Article – Discoveries section of MBE

Distinct patterns of selection in selenium-dependent genes between land and aquatic vertebrates

Gaurab K. Sarangi[1], Frédéric Romagné[1] and Sergi Castellano[1,2,3]

[1] *Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig 04103, Germany*

[2] *Genetics and Genomic Medicine Programme, Great Ormond Street Institute of Child Health, University College London (UCL), London WC1N 1EH, UK*

[3] *UCL Genomics, London WC1N 1EH, UK*

Corresponding author: Sergi Castellano at s.castellano@ucl.ac.uk

Abstract

Selenium (Se), a sparse element on earth, is an essential micronutrient in the vertebrate diet and its intake depends on its content in soils and waters worldwide. Selenium is required due to its function in selenoproteins, which contain selenocysteine (Sec), the 21st amino acid in the genetic code, as one of their constituent residues. Selenocysteine is analogous to the amino acid cysteine (Cys), which uses the abounding element sulfur instead. Despite the irregular distribution of Se worldwide, its distinct biochemical properties have made the substitution of Sec for Cys rare in vertebrate proteins. Still, vertebrates inhabited environments with different amounts of Se and may have distinctly adapted to it. To address this question, we compared the evolutionary forces acting on the coding sequences of selenoprotein genes and genes that regulate Se between vertebrate clades and between the selenium-dependent genes and their paralogs with Cys. We find that the strength of natural selection in genes that use or regulate Se is distinct between land vertebrates and teleost fishes and more variable than in the Cys paralogs, particularly in genes involved in the preferential supply of Se to some organs and the tissue-specific expression of selenoproteins. This is compatible with vertebrates adapting to Se scarcity in land and its abundance in waters. In agreement, teleost fishes duplicated and subfunctionalized or neofunctionalized selenoprotein genes and maintained their capacity for Se transport in the body, which declined (under neutrality) for millions of years in terrestrial vertebrates. Dietary Se has thus distinctly shaped vertebrate evolution.

Introduction

Vertebrates have adapted over the past 500 million years to most of the earth's vast range of environments. These environments differ widely in their geology and vary in the abundance of the chemical elements required by the various vertebrate species in them. Elements such as zinc, iron, manganese, copper, iodine and selenium (Se) are essential components of the vertebrate diet but are needed only in trace amounts, with their deficiency or excess having adverse health consequences (Mertz 1981; WHO 1987). Indeed, disorders from inadequate levels of these micronutrients in humans and other animals have long been known (Mills 1974; Rayman 2012).

Of these essential micronutrients, Se is unusual in that it has a narrow margin between nutritionally optimal and potentially toxic (Wilber 1980), making its uneven environmental distribution a challenge to the needs of the vertebrate diet. Diet is the most important source of Se and its intake depends on the Se content of the soil on which food is gathered, hunted or grown (Johnson, et al. 2010). Soil Se levels, in turn, depend largely on the underlying bedrock from which they are formed, which has created a patchwork of deficient, adequate and sometimes toxic areas across the world varying hundreds-fold in their Se levels. Importantly, human populations inhabiting some of the most Se-deficient environments in the world have adapted to its scarcity (White, et al. 2015). Levels of Se in waters worldwide also vary hundreds-fold (Selinus, et al. 2005) but oceans, seas and other aquatic environments are an environmental sink for land Se (Selinus, et al. 2005), with its inorganic forms being rapidly and efficiently bioaccumulated in phytoplankton (100-1,000,000 fold-enrichment from the water concentration (Stewart, et al. 2010) and converted into organic forms of Se that can enter the animal diet (Ogle, et al. 1988). Seafood is today a primary source of Se in the human diet (Sunde 2014). Thus, vertebrate species from different terrestrial and aquatic environments evolved under different levels of Se in their diets.

Selenium is required by vertebrates mainly due to its function in selenoproteins, which contain the amino acid selenocysteine (Sec) as one of their constituent residues. Sec, the 21st amino acid in the genetic code, is encoded by a UGA (STOP) codon and is analogous to the amino acid cysteine (Cys), which is encoded by the UGC and UGU codons. Sec has a Se-containing selenol group in place of the sulfur-containing thiol group

in Cys (Hatfield 1985), conferring different biochemical properties to selenoproteins (Gromer, et al. 2003; Steinmann, et al. 2010; Snider, et al. 2013). In consequence, substitutions between Sec and Cys in orthologous proteins are rare in vertebrates and under evolutionary constraint (Castellano, et al. 2009), although paralogous proteins with Cys instead of Sec exist in most vertebrate species. At least 28 selenoprotein genes existed in the vertebrate ancestor (Castellano, et al. 2005; Lobanov, et al. 2007; Mariotti, et al. 2012) but their number today varies widely, with teleost fishes having from 10 to 14 more selenoprotein genes than other vertebrate species. Moreover, the number of Sec residues in Selenoprotein P (SELENOP), which transports Se atoms from the liver to all other organs and tissues, also varies more than two-fold among vertebrates, with teleost fishes always in the upper part of this range (Lobanov, et al. 2008). Thus, teleosts fishes have larger selenoproteomes and potentially more Se transported *via* plasma than most other vertebrates.

It has been then suggested that teleost fishes have developed greater dependence on environmental Se, whereas other vertebrates have reduced their reliance on it (Lobanov, et al. 2007; Lobanov, et al. 2008). This presents us with the question of whether natural selection has distinctly shaped the use and regulation of this micronutrient in the different vertebrate clades. We addressed this question by comparing the evolutionary forces acting on the coding sequences of selenoprotein genes and genes involved in the regulation of Se between clades and to those acting on their paralogs with Cys along the vertebrate phylogeny. We find that the strength of natural selection has significantly changed across vertebrate lineages and clades for genes that use (selenoprotein genes) or regulate Se, while it is more uniform in the Cys-containing genes, which depend on sulfur and not Se. The strength of natural selection is uniquely variable in regulatory genes across the vertebrate phylogeny, particularly in genes that contribute to the hierarchy of Se supply to organs and tissue-specific expression of selenoproteins. The strength of selection is, in addition, uniquely variable in the selenoprotein genes of teleost fishes, further suggesting a complex nutritional history in a clade where selenoprotein gene duplications abound. Most of these gene duplications may have later split the ancestral function between them (subfunctionalization) but one or two of them may have acquired novel functions (neofunctionalization). Finally, we infer that the capacity to transport Se atoms *via* plasma

in SELENOP is under strong evolutionary constraint in the selenoprotein-rich teleost fishes, whereas it has evolved neutrally in mammals and most other non-fish lineages. This provides an evolutionary explanation to the conservation in teleost fishes and decline in most non-fish vertebrates of their capacity to transport Se through the body. We conclude that the essentiality of Se in vertebrate proteins (Castellano, et al. 2009) has made its environmental variation across the earth a distinct selective pressure for genes and proteins that use or regulate this micronutrient, with both its deficiency in humans (White, et al. 2015) and other terrestrial vertebrates and abundance in teleost fishes having left signatures of selection.

Results

Annotation, orthology, paralogy and alignment of vertebrate genes

Selenoprotein genes are missannotated in most databases due to their use of an in-frame UGA termination codon to encode Sec, the Se-containing amino acid. For this work, we manually curated our own computational annotations of 19 selenoprotein genes in 53 vertebrate genomes from our database SelenoDB 2.0 (Romagne, et al. 2014). These selenoprotein genes belong to 15 gene families (table 1 and supplementary fig. 1, Supplementary Material online). From the same database, we manually curated eight Cys-containing genes (selenoproteins' paralogs) belonging to four gene families. Seventeen genes that regulate Se were obtained from Ensembl 69 (Flicek, et al. 2013). Gene orthology assignments were taken from SelenoDB 2.0 and Ensembl.

We aligned the protein sequences of these orthologous vertebrate genes and used a probabilistic approach to remove amino acid positions whose alignment is uncertain, usually due to sequence gaps or sequence divergence in the same or nearby positions (Materials and Methods). In total, we aligned 24,191 orthologous amino acid positions of which 3,344 (13.8%) and 353 (1.5%) were excluded due to alignment uncertainty from gaps and sequence differences, respectively. Of the 24,494 amino acid positions analyzed, 5,046 are from selenoprotein genes, 14,133 from Se regulatory genes and 1,315 from Cys-containing genes.

We also manually curated 12 Sec-containing paralogs of selenoprotein genes in the teleost fishes belonging to nine protein families from our computational annotations in

SelenoDB 2.0. Gene paralogy assignments were taken from SelenoDB 2.0 and Ensembl. We aligned 1,612 paralogous amino acid positions of which 322 (20.0%) gap and 59 (3.7%) divergent amino acid positions were excluded due to alignment uncertainty. Finally, for the analysis of the transport domain of SELENOP, we aligned 481 amino acid positions, of which 87 (18.1%) gap and 53 (11.0%) divergent position, were removed from the vertebrate alignment, respectively, due to their uncertainty.

In this regard, it is worth mentioning that our probabilistic approach allowed us to keep those regions of the alignment with gaps and amino acid differences that are nevertheless confidently aligned in the multiple alignments of orthologous and paralogous genes (Materials and Methods). The codons encoding the 20,494 orthologous, the 1,231 paralogous and the 341 SELENOP amino acid positions were used in the evolutionary analyses presented in this work.

Genes with a variable dN/dS ratio along the vertebrate phylogeny

We first asked, separately for the groups of selenoprotein genes, their Cys-containing paralogs and the genes that regulate the metabolism and homeostasis of Se (table 1), whether the strength of natural selection acting on them has been different or the same across the vertebrate phylogeny (supplementary fig. 2, Supplementary Material online). To do this, we rephrased this question in terms of an evolutionary ratio: dN/dS. This ratio quantifies the strength of selection on protein sequences by comparing the rate of substitutions at synonymous sites (dS), which are presumed neutral, to the rate of substitutions at non-synonymous sites (dN), which are possibly under selection. For each orthologous gene, we contrasted a model that supports independent dN/dS ratios across vertebrate lineages (present-day and ancestral) with a model that supports the same dN/dS ratio for all lineages, and thus assumes no variation in this ratio in the vertebrate phylogeny (null model). In particular, we compared PAML's free-ratios model to the one-ratio model (Yang 2007) using a likelihood ratio test (Materials and Methods). We calculated P-values from these tests and compared their distribution in selenoprotein genes, their Cys-containing paralogs and the regulatory genes (supplementary fig. 3, Supplementary Material online) to each other using a one-sided Mann–Whitney U test, thus testing their deviation as a group from the null model. The use of P-values instead of the likelihood

ratios themselves is necessary to take into account the varying complexity of the alternative model among genes, which are sometimes unevenly annotated in vertebrate genomes (supplementary fig. 1, Supplementary Material online).

We found that a model with independent dN/dS ratios across vertebrate lineages was a significantly better fit for both selenoprotein genes ($P = 0.001$; Mann–Whitney U; one-sided test) and genes that regulate Se ($P = 0.0001$) than for the Cys-containing genes (fig. 1A). Thus, selenoprotein genes and genes that regulate Se, compared to genes that use sulfur and are independent of Se and its abundance across the world, experienced unusually large changes in the strength of selection among vertebrate species. Interestingly, the extent of variation of the dN/dS ratio across the vertebrate phylogeny was larger in the regulatory genes than in selenoprotein genes ($P = 0.0009$), with roughly twice as many regulatory genes contributing to the Mann–Whitney U differences with the Cys-containing genes than selenoprotein genes (Materials and Methods). This suggests that the regulation of selenium metabolism and homeostasis is particularly polygenic. Indeed, 11 genes (CELF1, EIF4A3, LRP8, LRP2, SARS2, SPS1, XPO1, SBP2L, PSTK, SBP2 and ELAVL1) contributed around 75% of the difference with the Cys-containing genes, whereas only six selenoprotein genes (TXNRD1, TXNRD3, SELENOI, SELENOO, TXNRD 2 and DIO2) are needed to contribute that much.

The genomes of the vertebrate species analyzed are of uneven sequence quality and, thus, the completeness of our gene annotation varies among species (supplementary fig. S1, Supplementary Material online). While the significance (in the form of a P-value) of our likelihood ratio tests already takes the unevenness of our gene annotation into account (Materials and Methods), we additionally resampled the number of species that can contribute to the variation of the dN/dS ratio across the vertebrate phylogeny for each gene in our analysis (Materials and Methods). We used 200 samples to compute the median and its 95% confidence interval of the significance of our comparisons above (fig. 1A) and concluded that differences in sequence and gene annotation quality among vertebrate genomes do not impact our results. Furthermore, the smaller number of Cys-containing paralogs should not diminish the significance of the observed differences in our test (Mann and Whitney 1947). Specially, because the P-values from the Cys-containing paralogs overlap, for the most part, with the small range of P-values from neutral simulations

(supplementary fig. 4, Supplementary Material Online; Materials and Methods). This suggests that the issue of Cys-genes compensating for the lack of Se while becoming more essential under its deficiency does not contribute much to the already small variation in their strength of selection, which is not far from its neutral expectation. It also suggests that known Cys-containing paralogs fairly represent the range of variation in the strength of selection typical of these genes, which do not depend on Se and seem to evolve quite uniformly across lineages.

Genes with variable dN/dS ratios among protein sites

We next asked, for the groups of selenoprotein genes, their Cys-containing paralogs and the genes that regulate Se (table 1), whether natural selection acting on them has been different or the same across their protein sequences. Again, we rephrased this question in terms of the dN/dS ratio. For each orthologous gene, we contrasted a model that estimates three independent dN/dS ratios by dividing protein sites into three categories and a model that estimates just one dN/dS for all sites, and thus assumes no variation in this ratio across the sequence of a protein (null model). To do this, we compared PAML's three-ratio site model M3 and the one-ratio site model M0 (null model) using a likelihood ratio test (Materials and Methods). We compared the distribution of log likelihood ratios from this test in selenoprotein genes, their Cys-containing paralogs and the regulatory genes to each other using again a one-sided Mann–Whitney U test. We found that a model that allows variation of the dN/dS ratio across protein sites was a significantly better fit for genes involved in the regulation of Se ($P = 0.026$; Mann–Whitney U; one-sided test) than for genes with Cys. The M3 model also fitted better the selenoprotein genes than their paralogs with Cys, although the difference was not significant ($P = 0.119$). The variation of the dN/dS ratio across the protein sequences of regulatory genes was larger than in selenoprotein genes, but not significantly so ($P = 0.093$). These results were largely in the direction of our analysis on the variability of the strength of natural selection across vertebrates, with regulatory and selenoprotein genes having the stronger variation along their protein sequences expected from their elevated dN/dS variation across the branches of the vertebrate phylogeny.

Vertebrate clades with highly variable dN/dS ratios

As our next step, we sought to identify whether the unusually variable dN/dS ratio in the groups of selenoprotein and regulatory genes stems from the entire vertebrate phylogeny or from individual clades. We divided the vertebrate phylogeny into five clades: primates, rodents, laurasiatheria (which includes bovids, pigs, horses, carnivores and others), sauria (birds and reptiles) and teleost fishes and, for each orthologous gene in each clade, we contrasted using a likelihood ratio test (Materials and Methods) a model that supports independent dN/dS ratios across lineages (PAML's free ratio model) with a model that supports the same dN/dS ratio for all lineages (null model). We then compared the fit of these models between clades and between each clade and the entire vertebrate phylogeny for the groups of selenoprotein and regulatory genes using a one-sided Mann–Whitney U test. This test is independent of the Cys-containing genes. We found that a model with independent dN/dS ratios across lineages was generally a significantly better fit in the entire vertebrate phylogeny than in each of the clades for selenoproteins ($P = 0.002$ for teleost fishes and $P < 5x10^{-5}$ for the other clades; Mann–Whitney U; one-sided test) and regulatory genes ($P < 1x10^{-5}$ for all clades). Thus, the dN/dS ratio in selenoprotein and regulatory genes significantly varied across the overall vertebrate phylogeny more than in any individual clade. This is not unexpected given the breadth of the evolutionary and nutritional histories of the entire phylogeny.

However, the dN/dS ratios of selenoprotein genes appeared to be unusually variable among teleost fishes (fig. 2A), making a model with independent dN/dS ratios a significantly better fit in these fishes than in any other vertebrate clade ($P = 0.0059$ vs. primates; $P = 0.0318$ vs. rodents; $P = 0.0172$ vs. laurasiatheria; $P = 0.0178$ vs. sauria). The variation of the dN/dS ratios in the regulatory genes of teleost fishes is also larger than in the other clades, although often not significantly so (fig. 2B). In any case, the variation in the dN/dS ratio of selenoprotein genes in teleost fishes varied as much as in their regulatory genes ($P = 0.464$ for their difference). In agreement, the same number of genes contributed around 75% of the difference with the Cys-containing genes in both regulatory (LRP2, SBP2, SARS2 and CELF1) and selenoprotein (TXNRD3, DIO2, GPX2 and SELENOI) genes. This contrasted with the significantly higher variation in the dN/dS ratio of genes that regulate Se in the entire vertebrate phylogeny ($P = 0.0009$) and its stronger polygenic

signature than in selenoprotein genes.

Variation of the dN/dS ratio in non-fish vertebrates

We further investigated whether the elevated variation in the dN/dS ratio of selenoprotein genes in teleost fishes impacts our previous results using orthologous genes across the vertebrate phylogeny. We thus asked whether the strength of natural selection acting on selenoprotein genes and genes that regulate Se in the non-fish lineages of the vertebrate phylogeny, compared to the Cys-containing paralogs, has been different (as found for the entire phylogeny) or the same. We find that a model with independent and variable dN/dS ratios across non-fish vertebrates was once more a significantly better fit for both selenoprotein genes ($P = 0.004$; Mann–Whitney U; one-sided test) and genes that regulate Se ($P = 0.0005$) than for Cys-containing genes (fig. 1B). Non-fish regulatory genes, in turn, had again more variable dN/dS ratios than selenoprotein genes ($P = 0.012$) (fig. 1B). Indeed, nine genes (LRP2, LRP8, SARS2, SPS1, XPO1, EIF4A3, CELF1 and SBP2) contributed around 75% of the difference with the Cys-containing genes, whereas only four selenoprotein genes (TXNRD1, TXNRD3, SELENOO, SELENOW1) are needed to contribute that much.

As before, the differences in sequence quality and gene annotation among vertebrate genomes did not influence these results (fig. 1B). We concluded that orthologous genes that regulate and use (selenoprotein genes) Se and, thus depend on the uneven distribution of this essential micronutrient throughout the world, had unusually large changes in the strength of selection acting on them in both non-fish and teleost fish vertebrates. The latter, however, had more variation in the strength of selection acting on selenoprotein genes. This may be related to the overall abundance and bioaccumulation of Se in waters of the world and the increase through duplications of genes using Se in the teleost fishes (Lobanov, et al. 2008).

Selenoprotein gene duplications in the teleost lineage

We thus studied seven selenoprotein gene copies in the teleost fishes that have kept Sec in their protein sequence since duplication. These are DIO3, GPX1, GPX3, GPX4, SELENOT, MSRB1 and SELENOU1. We traced the origin of these selenoprotein gene

paralogs to a lineage ancestral to the teleost fishes in our analysis and, hence, they are most likely the result of the whole-genome duplication event that happened in the teleost fish ancestor (Jaillon, et al. 2004). Interestingly, 25% (seven out of 28) selenoprotein duplicates are present in the genome of teleost fishes today, which is about twice the average retention rate observed in protein families after whole-genome duplication (Brunet, et al. 2006). This raises the question of the evolutionary mechanism responsible for the retention of this large number of duplicated genes. We therefore investigated two alternative models of evolution that can explain the retention of these duplicated genes, namely neofunctionalization and subfunctionalization (Innan and Kondrashov 2010). Neofunctionalization, where one duplicate evolves a new function in which natural selection can act upon, is expected to result in an asymmetry in the rate of evolution of the two gene copies. In contrast, subfunctionalization, where the duplicates subdivide the ancestral functions through the accumulation of loss of function mutations, is expected to result in higher but still similar rates of evolution between the gene copies.

For each paralogous gene, we compared these two models of evolution estimating independent dN/dS ratio in the lineages before the gene duplication and immediately after it (table 2). To do this, we used a pairwise maximum likelihood comparison in PAML (Materials and Methods). We found that the dN/dS ratio tends to increase somewhat after duplication but remains comparable between the gene copies for most duplicated genes, which is compatible with subfunctionalization. We analyzed in the same way the five lineage-specific duplications of selenoprotein genes in the teleost phylogeny (Mariotti, et al. 2012). These Se-dependent copies are SELENOT1b, SELENOO2, SELENOW2b, SELENOU1b and SELENOJ2. Again, we found for the most part small differences in the dN/dS ratio between the lineage-specific gene copies and between these and their ancestor (table 2), hence, suggesting subfunctionalization.

One mechanism (other than amino acid changes) through which subfunctionalization could also have occurred is changes in the time and tissue expression of duplicated genes. We thus investigated whether subfunctionalization could have led to (or be a consequence of) distinct expression patterns between gene copies from the whole-genome and lineage-specific duplications. Indeed, there is published evidence of differential time and tissue expression for DIO3 (Marelli, et al. 2016), GPX4 (Mendieta-

Serrano, et al. 2015), MSRB1 and SELENOT1 in the ZFIN database (Howe, et al. 2013) at http://www.zfin.org/. We did not find expression information for both copies of the other whole-genome duplications or lineage-specific duplications with symmetric rates of evolution.

Interestingly, one ancestral whole-genome duplication, GPX3, and one lineage-specific duplication, SELENOJ, have rates of evolution between gene copies and between these and their ancestor that differ more than two-fold (table 2). This is suggestive of neofunctionalization but the possible new roles of selenium in the activity of these proteins remain unexplored. Interestingly, GPx3 is a plasma protein with peroxidase activity and perhaps a minor role in the maternal-fetal transport of selenium (Burk, et al. 2013), and SELENOJ is a protein with a potential function in ADP-ribosylation that may have later acquired a structural role in the eye lens (Castellano, et al. 2005).

Transport of Se by selenoprotein P

We next tested the pattern of accumulation and retention of Sec residues in the transport domain of SELENOP over its vertebrate history. Selenoprotein P transports Se atoms in the form of Sec but their number varies widely across vertebrates (from 7 to 18 selenium atoms) as Sec is often substituted by the sulfur-containing Cys amino acid (Lobanov, et al. 2007; Lobanov, et al. 2008). We thus simulated half a billion years of vertebrate evolution in the Sec residues of SELENOP's transport domain and measured their neutral exchange with Cys. To derive the neutral expectation of the average number of Sec residues transported in the terrestrial and teleost fish lineages we performed 10,000 simulations using synonymous rates from SELENOP as a proxy for neutrality (Materials and Methods). We compared the observed averages to these distributions and found them to be consistent with neutral evolution in terrestrial lineages (observed average in non-fishes = 10.84; neutral average = 11.25; $P > 0.5$) and strong purifying selection in teleost fishes (observed average in teleost fishes = 15.57 Se atoms; neutral average = 11.47; $P < 0.0001$) (fig. 3). Furthermore, the comparison of the distribution of the actual number of Sec residues across species of teleost fishes ($P = 0.00002$) and non-fish vertebrates ($P = 0.5162$) to their neutral expectation supports the same conclusions (supplementary fig. 9, Supplementary Material online). Finally, to test our neutral model we repeated the

simulations and analysis shown in figure 3 using the synonymous rates from multiple genes across the genome (Castellano, et al. 2009). The results agree (supplementary fig. 10, Supplementary Material online).

Conservation of Se transport in teleost fishes may be related to the size of their selenoproteomes (Lobanov, et al. 2008) and their Se needs, which appear to be larger than in mammals and birds. Indeed, the number of Se atoms in SELENOP seems to positively correlates with the increasing requirements of Se (in mg per Kg of organic dry matter) from mammals to birds to teleost fishes (Penglase, et al. 2015).

Discussion

The varying amounts of Se across the earth environments has made its uneven dietary intake a potential selective pressure throughout vertebrate history. In a recent study, this question was addressed with respect to recent human evolution. Human populations living today in the extreme Se deficient areas of China have allele frequency shifts in the genes that use or regulate Se that are compatible with the action of positive selection (White, et al. 2015). Heart (Keshan) and bone (Kashin–Beck) diseases, which are treated with Se supplementation, are endemic to these areas of China. Thus, recent changes in the use or regulation of Se when deficient may have been adaptive. Similarly, variants in genes that regulate iodine metabolism may have allowed some Pygmy populations, which have unusually low levels of goiter today (Dormitzer, et al. 1989), to adapt to the iodine deficient environments they inhabit (Lopez Herraez, et al. 2009).

In this work, we investigated the role of Se in a larger timeframe, one spanning 500 Myr of vertebrate evolution and diversification into fishes, amphibians, reptiles, birds and mammals. Due to the misannotation of selenoprotein genes in major databases (Sec is encoded by an equivocal STOP codon), we relied on our own computational and manually curated gene annotations in vertebrate genomes (Romagne, et al. 2014). Using these annotations, we conducted evolutionary analyses on selenoprotein genes but also genes involved in the regulation of Se in 53 vertebrate species and compared them between clades and to the Se-independent paralogs of selenoprotein genes. If differences in dietary Se pose different selective pressures among vertebrate lineages, we expect the Se-related genes to have evolved under varying strengths of natural selection in vertebrates. At the same time,

we expect genes that do not depend on Se (Cys-containing genes) to be uninfluenced by its abundance throughout the world and, hence, to evolve under a more uniform strength of selection. We find this to be the case, with genes using or regulating Se having significantly higher variation in their dN/dS ratios across the different vertebrate lineages than the Cys-containing genes (fig. 1). From this we conclude that the uneven distribution of Se across earth environments and its distinct dietary availability among vertebrates may have posed the selective pressure accounting for this variation. This suggests that Se availability has shaped the evolution of vertebrates.

One caveat to this interpretation is the possibility that Cys-containing genes functionally compensate for Se variation and, thus, also depend on the variation in dietary Se among vertebrates. Another caveat is that differences in the essentiality of genes with Sec and Cys makes their comparison unsuitable. We argue that both of these factors, if substantial, would increase the variation in the strength of selection in the Cys-containing genes, making more difficult to obtain the observed differences. The reason is that Cys-containing genes should compensate more (and hence become more essential) in vertebrate species with less rather than more selenium. Thus, the variation in their dN/dS ratios across vertebrates should increase somewhat proportionally to the variation in dietary selenium among vertebrates, making the reported tests conservative. In any case, the variation in the strength of selection across vertebrate lineages is highest in the genes regulating the metabolism and homeostasis of Se, which is in agreement with the greater dN/dS ratio variation across protein sites in these genes. Regulatory variation, if adaptive, is likely to have a wider physiological impact than adaptations in selenoprotein genes. Another caveat is the smaller number of Cys-containing genes compared to selenium-dependent ones (supplementary fig. S3, Supplementary Material online). However, our test is robust to differences in the number of genes compared between groups (Materials and Methods). Furthermore, neutral simulations of Cys-containing genes support this notion (Results).

Tests that compare the groups of genes that use or regulate selenium among themselves across vertebrates are not subject to the caveats above. Such comparisons show that, within the vertebrate phylogeny, the teleost fish clade has a unique pattern of variation in the strength of natural selection. The dN/dS ratios of selenoprotein genes in this clade are unusually variable (as variable as in its regulatory genes), which is suggestive of

additional differences in the dietary history of the species in this clade. The source of this variation is not well understood but both the overall higher Se content in aquatic environments (probably due to increased bioaccumulation in the trophic web) compared to terrestrial ones, and the variety of environments fishes in the teleost clade inhabit (marine, brackish and fresh water environments) may contribute to it. This is best discussed in the context of the common and separate history of selenoprotein gene duplications in the teleost fishes. The whole-genome duplication event in the teleost fish ancestor (Jaillon, et al. 2004), which added seven selenoprotein genes to the teleost genome, agrees with the presumed higher availability of Se in waters than in lands around the world but provides no insight into Se differences among teleost fishes. The seven lineage-specific selenoprotein gene duplications in teleost species, however, are indicative of increased Se availability (compared to non-fishes, which have few lineage-specific duplications) but also of dietary differences among the fish lineages themselves. If so, other lineage-specific duplications may exist in unsequenced teleost genomes.

The model of gene duplication of selenoprotein genes in the teleosts is also informative. We find for the most part evidence of subfunctionalization from the rate of evolution of the duplicated gene copies, with the ancestral function of some duplicates being likely split by changes in their timing and pattern of gene expression (Innan and Kondrashov 2010). Neofunctionalization may also have occurred in two selenoprotein gene duplications (one ancestral and one lineage-specific) but whether Se itself acquired novel functional is unknown (table 2). In any case, the whole genome and lineage-specific duplications contributed to the larger selenoproteome found in teleost fishes today and, in turn, the expanding selenoproteome may have increased their need for Se (Lobanov, et al. 2008). This agrees with the inferred conservation (under purifying selection) of the transport capacity of Se from the liver to the other body organs in teleost fishes (fig. 3), which contrast with the loss (under neutrality) of the capacity to transport Se on SELENOP in most other terrestrial vertebrates (fig. 3). We note, however, that ultimately the total amount of Se transported via plasma depends on the expression of SELENOP. In particular, it depends on the expression levels of the full protein (with all Se atoms) and the expression levels of shorter isoforms (with fewer Se atoms) (Shetty, et al. 2014). Interestingly, the expression of the full protein is promoted by the availability of Se (Shetty, et al. 2014),

suggesting that Se-rich teleost fishes will tend to transport more Se atoms in SELENOP than most other land vertebrates.

We conclude that the sequence patterns described in this work support two modes of evolution in vertebrates, one for terrestrial vertebrates and another for aquatic ones. This has been hypothesized before on the basis of the different number of selenoprotein genes in vertebrate genomes and Se atoms in SELENOP (Lobanov, et al. 2007; Lobanov, et al. 2008) but no evolutionary tests on this hypothesis had been put forward. In doing so, a more nuanced and comprehensive picture of the role of Se in vertebrate evolution emerges. In particular, the evolutionary importance of the polygenic evolution of genes involved in the metabolism and homeostasis of Se in vertebrates under both its deficiency in the land, as previously reported for human populations in China (White, et al. 2015), and presumed abundance in the water. It becomes then significant that the specific SELENOP receptor for the brain, testis and bone (LRP8) (Pietschmann, et al. 2014), the only organs to preserve acceptable levels of Se under its deficiency, only contributes to adaptation when Se is scarce. That is in terrestrial vertebrates and in human populations in areas of China that do not provide enough dietary Se (White, et al. 2015). It also becomes significant that one of the regulatory process with the strongest adaptive signatures, across both vertebrate species and human populations and across all Se levels, is the differential expression of selenoproteins. In particular, the regulation of SBP2 expression by CUG-BP1, which determines selenoprotein expression in a tissue and Se level-dependent manner (Squires, et al. 2007). CUG-BP1 binds the proximal region of the 3'UTR of SBP2 and, in doing so, controls the stability and translatability of SBP2. The differential binding affinity of SBP2 to the SECIS (selenocysteine insertion sequence) RNAs of the different selenoproteins is responsible for some of the changes in their expression levels with varying Se status (Schomburg and Schweizer 2009). This suggests that the hierarchy of Se supply to the various organs, determined by the LRP8 receptor, and the hierarchy of selenoprotein expression across tissues, determined by SBP2 and its regulating proteins, has been targeted by natural selection in the distant and recent past to adapt to levels of Se in the vertebrate (and human) diet. Whether deficiency and abundance of other essential micronutrients leads to similar evolutionary patterns remains unknown.

Materials and Methods

Orthologous genes and vertebrate species

For our orthology analysis across the vertebrate phylogeny, we considered 44 vertebrate genes of which 19 are selenoprotein genes, eight are Cys-containing paralogs of the selenoprotein genes and 17 are genes involved in the regulation of Se (table 1 and supplementary fig. S1, Supplementary Material online). Note however that the selenoprotein genes SELENOP and SPS2 are grouped with the regulatory genes in our analyses due to their regulatory function. Note also that the selenoprotein gene SELENOW2 is only present in teleost fishes and amphibians and that the Cys-containing gene Rdx12 (also known as MIEN1) is only present in non-fish vertebrates (Mariotti, et al. 2012). We did not analyze the GPX6 gene because of its mosaic vertebrate distribution with Sec and Cys. We also excluded the seven genes (DIO3, GPX1, GPX3, GPX4, SELENOT, MSRB1 and SELENOU1) which have more than one copy in fishes due to an ancestral teleost-specific whole-genome duplication but have only one copy in the rest of the vertebrates. We investigated these seven genes (and other lineage-specific paralogs in teleost fishes) in a separate analysis concerning duplicated genes (table 2). We used the recently updated nomenclature for selenoprotein genes (Gladyshev, et al. 2016) throughout this work.

We retrieved the protein-coding sequences for the 19 selenoprotein and eight Cys-containing genes from 53 vertebrate genomes annotated in SelenoDB 2.0 (Romagne, et al. 2014). This database provides computational gene annotations for these genes, which we manually curated for this work. It is important to use these gene annotations for our analyses as the dual and seemingly ambiguous nature of UGA codons (coding for STOP or Sec) has led to many annotation errors in selenoprotein genes in major databases (*e.g.* truncated gene structures stopping at or skipping the Sec residue). For the 17 regulatory genes that don't have a Sec residue we used the gene annotations available in Ensembl 69 (Flicek, et al. 2013). Orthology assignments between vertebrate species were obtained from SelenoDB 2.0 and Ensembl 69.

The vertebrate genomes annotated in SelenoDB 2.0 encompass one coelacanth, seven teleost fishes, one amphibian, two reptiles, three birds and 39 mammals (monotremes, marsupials, bovids, pigs, carnivores, rodents, primates and others). These

provide a rich sample of Se nutritional histories in vertebrates across many of the earth's environments. The same species were used for the regulatory genes annotated in Ensembl 69. The sequence quality of the genomes of these species is however uneven (Flicek, et al. 2013) and, as a result, some genes are either partially annotated or not annotated in the genomes of some species (supplementary fig. 1, Supplementary Material online). Our analyses resampling these vertebrate species take into account the unevenness of the gene annotation among them.

Paralogous genes in teleost fishes

We also considered 12 Sec-containing paralogs of selenoprotein genes in the teleost fishes (table 2). Seven of these paralogous genes resulted from a whole-genome duplication event in the ancestor of teleost fishes. The date of this whole-genome duplication event remains uncertain but may have occurred around 300 Mya (Christoffels, et al. 2004; Hoegg, et al. 2004; Vandepoele, et al. 2004; Hurley, et al. 2007). The other five paralogous genes resulted from lineage-specific duplications that happened later. We retrieved and manually curated the protein-coding sequences for the 12 duplicated genes from the seven teleost fish genomes annotated in SelenoDB 2.0. Paralogy assignments were also obtained from SelenoDB 2.0.

Alignment of orthologous and paralogous genes

We used MAFFT (Katoh and Standley 2013) to align the protein sequences of orthologous genes with Sec or Cys (table 1) across the vertebrate phylogeny. We used the same program to align the protein sequences of the duplicated genes (table 2) in teleost fishes and the transport domain of SELENOP. To avoid wrongly aligned amino acid residues that could confound our downstream evolutionary analyses, we converted each protein multiple alignment into a Hidden Markov Model using HMMER (Eddy 1998), and use a forward-backward algorithm (Durbin, et al. 1998) to compute a posterior probability representing the degree of confidence in each individual aligned residue or gap for each protein sequence.

Using this approach, we removed any amino acid position from the multiple alignment with an average posterior probability (from all the protein sequences for that

position) of less than 0.9. These tend to be: 1) positions of the multiple alignment where gaps create alignment uncertainty in the same or nearby positions. Positions where sequences are missing in many species (typically at the edges of protein sequences due to missing gene annotations or insertions and deletions) can contribute to the alignment uncertainty and thus also be removed; and 2) positions of the alignment where sequence divergence leads to alignment uncertainty in the same or nearby positions and thus misalignments are possible (typically at the edges of protein sequences but also in their middle). The UGA codon encoding the Sec amino acid was treated as an ambiguity character. The codons encoding the remaining amino acid positions (average posterior probability of 0.9 or higher) in the multiple alignment were used in the subsequent PAML (Yang 2007) analyses. Note that these include the codons of amino acid positions whose gaps and amino acid differences are convincingly aligned.

PAML analysis

We used the program CodeML provided in the package PAML (Yang 2007) for all of our codon analyses in the vertebrate phylogeny (supplementary fig. 2, Supplementary Material online). CodeML, uses the parameter dN/dS as a measure of the strength and mode of natural selection acting on proteins, where, dN is the rate of non-synonymous substitution per non-synonymous site and dS is the rate of synonymous substitution per synonymous site. For our analysis, we employed both branch and site models which come inbuilt with CodeML. Confidently aligned gaps are allowed in the multiple alignments analyzed with PAML (cleandata = 0).

PAML branch models

The various branch models in PAML allow dN/dS ratios to vary among the branches in the phylogeny. One special case of branch model is the free-ratio model (model=1). In this model, for each of the branches in the phylogenetic tree an independent dN/dS ratio is estimated. This model is then compared to a null model (M0 model, model=0, NSsites=0), where only one dN/dS ratio is estimated for all branches. The difference in likelihood between the models is calculated as a likelihood ratio (in log space) and used, along with the difference in the number of parameters between the models (which depends on the

number of vertebrate genomes contributing annotations for a gene), to calculate its significance (in the form of a P-value). That is, the degrees of freedom of the likelihood ratio test is the number of parameters in the free model (which vary per vertebrate gene according to the annotated branches in the vertebrate phylogeny) minus one (the single parameter in the null model representing all branches in the vertebrate phylogeny). This has the advantage of taking into account the variation in the quality of the genome sequences and annotations used per vertebrate gene. The P-values from the likelihood ratio test are not significantly correlated with gene length.

We used this likelihood ratio test for our analysis on the variation of the strength of natural selection between the groups of selenoprotein genes, their Cys-containing paralogs and the genes that regulate Se (table 1) across the vertebrate phylogeny. We compared the distribution of P-values in each of these groups of genes to each other using a one-sided Mann–Whitney U test, thus testing their skewness (supplementary fig. 3, Supplementary Material online). Making this group comparison is important as the significance of the (log) likelihood ratio tests per gene is dependent on the specification of the null model (one single dN/dS rate for the whole vertebrate phylogeny). Significant deviations (see Results and fig. 1) between the P-value distributions indicate differences in the extent of variation of the dN/dS ratio in the groups of selenoproteins, their Cys-containing paralogs and the genes that regulate Se. Importantly, the unequal number of genes in these groups (the Cys- group has fewer genes) does not detract from the significance of their observed differences in our test (Mann and Whitney 1947).

Nevertheless, we explored the range of the P-values from the eight Cys-containing genes in the likelihood ratio test (supplementary fig. 3, Supplementary Material online). These P-values represent the significance of the variation in these genes of the strength of selection across vertebrates and come from the comparison of the alternative (one dN/dS ratio per branch) to the null model (one dN/dS ratio in the phylogeny) in our test. We repeatedly sampled and simulated the history under neutrality of Cys-containing-like genes to recalculate our likelihood ratio test: 1) starting from the codon alignments of each of the available Cys-containing paralogs, we used PhyloBayes (Lartillot and Philippe 2004) with default parameters to obtain posterior samples along the vertebrate phylogeny for each of these genes. We took six samples per gene, randomly changing in each of them the

substitution rates (branch lengths) in their phylogenies. Branch length were set to the sum of the dN and dS rates previously estimated per branch in our alternative model (free ratio model). Adding dN to dS creates some minor variability to the substitution rates among lineages. In addition, the first sample adds variation to the sum of the dN and dS rates previously estimated per branch in our alternative model (free ratio model) by sampling a gamma distribution in which 50% of the probability density is below or higher than one (alpha = 1.315). We used these samples to randomly scale branch lengths (dN + dS), proportionally decreasing or increasing them according to the value of the sample (increasing when it is higher than one and decreasing when it is lower than one). The second and third samples used a gamma distribution in which roughly 63% and 37% (alpha = 1) or the other way around (alpha = 1.780) are below and over one, respectively; 2) Based on these posterior samples, we used PhyloBayes to simulate sequence alignments to recalculate our likelihood test comparing the alternative to the null model with PAML (supplementary fig. 4, Supplementary Material online). The bulk of the distribution of P-values in this test from the simulated Cys-containing genes show no significant variation in the strength of selection among lineages. This is expected under neutrality as the dN/dS ratios should not change among lineages but for the stochasticity of substitution rates across phylogenies included in our simulations (*e.g.* a very short branch may increase or decrease its dN/dS ratio by chance).

The P-values in the regulatory and selenoprotein genes smaller than the lowest P-value in the Cys-containing genes were ranked from most to least significant. We measured the contribution of the P-value of each gene to the significance of the Mann–Whitney U test as its fraction of the sum of the ranked (in log space) P-values . Cumulative contributions are the sum of these fractions over genes and are given in the text as percentages (Results). We used the same likelihood ratio test to compare individual vertebrate clades (primates, rodents, laurasiatheria, sauria and teleost fishes) to each other and to the whole of the vertebrate phylogeny. As before, we compared the clade and vertebrate distribution of P-values in the groups of selenoproteins and the genes that regulate Se using a Mann–Whitney U test, and interpreted their skewness accordingly (Results). Nine vertebrate species that neither belong to the clades above nor form a clade

themselves (they are polyphyletic) were not included in the clade analysis (supplementary fig. 1, Supplementary Material online).

PAML pairwise maximum likelihood comparison

CodeML when used with the option "runmode = -2" makes pairwise maximum likelihood comparisons between sequences. We used this feature to investigate the rate of evolution of paralogous genes before and after the whole-genome duplication in the ancestor of the teleost fishes and the lineage-specific duplications within the teleost clade (supplementary fig. 2, Supplementary Material online). For each set of paralogous genes, we reconstructed four ancestral sequences using PAML: 1) one sequence belonging to the ancestor of the paralogous genes and the outgroup, before the (whole-genome or lineage-specific) gene duplication and before separation from the outgroup; 2) one sequence belonging to the ancestor of the paralogous genes, before duplication and after separation from the outgroup; and 3) two sequences belonging to the two gene copies immediately after duplication and before speciation. We then made three pairwise comparisons between these four ancestral sequences: 1) one comparing the ancestor of the paralogous genes and the outgroup against the ancestor of the paralogous genes (excluding the outgroup); 2) two comparing the ancestor of the paralogous genes (excluding the outgroup) against the two gene copies immediately after duplication.

PAML site models

The various site models in PAML allow dN/dS ratios to vary among sites (codons). For our analysis on the variability of the strength of natural selection across sites of selenoprotein genes, their Cys-containing paralogs and the genes that regulate Se (table 1), we used the site-model M3 (model=0, NSsites=3). In this model, sites are divided into three categories with independent dN/dS ratios estimated for each category. We compared this model with a null model in which only one dN/dS ratio is estimated for all sites (M0, model=0, NSsites=0). The difference in likelihood between the models is calculated as a likelihood ratio (in log space) and, since the difference in the number of parameter between the models is always the same (two), we used the likelihood ratios themselves to assess significance. We compared the distribution of likelihood ratios in each groups of genes to

each other using a one-sided Mann–Whitney U test, thus testing their skewness. Significant deviations between these likelihood ratio distributions indicate differences in the extent of variation of the dN/dS ratio across codons in each group of genes (Results).

Resampling gene annotations across vertebrate genomes

The varying quality of the vertebrate genomes used in this work makes our analyses challenging. This is because the completeness of ours and Ensembl's gene annotation is dependent on the completeness of the genome sequences of each vertebrate species. Thus, the vertebrate genomes analyzed varied from gene to gene, with the power of our gene-by-gene (log) likelihood tests and, hence, our ability to detect true differences among clades or gene categories ultimately depending on the sequence divergence of each of these orthologous genes. PAML tests are generally less powerful with closely related sequences and less reliable with very far ones. To assess this seeming bias we designed a resampling scheme that takes into account the underlying contribution of each vertebrate genome to the annotation of the 44 orthologous genes analyzed in this work (supplementary fig. S1, Supplementary Material online). To do this: 1) we make a list of the number of vertebrate genomes annotating each of these genes and then reassign the number of genomes contributing to the annotation of each gene by random shuffling the numbers in our list. The minimum number of genomes contributing annotations for a gene is eight and the maximum 42; 2) for each gene, we randomly select as many genomes contributing gene annotations as indicated in our shuffled list. When the number of genomes in our list for a gene is larger than the number of available genomes annotating the gene, we take all of the available gene annotations; and 3) we resample 200 times, perform our analyses in each sample and used them to build the 95% confidence interval around the median of the significance in the variation of the dN/dS ratio between selenoprotein genes, their Cys-containing paralogs and genes that regulate Se (Results and fig. 1).

Neutrality test on the transport of Se

The neutrality test is based on the comparison of the average number of Se atoms (Sec residues) present in the transport domain of SELENOP in teleost fishes and non-fish vertebrates to their neutral expectation (fig. 3), which was obtained using neutral

simulations of the evolution of ancestral Sec or Cys codons along the vertebrate phylogeny. To derive the distribution of the average number of Sec residues in the transport domain of SELENOP under neutrality, we: 1) reconstructed the ancestral states of the Sec and Cys residues. To do this: 1) we aligned the 17 codons with orthologous Sec/Cys codons in the vertebrate phylogeny) in the SELENOP transport domain of 32 species (supplementary fig. S6, Supplementary Material online). One Sec codon with no orthology in the frog (*Xenopus tropicalis*) and another one in some non-fish vertebrates (Lobanov, et al. 2008) were not included in the alignment; 2) we used PAML to reconstruct the Sec/Cys state of each codon in the ancestral nodes of the phylogeny for the 32 vertebrate species (supplementary fig. 7, Supplementary Material online); and 3) run neutral simulations of the evolution of the ancestral Sec/Cys states from the root of the vertebrate phylogeny (node N1 in supplementary fig. 7, Supplementary Material online) using a continuous time Markov Chain model of sequence evolution that assumes independence between sites. We performed 10,000 MCMC simulations, with a modified version of Seq-Gen v1.3.2 (Castellano, et al. 2009), in which strongly deleterious mutations (other than between Sec and Cys codons) are immediately eliminated from the population and do not contribute to sequence divergence. We used the standard Hasegawa-Kishino-Yano model of nucleotide evolution (Hasegawa, et al. 1985) whose instantaneous rate of evolution is comprised of a transition/transversion ratio (TS/TV) set to 1.8 (Rosenberg, et al. 2003) with equilibrium nucleotide frequencies set to A = 0.26, T = 0.26, C = 0.24, and G = 0.24, as estimated from 4-fold degenerate sites in 10 vertebrate species ranging from human to *Takifugu* (Margulies, et al. 2005). Branch lengths were set to a proxy of the mean number of neutral mutations per site, the number of synonymous substitutions per site, as estimated by PAML (supplementary fig. 8, Supplementary Material online) using a codon alignment of the SELENOP transport domain of the 32 species considered in our analysis. Positions whose alignment was uncertain were removed using the described probabilistic approach. The distribution of the average number of Se atoms in the transport domain of SELENOP is shown in figure 3 for teleost fishes and terrestrial vertebrates. The same number of simulations were run using this time branch lengths from the number of synonymous substitutions per site from multiple genes across the genome (Castellano, et

al. 2009). The distribution of the average number of Se atoms in the transport domain of SELENOP is shown in supplementary fig. 10, Supplementary Material online.

Acknowledgments

The authors thank Aida M. Andrés for comments on the evolutionary analyses. This work was supported by the Max Planck Society and University College London. Part of this study was funded by the NIHR HS&DR Programme (14/21/45) and supported by the NIHR GOSH BRC. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Tables

Genes
Selenoproteins
Glutathione peroxidase 2 (GPX2)
Iodothyronine deiodinase (DIO) 1 and 2
Selenoprotein F (SELENOF)
Selenoprotein H (SELENOH)
Selenoprotein I (SELENOI)
Selenoprotein K (SELENOK)
Selenoprotein M (SELENOM)
Selenoprotein N (SELENON)
Selenoprotein O (SELENOO)
Selenoprotein S (SELENOS)
Selenoprotein V (SELENOV)
Selenoprotein W (SELENOW1 and teleost fishes SELENOW2)
Thioredoxin reductase (TXNRD) 1, 2 and 3
Cys-containing paralogs
Glutathione peroxidase (GPX) 5, 7, and 8
Methionine-R-sulfoxide reductase (MSRB) 2 and 3
Selenoprotein U (SELENOU) 2 and 3
Rdx12
Regulatory
Transport and uptake of Se into cells
Selenoprotein P (SELENOP)
LRP8 (ApoER2)
LRP2 (Megalin)
Metabolism of Se
Selenocysteine lyase (SCLY)
Selenium binding protein 1 (SELENBP1)
Biosynthesis of Sec
O-phosphoryl tRNASec kinase (PSTK)
Selenophosphate synthetase 2 (SPS2)
O-phosphoseryl-$tRNA^{Sec}$ selenium transferase (SEPSECS)
Seryl-tRNA synthethase (SARS2)
Incorporation of Sec into proteins
CUGBP, Elav-like family member 1 (CELF1)
Elongation factor for Sec (eEFSec)
Eukaryotic translation initiation factor 4A3 (EIF4A3)
ELAV like RNA binding protein 1 (ELAVL1)
Ribosomal protein L30 (RPL30)
SECIS binding protein 2 (SBP2 and SBP2L)
Selenophosphate synthetase 1 (SPS1)
$tRNA^{Sec}$ 1 associated protein 1 (TRNAU1AP) (SECp43)
Exportin 1 (XPO1)

Table 1. Candiadate genes grouped according to their type and the biological process they participate in. SELENOP and SPS2 are selenoproteins with a regulatory role.

		dN/dS			Fold change		
		Pre-duplication	Post-duplication		Ancestor vs		Copy 1 vs Copy 2
	Genes	Ancestor	Copy 1	Copy 2	Copy 1	Copy 2	
Ancestral whole-genome duplication	**DIO3**	0.098	0.1239	0.1296	1.322	1.264	1.046
	GPX3	0.138	0.192	0.6283	1.387	4.540	3.272
	GPX4	0.199	0.1602	0.221	0.804	1.109	1.380
	SELENOT	0.096	0.1077	0.1083	1.122	1.128	1.006
	SELENOU1	0.149	0.2664	0.2835	1.794	1.909	1.064
Lineage-specific duplication	**SELENOT1**	0.1338	0.0954	0.0731	0.713	0.546	0.766
	SELENOU1b	0.266	0.3443	0.3674	1.292	1.379	1.067
	SELENOJ	0.1201	0.2734	0.1507	2.276	1.255	0.551
	SELENOO	0.1312	0.1067	0.0654	0.813	0.498	0.613

Table 2. Rate of evolution before and after the specific gene duplications in the teleost fishes (supplementary fig. 5, Supplementary Material online). GPX1, MSRB1 and SELENOW2 are not shown as one of the copies of these genes lacks either synonymous or non-synonymous changes.

Figures

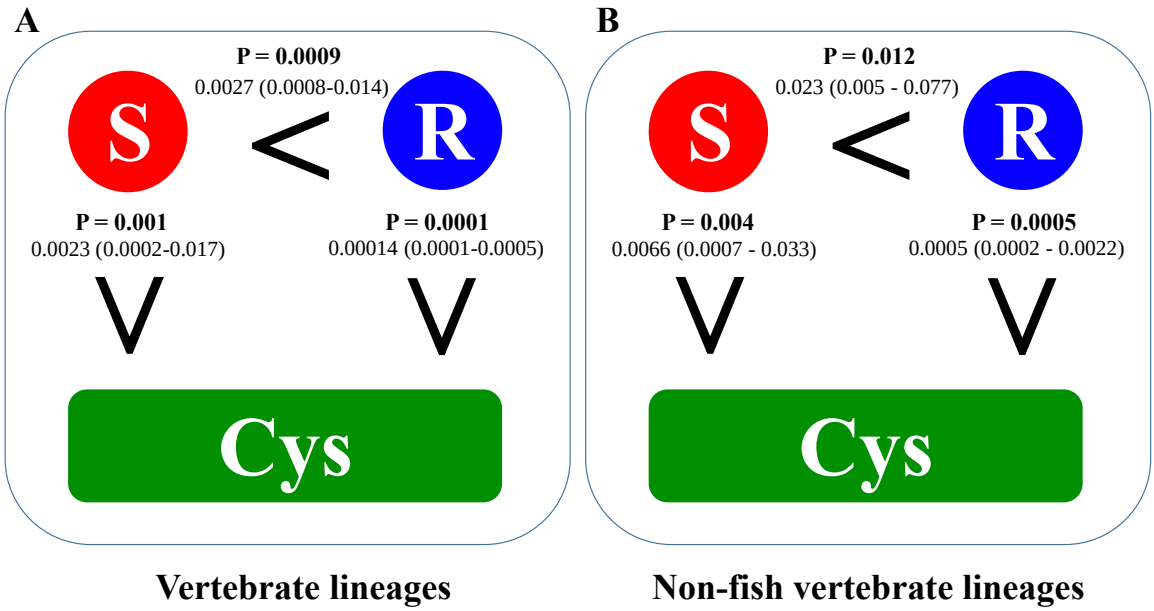

Figure 1. Variation in the strength of natural selection, as measured by the dN/dS ratio, between the groups of selenoprotein genes (S for selenoprotein), their Cys-containing paralogs (Cys) and the genes that regulate the metabolism and homeostasis of Se (R for regulatory). The significance of these comparisons is given for the available gene annotation in 44 vertebrate species and for 200 random samples taken from these gene annotations (median and 95% confidence interval). (**A**) vertebrate lineages and (**B**) non-fish (terrestrial) vertebrate lineages only.

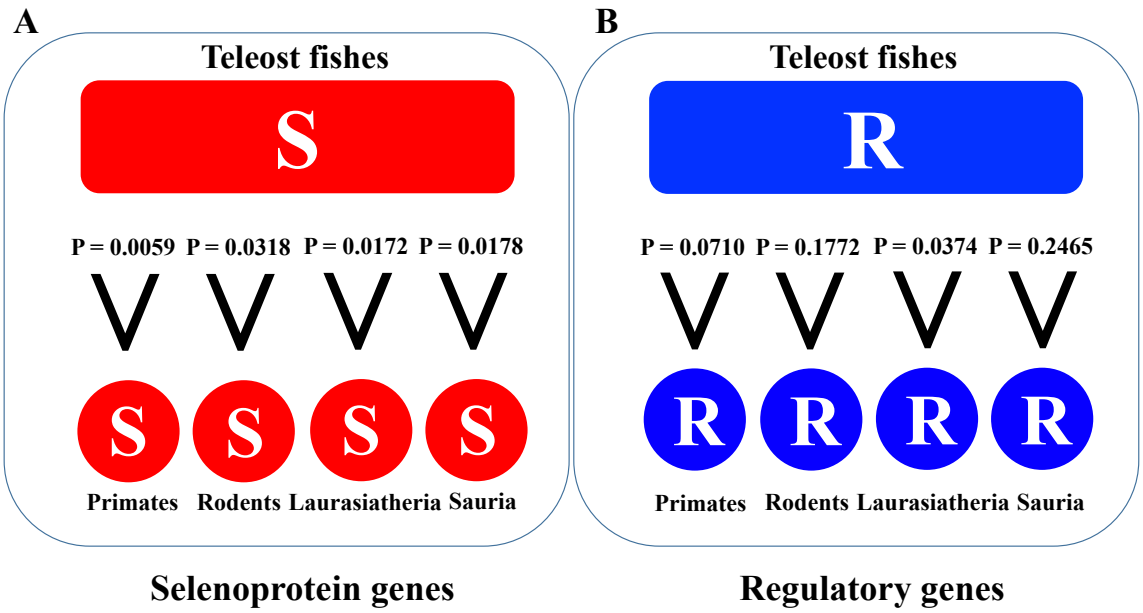

Figure 2. Variation in the strength of natural selection, as measured by the dN/dS ratio, between the teleost fish and the different terrestrial vertebrate clades. (**A**) selenoprotein genes and (**B**) genes that regulate the metabolism and homeostasis of Se.

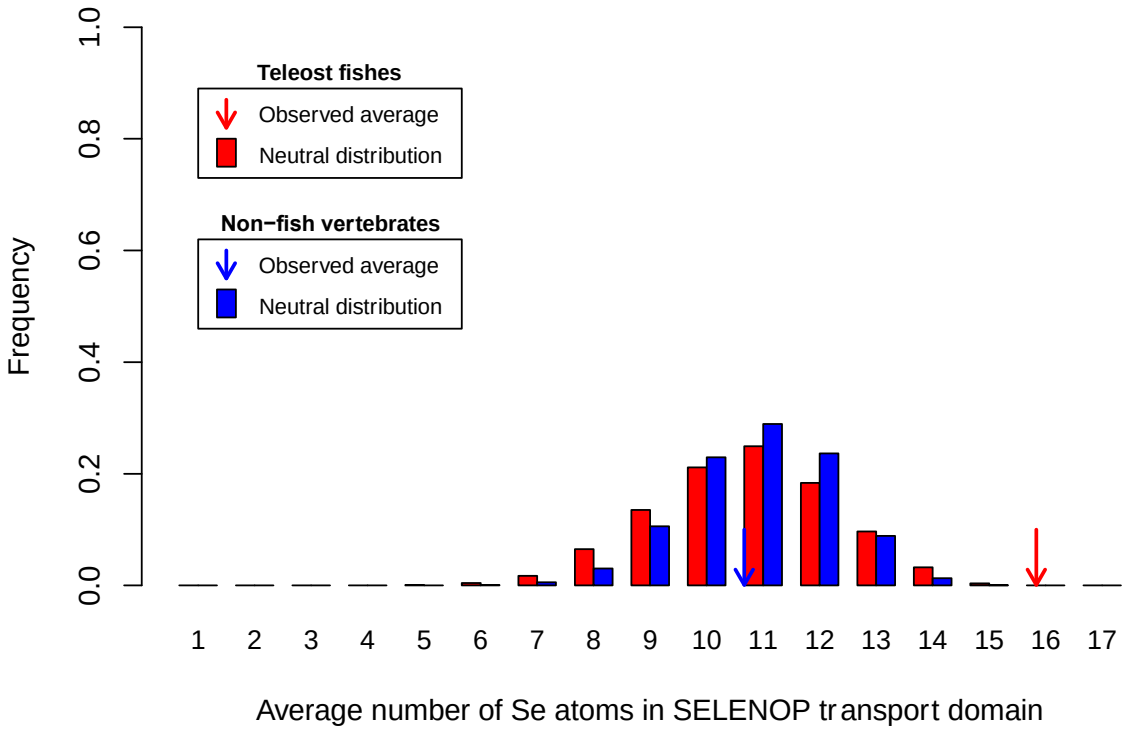

Figure 3. Expected distribution under neutrality for the exchange of Se and sulfur (in the form of Sec and cysteine, respectively) in the transport domain of selenoprotein P (SELENOP) of teleost fishes and non-fish vertebrates. The distribution of the average number of Sec in the neutral simulations is shown and compared to the observed one. The neutral distributions are the result of 10,000 simulations of the divergence process along the vertebrate phylogeny (supplementary fig. 8, Supplementary Material online).

References

Brunet FG, Roest Crollius H, Paris M, Aury JM, Gibert P, Jaillon O, Laudet V, Robinson-Rechavi M. 2006. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. Mol Biol Evol 23:1808-1816.

Burk RF, Olson GE, Hill KE, Winfrey VP, Motley AK, Kurokawa S. 2013. Maternal-fetal transfer of selenium in the mouse. FASEB J 27:3249-3256.

Castellano S, Andres AM, Bosch E, Bayes M, Guigo R, Clark AG. 2009. Low exchangeability of selenocysteine, the 21st amino acid, in vertebrate proteins. Mol Biol Evol 26:2031-2040.

Castellano S, Lobanov AV, Chapple C, Novoselov SV, Albrecht M, Hua D, Lescure A, Lengauer T, Krol A, Gladyshev VN, et al. 2005. Diversity and functional plasticity of eukaryotic selenoproteins: identification and characterization of the SelJ family. Proc Natl Acad Sci U S A 102:16188-16193.

Christoffels A, Koh EG, Chia JM, Brenner S, Aparicio S, Venkatesh B. 2004. Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. Mol Biol Evol 21:1146-1151.

Dormitzer PR, Ellison PT, Bode HH. 1989. Anomalously low endemic goiter prevalence among Efe pygmies. Am J Phys Anthropol 78:527-531.

Durbin R, Durbin R, Eddy S, Krogh A, Mitchison G. 1998. Biological sequence analysis : probabilistic models of proteins and nucleic acids. Cambridge, United Kingdom: Cambridge University Press.

Eddy SR. 1998. Profile hidden Markov models. Bioinformatics 14:755-763.

Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, et al. 2013. Ensembl 2013. Nucleic Acids Res 41:D48-55.

Gladyshev VN, Arner ES, Berry MJ, Brigelius-Flohe R, Bruford EA, Burk RF, Carlson BA, Castellano S, Chavatte L, Conrad M, et al. 2016. Selenoprotein Gene Nomenclature. J Biol Chem 291:24036-24040.

Gromer S, Johansson L, Bauer H, Arscott LD, Rauch S, Ballou DP, Williams CH, Jr., Schirmer RH, Arner ES. 2003. Active sites of thioredoxin reductases: why selenoproteins? Proc Natl Acad Sci U S A 100:12618-12623.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160-174.

Hatfield D. 1985. Suppression of Termination Codons in Higher Eukaryotes. Trends in Biochemical Sciences 10:201-204.

Hoegg S, Brinkmann H, Taylor JS, Meyer A. 2004. Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. J Mol Evol 59:190-203.

Howe DG, Bradford YM, Conlin T, Eagle AE, Fashena D, Frazer K, Knight J, Mani P, Martin R, Moxon SA, et al. 2013. ZFIN, the Zebrafish Model Organism Database: increased support for mutants and transgenics. Nucleic Acids Res 41:D854-860.

Hurley IA, Mueller RL, Dunn KA, Schmidt EJ, Friedman M, Ho RK, Prince VE, Yang Z, Thomas MG, Coates MI. 2007. A new time-scale for ray-finned fish evolution. Proc Biol Sci 274:489-498.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 11:97-108.

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, et al. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 431:946-957.

Johnson CC, Fordyce FM, Rayman MP. 2010. Symposium on 'Geographical and geological influences on nutrition': Factors controlling the distribution of selenium in the environment and their impact on health and nutrition. Proc Nutr Soc 69:119-132.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772-780.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095-1109.

Lobanov AV, Fomenko DE, Zhang Y, Sengupta A, Hatfield DL, Gladyshev VN. 2007. Evolutionary dynamics of eukaryotic selenoproteomes: large selenoproteomes may associate with aquatic life and small with terrestrial life. Genome Biol 8:R198.

Lobanov AV, Hatfield DL, Gladyshev VN. 2008. Reduced reliance on the trace element selenium during evolution of mammals. Genome Biol 9:R62.

Lopez Herraez D, Bauchet M, Tang K, Theunert C, Pugach I, Li J, Nandineni MR, Gross A, Scholz M, Stoneking M. 2009. Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. PLoS One 4:e7888.

Mann HB, Whitney DR. 1947. On a Test of Whether One of 2 Random Variables Is Stochastically Larger Than the Other. Annals of Mathematical Statistics 18:50-60.

Marelli F, Carra S, Agostini M, Cotelli F, Peeters R, Chatterjee K, Persani L. 2016. Patterns of thyroid hormone receptor expression in zebrafish and generation of a novel model of resistance to thyroid hormone action. Mol Cell Endocrinol 424:102-117.

Margulies EH, Program NCS, Maduro VV, Thomas PJ, Tomkins JP, Amemiya CT, Luo M, Green ED. 2005. Comparative sequencing provides insights about the structure and conservation of marsupial and monotreme genomes. Proc Natl Acad Sci U S A 102:3354-3359.

Mariotti M, Ridge PG, Zhang Y, Lobanov AV, Pringle TH, Guigo R, Hatfield DL, Gladyshev VN. 2012. Composition and evolution of the vertebrate and mammalian selenoproteomes. PLoS One 7:e33066.

Mendieta-Serrano MA, Schnabel D, Lomeli H, Salas-Vidal E. 2015. Spatial and temporal expression of zebrafish glutathione peroxidase 4 a and b genes during early embryo development. Gene Expr Patterns 19:98-107.

Mertz W. 1981. The essential trace elements. Science 213:1332-1338.

Mills CF editor. International Symposium on Trace Element Metabolism in animals. 1974 Baltimore.

Ogle RS, Maier KJ, Kiffney P, Williams MJ, Brasher A, Melton LA, Knight AW. 1988. Bioaccumulation of Selenium in Aquatic Ecosystems. Lake and Reservoir Management:165-173.

Penglase S, Hamre K, Ellingsen S. 2015. The selenium content of SEPP1 versus selenium requirements in vertebrates. PeerJ 3:e1244.

Pietschmann N, Rijntjes E, Hoeg A, Stoedter M, Schweizer U, Seemann P, Schomburg L. 2014. Selenoprotein P is the essential selenium transporter for bones. Metallomics 6:1043-1049.

Rayman MP. 2012. Selenium and human health. Lancet 379:1256-1268.

Romagne F, Santesmasses D, White L, Sarangi GK, Mariotti M, Hubler R, Weihmann A, Parra G, Gladyshev VN, Guigo R, et al. 2014. SelenoDB 2.0: annotation of selenoprotein genes in animals and their genetic diversity in humans. Nucleic Acids Res 42:D437-443.

Rosenberg MS, Subramanian S, Kumar S. 2003. Patterns of transitional mutation biases within and among mammalian genomes. Mol Biol Evol 20:988-993.

Schomburg L, Schweizer U. 2009. Hierarchical regulation of selenoprotein expression and sex-specific effects of selenium. Biochim Biophys Acta 1790:1453-1462.

Selinus O, Alloway B, Centeno JA, Finkleman R, B,, Fuge R, Lindh U, Smedley P. 2005. Essentials of Medical Geology: Impacts of the Natural Environment on Public Health. Burlington: Elsevier Academic Press.

Shetty SP, Shah R, Copeland PR. 2014. Regulation of selenocysteine incorporation into the selenium transport protein, selenoprotein P. J Biol Chem 289:25317-25326.

Snider GW, Ruggles E, Khan N, Hondal RJ. 2013. Selenocysteine confers resistance to inactivation by oxidation in thioredoxin reductase: comparison of selenium and sulfur enzymes. Biochemistry 52:5472-5481.

Squires JE, Stoytchev I, Forry EP, Berry MJ. 2007. SBP2 binding affinity is a major determinant in differential selenoprotein mRNA translation and sensitivity to nonsense-mediated decay. Mol Cell Biol 27:7848-7855.

Steinmann D, Nauser T, Koppenol WH. 2010. Selenium and sulfur in exchange reactions: a comparative study. J Org Chem 75:6696-6699.

Stewart R, Grosell M, Buchwalter DB, Fisher NS, Luoma S, Mathews T, Orr PL, Wang WX. 2010. Bioaccumulation and Trophic Transfer of Selenium. In: Environment. EAoSitA, editor. Boca Raton: SETAC in collaboration with CRC Press. p. 93-139.

Sunde RA. 2014. Selenium. In. Modern Nutrition in Health and Disease. Philadelphia: Wolters Kluwer Health.

Vandepoele K, De Vos W, Taylor JS, Meyer A, Van de Peer Y. 2004. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. Proc Natl Acad Sci U S A 101:1638-1643.

White L, Romagne F, Muller E, Erlebach E, Weihmann A, Parra G, Andres AM, Castellano S. 2015. Genetic adaptation to levels of dietary selenium in recent human history. Mol Biol Evol 32:1507-1518.

WHO. 1987. Selenium. In. Geneva: World Health Organization.

Wilber CG. 1980. Toxicology of selenium: a review. Clin Toxicol 17:171-230.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-1591.