Supplementary Notes to: Processive recoding and metazoan evolution of Selenoprotein P: up to 132 UGAs in molluscs

Janinah Baclaocos[1‡], Didac Santesmasses[1,2,3‡], Marco Mariotti[1,2,3‡*], Katarzyna Bierła[4], Michael B. Vetick[5], Sharon Lynch[6], Rob McAllen[6], John J. Mackrill[7], Gary Loughran[1], Roderic Guigó[2], Joanna Szpunar[4], Paul R. Copeland[5], Vadim N Gladyshev[3], John F. Atkins[1]

Supplementary Note 1: Sec to Cys conversions in vertebrate *SelenoP2*

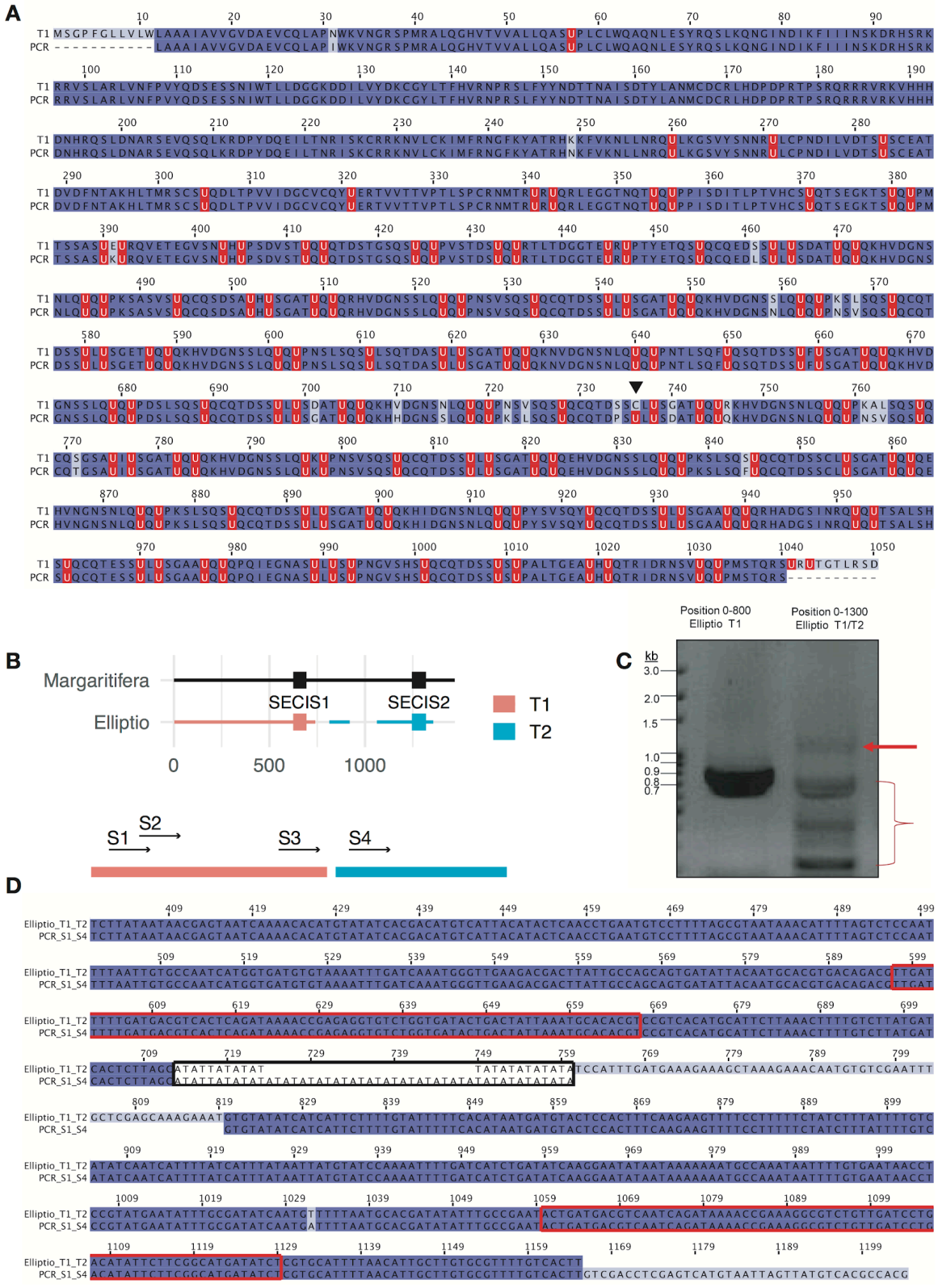
Inspecting the reconstructed protein tree of *SelenoP* genes (Supplementary Fig 1) as well as the species tree (Fig 2), we noticed the occurrence of events wherein the only Sec-UGA codon in *SelenoP2*, corresponding to UGA1 in *SelenoP1*, was converted to a Cys codon. One such event was observed within Amphibia: all Anura (including frogs) carry a *SelenoP2* gene without Sec-UGA codons, while salamanders (Caudata) possess a single-Sec codon *SelenoP2* like the rest of non-placental vertebrates. The mRNA sequences of Anura *SelenoP2* lack SECIS elements, and form a cluster separated from the rest of *SelenoP2* genes, at the base of all vertebrate *SelenoP* sequences (Supplementary Fig 1). This observation is consistent with increased protein divergence concomitant with Sec to Cys conversion, resulting in long-branch attraction in tree reconstruction. Increased divergence occurred also for various other (but not all) Sec to Cys conversions across metazoa. We observed another UGA1 to Cys codon conversion in *SelenoP2* of Galeomorphii sharks (cartilaginous fish), supported by numerous sequences. This also appears to have occurred concomitant with SECIS loss, but it did not result in obvious traces of increased protein divergence. Other fishes possess a *SelenoP1* or *SelenoP2* with a Cys codon replacing UGA1; however, these constitute isolated cases supported by single sequences.

Supplementary Note 2: Validation of 132 UGAs in *SelenoP* of *Elliptio complanata*

To confirm the exceptional number of Sec-UGAs in *Elliptio complanata SelenoP* mRNA, we extracted RNA from these mussels, then performed RT-PCR followed by Sanger sequencing (see Methods). Results are shown in Supplementary Note 2 Figure, here below. While we identified 131 UGAs in the genome sequence, we observed an additional one with RT-PCR, resulting from a Cys codon converted to Sec (Supplementary Note 2 Figure, A). This suggests that variants with diverse numbers of Sec-UGAs are present in *E. complanata* populations. The *SelenoP* transcript that we initially identified in the genome contained only a single SECIS element. A separate contig containing the missing SECIS 2 was fished out from the transcriptome using the sequence of a related species (*Margaritifera margaritifera*). Sequence searches with SECIS 2 from *M. margaritifera* allowed us to identify a separate *E. complanata* transcriptome contig that contained the missing SECIS 2, and enabled its sequence determination. We ascribe the apparent disconnection of the two contigs to an assembly artefact due to a stretch of highly repetitive AT-rich sequence between the two SECIS elements (Supplementary Note 2 Figure, B-D).

Supplementary Note 2 Figure (next page): *Elliptio complanata SelenoP* mRNA has 132 UGAs and 2 SECIS elements. (A) Amino acid sequence alignment of *Elliptio complanata* SelenoP predicted from its transcriptome assembly (NCBI GAHW00000000.1; transcript T1 = GAHW01001216.1) and the translation of the consensus sequence obtained by PCR sequencing. Red denotes Sec residues, dark blue are matched residues and light blue are amino-acid substitution. Position 736 (arrowhead) shows a Cys to Sec conversion (B) Location of SECIS 1 in T1 and SECIS 2 in a separate contig T2 (NCBI GAHW01060232.1), aligned to the 3'UTR of *SelenoP* from the related freshwater pearl mussel *Margaritifera margaritifera* (NCBI GFHD01004612.1). (C) Agarose gel of PCR amplification of *E. complanata SelenoP* 3'UTR from synthesised *E. complanata* cDNA. Lane 1 shows product amplification of primers spanning *E. complanata* T1 from positions 0-796 and lane 2 shows product amplification from positions 0 to 1300 spanning sequences of T2. Red arrow points to position of product excised for sequencing while red bracket depicts products from possible alternative primer binding sites. (D) Nucleotide sequence alignment using S1 and S4 sequencing primers binding at position indicated on top. Sequences of SECIS 1 (597-667) and SECIS 2 (1060-1129) are boxed in red; AT repeats (713-760) are boxed in black.

Supplementary Note 2 Figure



Supplementary Note 3: Sec to Cys conversion of *SelenoP* in gastropods

The gastropod owl limpet (*Lottia gigantea*; Patellogastropoda) has a Sec codon at the characteristic position of UGA1. However, in the gastropod lineage comprising Heterobranchia and Caenogastropoda, we observed that this Sec was replaced by a Cys codon. Their sequences form a separate cluster in the gene tree (Supplementary Fig 1), suggestive of increased divergence rate. In gastropods as a whole, we did not detect more than one SECIS, despite the presence of two SECIS elements in many non-gastropod multi-Sec codon mollusc *SelenoPs*. While the phylogenetic relationships of gastropods are still uncertain, the placement of the Patellogastropoda *SelenoP* (owl limpet) together with Heterobranchia and Caenogastropoda apparently contradicts the most widely accepted taxonomy, as Patellogastropoda is considered an early-branching gastropod (Zapata et al. 2014). This suggests a complex gene history for gastropod *SelenoP*.

Supplementary Note 3 reference:

- Zapata F, Wilson NG, Howison M, Andrade SC, Jorger KM, Schrodl M, et al. Phylogenomic analyses of deep gastropod relationships reject Orthogastropoda. Proceedings Biological sciences. 2014;281:20141739.

Supplementary Note 4: Divergence of *SelenoP* in arachnida

Within arachnida, we observed three phylogenetic clusters located in a basal position in the reconstructed protein tree (Supplementary Fig 1). Despite this topology, there is clear homology between the C-terminal domains of all these clusters (Supplementary Fig 2). This strongly indicates that these clusters actually constitute a single orthologous group with diverse divergence rates, resulting in long-branch attraction. The first cluster is comprised of sequences from Araneae (spiders, such as *Parasteatoda tepidariorum*). These genes carry multiple Sec codons and two SECIS elements. UGA1 is present at its common location, while the other UGAs are in the distal region and feature obvious modularity (see below). The second cluster consists of sequences from Ixodoidea (ticks, such as *Ixodes scapularis*). These genes have a single UGA which is at the classic UGA1 position. SECIS was identified in some sequences, but not all, possibly due to incomplete transcripts or high SECIS divergence. The third arachnid *SelenoP* cluster consisted of Acariformes (mites, such as *Leptotrombidium deliense*). This group included both genes with an in-frame UGA corresponding to UGA1, and genes with diverse codons in its place. Surprisingly, we could not detect any SECIS elements in this group. Future research will clarify whether these genes encode selenoproteins, or if their mRNAs are translated through a different, perhaps novel, mechanism of UGA recoding.

Supplementary Note 5: *SECISBP2* orthology between vertebrates and oysters

In addition to *SECISBP2*, humans also have a paralog in the same family, *SECISBP2L*. These two genes likely originated by gene duplication in the vertebrate branch (Donovan and Copeland, 2009). In vertebrates, SECISBP2 supports Sec insertion, although it is not required in some cases and has a role in mRNA stability (Seeher et al., 2014; Fradejas-Villar et al., 2017). SECISBP2L does not support Sec insertion on its own, and its function is still unclear (Donovan and Copeland, 2012). Invertebrates contain a single gene in this protein family, corresponding to the metazoan ancestral state, prior to the vertebrate duplication. This gene shows higher sequence similarity to vertebrate *SECISBP2L* than *SECISBP2* (Donovan and Copeland, 2009; Supplementary Fig 7). However, the protein from the annelid *Capitella teleta* is competent for binding the human SECIS elements studied (Donovan and Copeland, 2012), and thus is clearly functional orthologous to vertebrate *SECISBP2*. Because of this, we will refer to the *M. gigas* homolog as *SECISBP2*, though the name *SECISBP2L* (used in Donovan and Copeland, 2012) would have been just as valid.

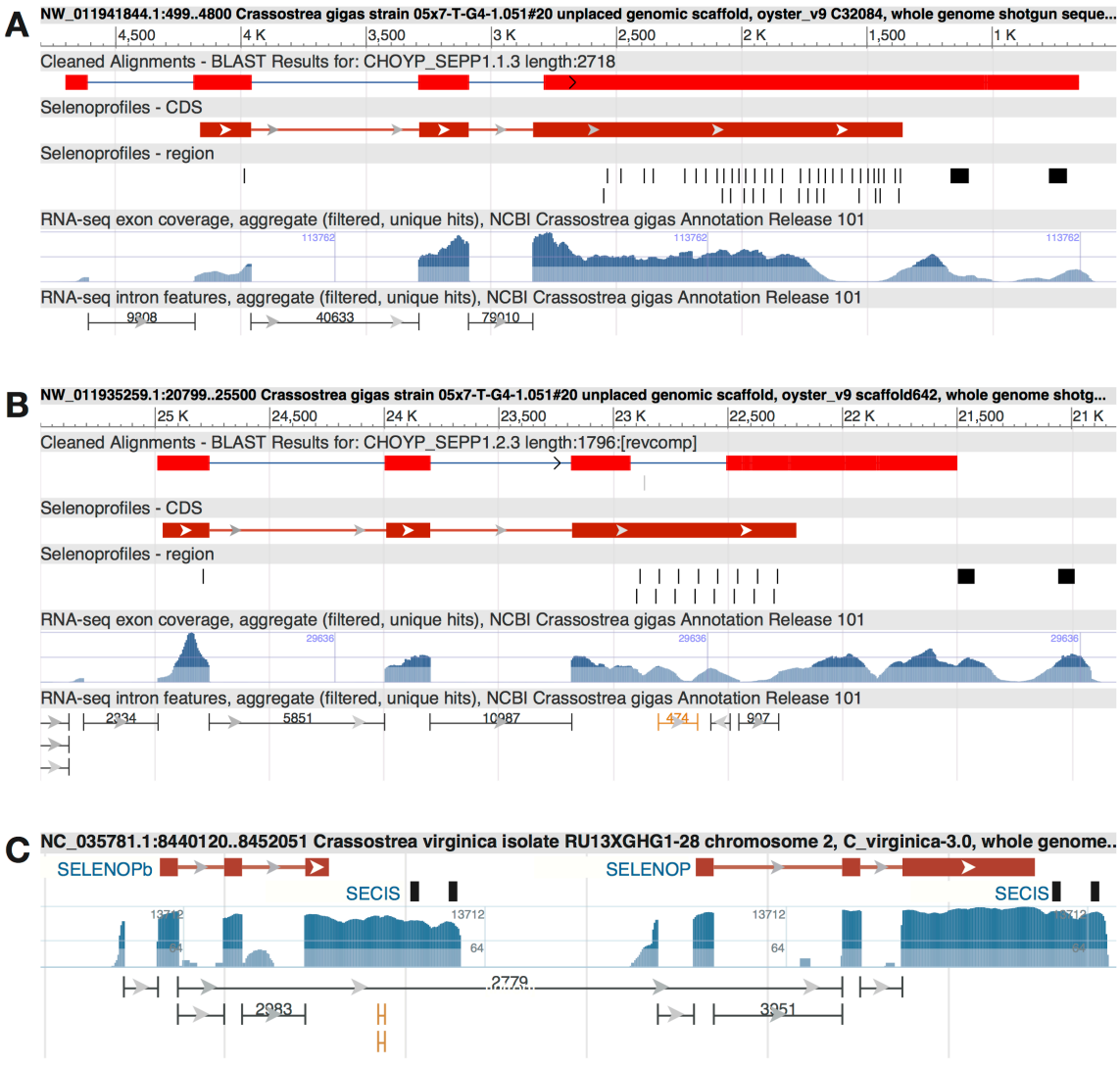
Supplementary Note 5 references:

- Donovan J, Copeland PR. Evolutionary history of selenocysteine incorporation from the perspective of SECIS binding proteins. BMC evolutionary biology. 2009;9:229.
- Seeher S, Atassi T, Mahdi Y, Carlson BA, Braun D, Wirth EK, et al. Secisbp2 is essential for embryonic development and enhances selenoprotein expression. Antioxidants & redox signaling. 2014;21:835-49.
- Fradejas-Villar N, Seeher S, Anderson CB, Doengi M, Carlson BA, Hatfield DL, et al. The RNA-binding protein Secisbp2 differentially modulates UGA codon reassignment and RNA decay. Nucleic acids research. 2017;45:4094-107.
- Donovan J, Copeland PR. Selenocysteine insertion sequence binding protein 2L is implicated as a novel post-transcriptional regulator of selenoprotein expression. PloS one. 2012;7:e35581.

Supplementary Note 6: *SelenoP* duplication in oysters

Besides the oyster 46-UGA *SelenoP* gene described in the text and subjected to several experimental analyses, we identified a second gene in this protein family in the *M. gigas* genome, which was also confirmed in the transcriptome (Supplementary Note 6 Figure, below). The second gene, here referred to as oyster *SelenoPb* (SELENOP.1 in Supplementary Table 1), had two SECIS elements, but the number of in-frame UGAs was unclear: the predicted CDS from the genome had 17 UGAs, whereas the assembled transcript had only 5. The C-terminal domain, which contains multiple instances of a repeat containing two UGAs, was shorter in the transcript. We attributed the differences between the two sequences to an assembly artefact in either the genome or transcriptome. A *SelenoP* paralog was also found in the genome of a related species, eastern oyster *Crassostrea virginica*. In this case, the second gene has two UGAs and two SECIS elements. The exonic structure is conserved between oyster *SelenoPb* and *SelenoP* in both *M. gigas* and *C. virginica*. *SelenoP* and *SelenoPb* are in tandem in *C. virginica*, but not in *M. gigas*. The remainder of bivalves analyzed here had only one gene in this protein family. These observations suggest that oyster *SelenoPb* appeared by tandem duplication in the common ancestor of these two *Crassostrea* species (*M. gigas* and *C. virginica*).

Supplementary Note 6 Figure: Duplication of *SelenoP* in *Crassostrea*. (A) Intron exon disposition at the *SelenoP* locus in the *M. gigas* genome. Tracks (top to bottom): (1) Blast alignment of the transcript sequence (CHOYP_SEPP1.1.3) onto the genome; (2) our prediction of *SelenoP* coding sequence (CDS); (3) position of the 46 UGAs (thin lines), shown in two different rows to avoid overlaps, and position of SECIS 1 and 2 (blocks); (4) RNA-seq aggregate exon coverage; and (5) RNA-seq intron features. (B) Counterpart for *SelenoPb* with the same tracks as panel A. The top track shows the transcript CHOYP_SEPP1.2.3 mapped onto the genome. Note that the transcript lacks a portion of the C-terminal domain predicted in the genome (track 2) and includes only the last four UGAs (track 3). The transcript also lacks the two SECIS elements, although the RNA-seq coverage (track 4) shows transcription up to SECIS 2. (C) *C. virginica*: Genomic view of *SelenoPb* (left) and *SelenoP* (right) loci which are located in tandem on chromosome 2. Tracks (top to bottom): (1) CDS prediction (red) and SECIS 1 and 2 (black). (2) RNA-seq aggregate exon coverage; and (3) RNA-seq intron features. Images generated using NCBI GDV browser. The corresponding contigs and assembly versions used are indicated on top of each panel.

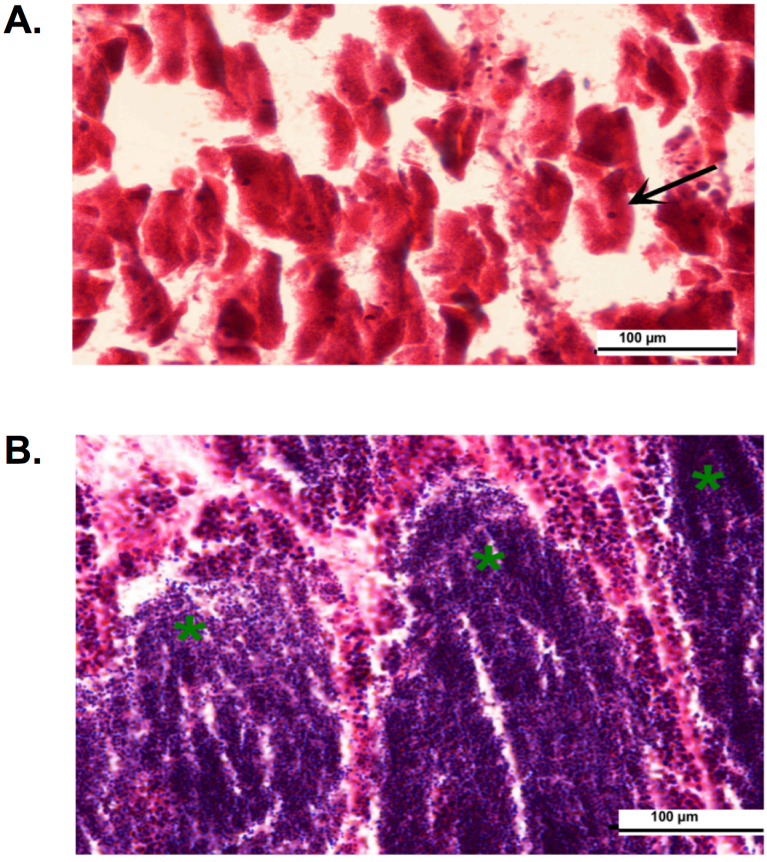


Supplementary Note 7: Oysters for selenium supplementation experiments

It is known that mammalian brain and testes preferentially take up long forms of SelenoP from plasma (Kurokawa et al. 2014). Analysis of publicly available RNA-seq datasets revealed higher expression levels of *SelenoP* RNA in fully developed male oysters compared to both their female counterparts, and also oysters at earlier stages of development (Zhang et al., 2012; Riviere et al., 2015). To study high level SelenoP translation, we thus chose to utilize adult male oysters with developed gonads, which we identified through nuclei-staining of male gametes and mature follicle sacs from histological preparations (Supplementary Note 7 Figure, below).

Supplementary Note 7 references:

- Kurokawa S, Eriksson S, Rose KL, Wu S, Motley AK, Hill S, et al. Sepp1(UF) forms are N-terminal selenoprotein P truncations that have peroxidase activity when coupled with thioredoxin reductase-1. Free radical biology & medicine. 2014;69:67-76.
- Zhang G, Fang X, Guo X, Li L, Luo R, Xu F, et al. The oyster genome reveals stress adaptation and complexity of shell formation. Nature. 2012;490:49-54.
- Riviere G, Klopp C, Ibouniyamine N, Huvet A, Boudry P, Favrel P. GigaTON: an extensive publicly searchable database providing a new reference transcriptome in the pacific oyster Crassostrea gigas. BMC bioinformatics. 2015;16:401.



Supplementary Note 7 Figure: Histology of oyster gonads stained with haematoxylin and eosin stains. (A) The female gonad is stained with varying shades of red/pink. The arrow points to the blue-black stain of the nuclei of primary oocytes. (B) Male oyster gonads show purple staining of sperm cells nuclei. Green asterisks point to the follicle sacs in which the sperm cells are stored.

Supplementary Note 8: Assessment of ribosome profiling data

We assessed the quality of the high-throughput sequencing reads used in this work. Four libraries were prepared, two for riboseq and two for RNA-seq, from the non-supplemented and Se-supplemented samples (Methods). The total number of reads obtained is shown below in Supplementary Note 8 table. The Supplementary Note 8 figure (next page) contains panels showing various features of sequencing data, referenced hereafter. Riboseq shows a narrow distribution of RPF length centered at ~30bp as expected, with no substantial differences between supplementation groups (panel A). Mapped RPFs show moderate phasing in CDS, that is absent in RNAseq (B). A metagene analysis shows that the great majority of riboseq reads map to annotated CDS regions, with a sharp decrease corresponding to translation start and stop sites (C). RNAseq, in contrast, show continuous coverage. From the metagene analysis, we noticed an extra peak overlapping the start site (C, position 0), which is more pronounced in the non-supplemented sample. The peak is also present when reads are aligned by the second in-frame AUG codon in their coding sequence (D). We thus ascribe this peak to context dependent bias of ribosome profiling, due to sequence preference of the enzymes used for digestion or ligation. Analogously, the analysis of other codons shows a pattern of codon-specific enrichment or depletion (not shown). We do not know why the degree of this bias differs between supplementation groups. It may be a biological effect of added selenium, or a chemical effect on the riboseq protocol, or simply an artifactual batch effect (though all samples were prepared in parallel). Nevertheless, while evident from metagene analysis where thousands of sequences are stacked together aligned by their AUG, this is a very mild effect for any given gene and position. This is evident from the inspection of individual ribosome coverage maps of a few house-keeping genes (F). Thus, for any given gene, the biological effects of Se-related regulation most likely still dominate the differences between supplementation groups. For *SelenoP* (E), there is only a negligible difference of RPF density at position 0 (starting AUG) between samples; in contrast, we see a remarkable decrease of a peak at the -12 position upon selenium supplementation. This corresponds to a robust difference in magnitude of raw read counts, from >100 in the non-supplemented sample to <10 in Se-supplemented (G), which cannot be accounted by context dependent bias. We believe this reflects a selenium-dependent regulatory mechanism of translation initiation, likely involving the adjacent ISL structure.

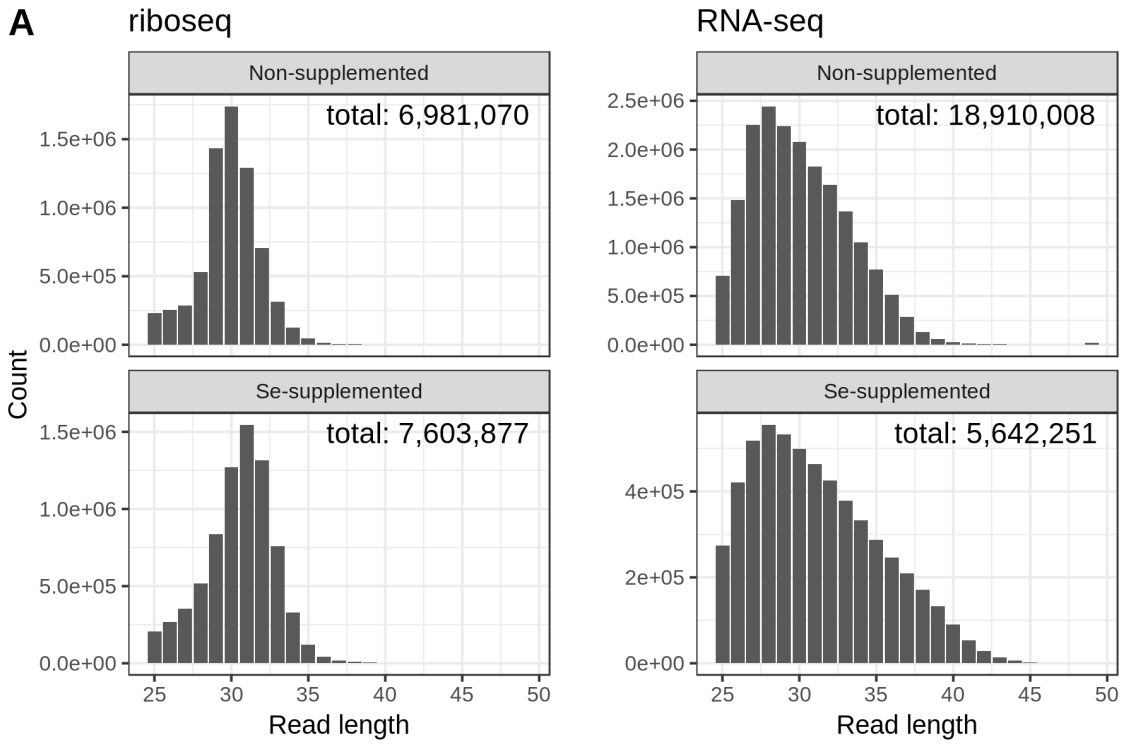
Supplementary Note 8 Table:

	riboseq		RNA-seq	
	Non-supp	Se-supp	Non-supp	Se-supp
Total reads	42,968,039	56,454,982	24,125,976	7,308,988
Mapped reads* (%)	6,981,070 (16.2)	7,603,877 (13.5)	18,910,008 (78.4)	5,642,251 (77.2)
Genes** (%)	25,802 (82.9)	24,997 (80.3)	23,001 (73.9)	23,549 (75.6)

* reads mapping to CDS for Riboseq; full mRNA for RNA-seq

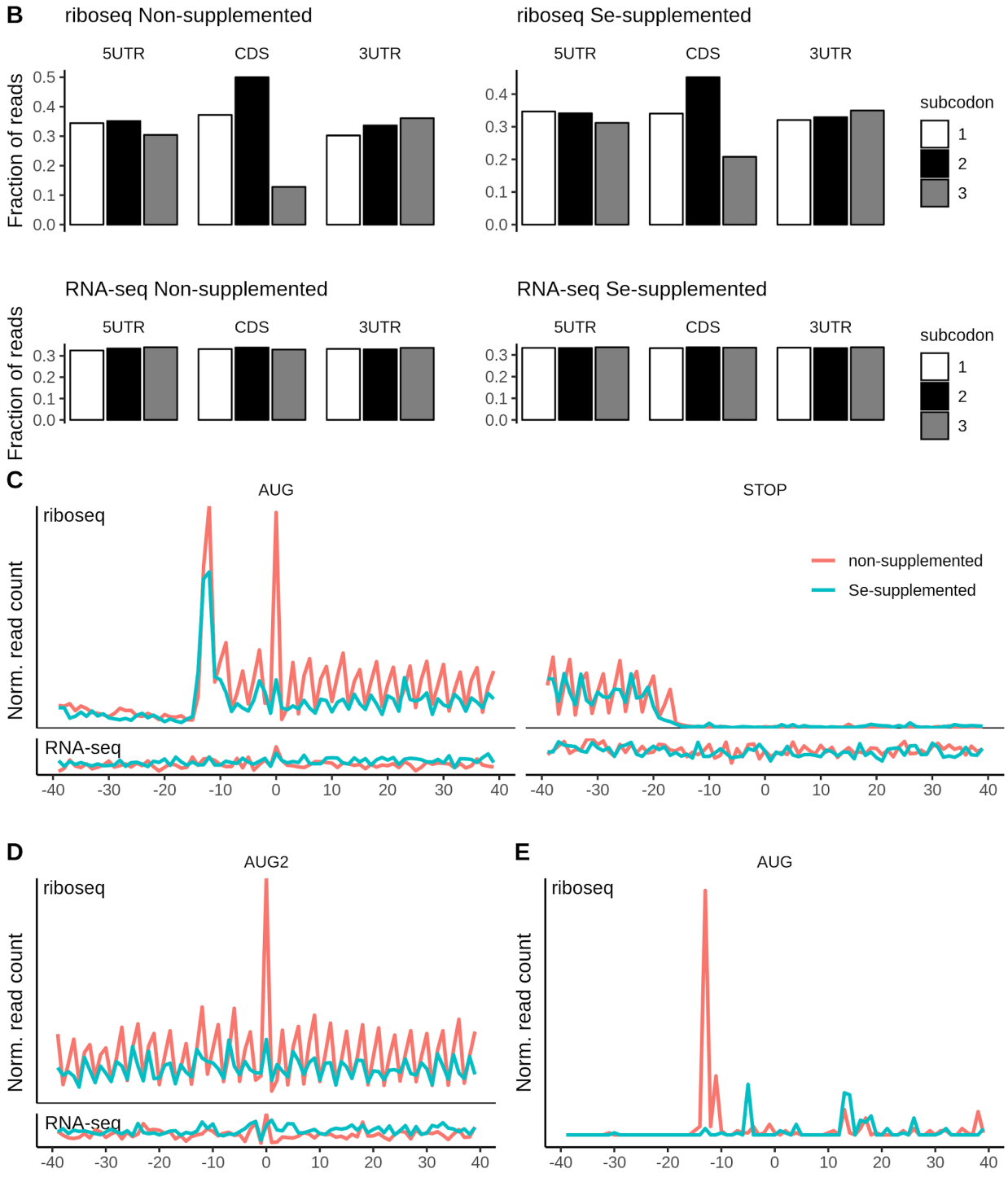
**RPFKM (riboseq) or RPKM (RNA-seq) >= 1; percentage computed over the total number of protein coding genes in the transcriptome annotation: 31,142.

Supplementary Note 8 Figure:
(A) Length distribution in riboseq and RNA-seq fragments (i.e., reads with adapters removed), for non-supplemented and Se-supplemented libraries. Only reads mapping to the coding sequence are shown for riboseq, whereas all mapped reads are shown for RNA-seq. The total number of reads included in the distribution is indicated. (B-E) Genome-wide metagene analysis of ribosome profiling footprints and RNA-seq reads for quality control. Only reads with length between 28 and 32 were considered for riboseq. (B) Fraction of ribosome footprints and RNA-seq reads per frame for non-supplemented and selenium-supplemented samples. (C) Abundance of ribosome footprints and RNA-seq reads (5' end) relative to the start (left) and stop (right) codons across all mRNAs in the transcriptome assembly of *M. gigas*. Ribosome footprints abruptly start at position -12 from start and end at position -15 from stop, corresponding to ribosomal P- and A-site. (D) Abundance of ribosome footprints (5' end) relative to the second AUG codon in all transcripts, and (E) relative to initiation AUG codon in *SelenoP*. (F) riboseq and RNA-seq mRNA coverage for four housekeeping genes from the transcriptome assembly of *M. gigas*. (G) Raw riboseq read counts for *SelenoP* mRNA in non-supplemented and Se-supplemented samples.



(Figure continues in next page)

Supplementary Note 8 Figure (continued):



(Figure continues in next page)

Supplementary Note 8 Figure (continued):