaction model (S ~ P + H*P + 0*), where S is the number of predicted selenoproteins (Selenoprofiles), P the number of annotated proteins (Ensembl), and H a binary variable for aquatic vs. terrestrial habitat. The "+0" denotes that the intercept was constrained to zero. Models included simple linear regression (LM) and phylogenetic generalized least squares (PGLS) under either Brownian motion (BM, neutrality) or Ornstein–Uhlenbeck (OU, balancing selection). All fish were considered aquatic; in some models, Cetacea and/or Amphibia were also classified as aquatic (see the third column). The fourth column reports the P-value of the habitat coefficient, and the fifth column the likelihood ratio test P-value comparing each model to its corresponding null model without H (S ~ P + 0).*

SUPPLEMENTARY DATA

Figure S1. Selenoproteome evolution across vertebrate species. *Each row represents a species and each column corresponds to a selenoprotein family. Rectangles represent individual Selenoprofiles predictions, color-coded according according to their Selenoprofiles labels, i.e. the codon found aligned to the Sec position: blue for Sec-UGA,*

1230 *orange for Cys, pink for unaligned). Selected lineages denoted with a colored*
1231 *background are labelled on the left.*

1232 ***Figure S2. Presence of genes within the SEPHS2 syntenic block across Sauropsidan***
1233 ***genome assemblies.*** *The species tree of Sauropsida (plus platypus, used as outgroup)*
1234 *is displayed on the left. Three columns display the presence of genes SEPSH2 (blue, third*
1235 *column) and the genes consistently located as its neighbors in tetrapods, G6PD (first*
1236 *column) and ikbkg (second column). The third column also shows the presence of*
1237 *SEPHS1 (purple), located in a different chromosomic location, here serving as positive*
1238 *control. Note the scattered lack of G6PD, ikbkg, and SEPHS2 from many bird genomes,*
1239 *attributed to imperfect quality of genome assemblies, as discussed in the text.*

1240 ***Figure S3. Gene tree of the Glutathione peroxidase (GPX) family.*** *Subfamilies are color-*
1241 *labeled as follows: GPX4 in orange, GPX4b in yellow, GPX3b in purple, GPX3 in green,*
1242 *GPX6 in pink, GPX5 in light purple, GPX2 in beige, GPX1 in brown, and GPX1b in blue.*
1243 *Refer to Figures 2 and 3 for details.*

1244 ***Figure S4. Gene tree of the Iodothyronine deiodinase (DIO) family.*** *Subfamilies are*
1245 *color-labeled as follows: DIO1 in orange, DIO2 in yellow, DIO3 in green and DIO3b in*
1246 *purple. See also Figures 2c (layout explanation) and 4 (compressed tree).*

1247 ***Figure S5. Iodothyronine deiodinase (DIO) evolution across Salmonidae species.*** *Each*
1248 *row represents a species and each column corresponds to a selenoprotein family.*
1249 *Nodes represent individual Selenoprofiles predictions and are color-coded according*
1250 *to the codon aligned at the Sec position: blue for Sec-UGA, orange for Cys.*

1251 ***Figure S6. Gene tree of the Thioredoxin reductase (TXNRD) family.*** *Subfamilies are*
1252 *color-labeled as follows: TXNRD1 in green, TXNRD2 in orange, TXNRD3 in yellow. See*
1253 *also Figures 2c (layout explanation) and 5 (compressed tree).*

1254 ***Figure S7. Analysis of synteny of TXNRD1 across vertebrates.*** The figure displays the
1255 genomic neighborhoods of *TXNRD1* and/or *CHST11* (see Methods) in representative
1256 genomes. Same-colored gene structures share the same annotated gene name. The
1257 species tree is shown on the left. Species highlighted in grey do not contain the *TXNRD1*
1258 gene.

1259 ***Figure S8. Analysis of synteny of TXNRD3 across vertebrates.*** The figure displays the
1260 genomic neighborhoods of *TXNRD3* and/or *CHST13* (see Methods) in representative
1261 genomes. Same-colored gene structures share the same annotated gene name. The
1262 species tree is shown on the left. Species highlighted in grey do not contain the *TXNRD1*
1263 gene.

1264 ***Figure S9. Vertebrate species tree representing the presence/absence of Grx domain in***
1265 ***TXNRD1/TXNRD3 sequences.*** *Colors are assigned according to the TXNRD gene tree.*

1266 ***Figure S10. Gene tree of the Selenoprotein J (SELENOJ) family.*** *Subfamilies are color-*
1267 *labeled as follows: SELENOJ in green and SELENOJ2 in yellow. See also Figures 2c*
1268 *(layout explanation) and 6 (compressed tree).*

1269 ***Figure S11. Selenoprotein J (SELENOJ) evolution across Actinopterygii species.*** *Each row*
1270 *represents a species and each column corresponds to a selenoprotein family. Nodes*
1271 *represent individual Selenoprofiles predictions and are color-coded according to their*
1272 *amino acid content at the UGA position: blue for Sec-containing sequences, orange for*
1273 *Cys-containing sequences, and pink for other classifications inferred by Selenoprofiles.*

1274 ***Figure S12. Gene tree of the Selenoprotein O (SELENOO) family.*** *Subfamilies are color-*
1275 *labeled as follows: SELENOO in green and SELENOO2 in orange. Nodes highlighted in*
1276 *brown and light purple indicate sequences from Tetraodontidae and Oryziinae species,*
1277 *respectively. See also Figures 2c (layout explanation) and 6 (compressed tree).*

1278 ***Figure S13. Selenoprotein O (SELENOO) evolution across Actinopterygii species.*** *Each*
1279 *row represents a species and each column corresponds to a selenoprotein family.*
1280 *Nodes represent individual Selenoprofiles predictions and are color-coded according*
1281 *to their amino acid content at the UGA position: blue for Sec-containing sequences,*
1282 *orange for Cys-containing sequences, and pink for other classifications inferred by*
1283 *Selenoprofiles.*

1284 ***Figure S14. Gene tree of the Methionine sulfoxide reductase B (MSRB) family.*** *See also*
1285 *Figures 2c (layout explanation) and 6 (compressed tree).*

1286 ***Figure S15. Methionine sulfoxide reductase B (MSRB) evolution across Actinopterygii***
1287 ***species.*** *Each row represents a species and each column corresponds to a*
1288 *selenoprotein family. Nodes represent individual Selenoprofiles predictions and are*
1289 *color-coded according to their amino acid content at the UGA position: blue for Sec-*
1290 *containing sequences, orange for Cys-containing sequences, and pink for other*
1291 *classifications inferred by Selenoprofiles.*

1292 ***Figure S16. Gene tree of the Selenoprotein T (SELENOT) family.*** *Subfamilies are color-*
1293 *labeled as follows: SELENOT1 in green and SELENOT2 in yellow. Nodes highlighted in*
1294 *light blue indicate sequences from Salmonidae species, while lightpink and grey*
1295 *indicate sequences from Cyprinoidei and Otophysi species, respectively. See also*
1296 *Figures 2c (layout explanation) and 7 (compressed tree).*

1297 ***Figure S17. Gene tree of the Selenoprotein U (SELENOU) family.*** *Subfamilies are color-*
1298 *labeled as follows: SELENOU1a in yellow, SELENOU1a2 in orange, SELENOU1b in green,*
1299 *SELENO1c in purple and Cysteine-homolog PRLX2A in beige. Nodes highlighted in light*
1300 *blue indicate sequences from Salmonidae species, while lightpink indicate sequences*
1301 *from Cyprinoidei. See also Figures 2c (layout explanation) and 7 (compressed tree).*

1302 ***Figure S18. Gene tree of the Selenoprotein V/W (SELENOV/W) family.*** *Subfamilies are*
1303 *color-labeled as follows: SELENOW1 from Actinopterygii species in green, SELENOW1*
1304 *from tetrapoda species in beige, SELENOV in purple, SELENOW2a in yellow and*
1305 *SELENOW2c in orange. Nodes highlighted in light blue indicate sequences from*
1306 *Salmonidae species, while light pink and light grey indicate sequences from Cyprinoidei*
1307 *and Cyprinodontoidei species, respectively. See also Figures 2c (layout explanation)*
1308 *and 8 (compressed tree).*

1309 ***Figure S19. Analysis of synteny of SELENOW1 across vertebrates.*** The figure displays the
1310 genomic neighborhoods of *SELENOW1* and/or *rsph4a* (see Methods) in representative
1311 genomes. Same-colored gene structures share the same annotated gene name. The
1312 species tree is shown on the left. Species highlighted in grey do not contain *SELENOW1*
1313 gene.

1314 ***Figure S20. Analysis of synteny of SELENOW2 across vertebrates.*** The figure displays
1315 the genomic neighborhoods of *SELENOW2* and/or *TRPM7* (see Methods) in
1316 representative genomes. Same-colored gene structures share the same annotated gene
1317 name. The species tree is shown on the left. Species highlighted in grey do not contain
1318 *SELENOW2* gene.

1319 ***Figure S21. Analysis of synteny of MIEN1 across vertebrates.*** The figure displays the
1320 genomic neighborhoods of *MIEN1* (see Methods) in representative genomes. Same-
1321 colored gene structures share the same annotated gene name. The species tree is shown
1322 on the left.

1323 ***Figure S22. Gene tree of the Selenoprotein P (SELENOP) family.*** *Nodes highlighted in*
1324 *light blue indicate sequences from Salmonidae species. See also Figures 2c (layout*
1325 *explanation) and 8 (compressed tree).*

1326 ***Figure S23. Gene tree of the Selenoprotein E (SELENOE) family.*** *See also Figures 2c*
1327 *(layout explanation) and 9 (compressed tree).*

1328 ***Figure S24. Gene tree of the Selenoprotein F (SELENOF) family.*** *Nodes highlighted in*
1329 *light blue indicate sequences from Salmonidae species See also Figures 2c (layout*
1330 *explanation) and 9 (compressed tree).*

1331 ***Figure S25. Gene tree of the Selenoprotein H (SELENOH) family.*** *Nodes highlighted in*
1332 *light blue indicate sequences from Salmonidae species. See also Figures 2c (layout*
1333 *explanation) and 9 (compressed tree).*

1334 ***Figure S26. Gene tree of the SEPHS family.*** *See also Figure 2c (layout explanation).*

1335 ***Table S1. Comparison of Glutathione Peroxidase (GPX) gene nomenclature between***
1336 ***Xue et al. and this study.*** *The table lists the gene names assigned to GPX family*
1337 *members in Xue et al. and the corresponding nomenclature used in this study (Ticó et*
1338 *al.). The third column provides relevant notes on differences in naming, such as updated*
1339 *classification, subfamily assignment, or gene duplication interpretations.*

1340 ***Table T2. Detection of SELENOW1, SELENOW2 and MIEN1 transcripts across vertebrate***
1341 ***species.*** *Boolean values denote presence (True) or absence (False) of RNA-derived*
1342 *sequences assigned to each family based on similarity scoring against the curated*
1343 *reference dataset extracted from the SELENOW gene tree. Corresponding taxonomic*
1344 *lineage is provided for each species.*

1345 ***Table T3. List of NCBI species and their corresponding genome assembly accession IDs***
1346 ***used in this study.***

1347 ***Table T4. Substitution models selected by IQ-TREE for phylogenetic tree reconstruction***
1348 ***in each selenoprotein family.***

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

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