MAIN TEXT

Identification of LINE retrotransposons and long non-coding RNAs expressed in the octopus brain

Giuseppe Petrosino[1,*], Giovanna Ponte[1,*], Massimiliano Volpe[1,*], Ilaria Zarrella[1], Concetta Langella[1], Giulia Di Cristina[1], Sara Finaurini[2], Monia T. Russo[3], Swaraj Basu[1], Francesco Musacchia[1], Filomena Ristoratore[1], Dinko Pavlinic[4], Vladimir Benes[4], Maria I. Ferrante[3], Caroline Albertin[5], Oleg Simakov[6,7], Stefano Gustincich[8,2], Graziano Fiorito[1,#], Remo Sanges[1,2,8,#]

[1] Department of Biology and Evolution of Marine Organisms, Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Napoli (SZN), Italy.

[2] Neurobiology Sector, Scuola Internazionale Superiore di Studi Avanzati (SISSA), Via Bonomea 265, 34136 Trieste, Italy.

[3] Department of Integrative Marine Ecology, Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Napoli (SZN), Italy.

[4] Scientific Core Facilities & Technologies, GeneCore, European Molecular Biology Laboratory (EMBL), Meyerhofstrasse 1, 69117 Heidelberg, Germany.

[5] Marine Biological Laboratory (MBL), Woods Hole, Massachusetts, USA.

[6] Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa 9040495, Japan.

[7] Department of Molecular Evolution and Development, Wien University, Althanstraße 14 (UZA I), 1090 Wien, Austria.

[8] Central RNA Laboratory, Istituto Italiano di Tecnologia (IIT), Via Enrico Melen 83, 16152 Genova, Italy.

* These authors contributed equally to this work

corresponding authors:

Remo Sanges, e-mail: remo.sanges@gmail.com

Graziano Fiorito, e-mail: graziano.fiorito@szn.it

Abstract

Transposable elements (TEs) widely contributed to the evolution of genomes allowing genomic innovations, generating germinal and somatic heterogeneity and giving birth to long non-coding RNAs (lncRNAs). These features have been associated to the evolution, functioning and complexity of the nervous system at such a level that somatic retrotransposition of long interspersed element (LINE) L1 has been proposed to be associated to human cognition. Among invertebrates, octopuses are fascinating animals whose nervous system reaches a high level of complexity achieving sophisticated cognitive abilities. The sequencing of the genome of the *Octopus bimaculoides* revealed a striking expansion of TEs which were proposed to have contributed to the evolution of its complex nervous system. We recently found a similar expansion also in the genome of *Octopus vulgaris*. However a specific search for the existence of full-length transpositionally competent TEs has not been performed in this genus. Here we report the identification of LINE elements competent for retrotransposition in *Octopus vulgaris* and *Octopus bimaculoides* and show evidence suggesting that they might be active in the brain. Transcription and translation measured for one of these elements resulted in specific signals in neurons belonging to areas associated with behavioral plasticity. We also report the transcription of thousands of lncRNAs and the pervasive inclusion of TE fragments in coding and non-coding transcripts in both *Octopus* species, further testifying the crucial activity of TEs in the evolution of the octopus genomes. We speculate that a convergent evolutionary process driven by LINE elements led to the evolution of sophisticated cognitive abilities in this genus.

Introduction

Transposable elements (TEs) have contributed to the evolution of specific functions in a variety of biological systems and have given birth to a large fraction of vertebrate long non-coding RNAs (lncRNAs) (Kazazian 2004; Kelley and Rinn 2012; Kapusta et al. 2013). Among TEs, retrotransposons move via a copy-and-paste mechanism using an RNA intermediate. The long interspersed element (LINE) L1, a non-LTR retrotransposon that predated the human genome, is active during neuronal differentiation (Muotri et al. 2005) and causes somatic mosaicism establishing genomic variability in the brain (Baillie et al. 2011). Somatic L1 insertions are suggested to alter the transcriptional output of individual neurons, eventually affecting neuronal plasticity and behavior (Singer et al. 2010). Transposition-driven genomic heterogeneity has also been documented in invertebrates including the neurons of mushrooms bodies of *Drosophila melanogaster* (Perrat et al. 2013) where they have been suggested to drive behavioral variability in individual flies. Negative regulators of retrotransposons are reported to be expressed in specific subgroups of neurons in *Aplysia californica* (Rajasethupathy et al. 2012), *Drosophila melanogaster* (Perrat et al. 2013) and in the mouse brain (Lee et al. 2011), further supporting the idea that, in nervous systems, retrotransposons activity is finely regulated in a broad range of organisms and does not simply constitute noise. Our understanding of the activities of TEs in metazoa genomes is nevertheless still far to be complete and the number of cellular events in which they have an influence is constantly growing. Novel findings allow us to increase our comprehension but also add layers of complexity to the topic. L1 elements, for example, have been demonstrated to have specific activity also at the non-coding level in the regulation of transcription and the organization of the genome (Jachowicz et al. 2017; Percharde et al. 2018) and to be a component of extra-chromosomal DNA (Thomas et al. 2017).

Among invertebrates, the cephalopod mollusk *Octopus vulgaris* is known for the richness of its behavioral repertoire achieving sophisticated vertebrate-like plasticity and neural control (Shomrat et al. 2008; Edelman and Seth 2009; Young 1991; Turchetti-Maia et al. 2017). The remarkable innovations in the morphological and functional organization of its nervous system (Edelman and Seth 2009; Shigeno et al. 2018; Styfhals et al. 2019) are linked to the evolution of an unprecedented complexity coded both at cellular and molecular

levels (Shigeno et al. 2018; Shigeno and Ragsdale 2015; Gray and Young 1964; Wang and Ragsdale 2017; Styfhals et al. 2019). In the common octopus, about 500 million nerve cells constitute the nervous system, with about 300 million composing the nervous system in the arm and about 200 million nerve cells in what is consider to be the central brain (Young 1963). This number results to be ten thousand times higher than that found in the sea hare Aplysia, and still remains two hundred times higher when compared to the number of neurons present in the brain of the honeybee *Apis mellifera* (reviewed in Borrelli and Fiorito 2008). Octopus and their allies are also key examples of innovations in the evolution of invertebrates. Such innovations have been recognized not only at the level of the *Bauplan*, but include also extraordinary examples of the transcriptional and structural outputs of its genome such as large cadherin genes encoding over 70 extracellular cadherin domains, unprecedented expansions of gene families crucial for regulation, signaling and cell comunication (e.g., protocadherins, zinc finger proteins, G-protein coupled receptors), birth of many novel octopus-specific genes, the existence of a vascular endothelial growth factor (VEGF) pathway, reflectin genes originated by horizontal gene transfer, differential arrangements of key developmental genes and extensive RNA editing capabilities (Albertin et al. 2015; Guan et al. 2017; Wang and Ragsdale 2017; Ritschard Elena A. et al. 2019; Lee et al. 2003; Liscovitch-Brauer et al. 2017).

The sequencing of the *Octopus bimaculoides* genome (Albertin et al. 2015) revealed an expansion of TEs and specific gene families related to transcriptional regulation and neuronal connectivity. Although no specific analysis was performed to identify full-length potentially active TEs and non-coding transcripts were discarded from the definition of the reference transcriptome, it was suggested that TEs are active in *O. bimaculoides* because a substantial fraction of RNAseq reads resulted in overlap with TEs fragments annotated in non-coding intergenic regions. Expansion of TEs has also been found in the genome of *Octopus minor* (Kim et al. 2018) and the other sequenced cephalopod species such as *Euprymna scolopes* (Belcaid et al. 2019) and the giant squid *Architeuthis dux* (da Fonseca et al. 2020). Performing a survey of the *O. vulgaris* genome we have confirmed the expansion of TEs also in this species (Zarrella et al. 2019). The current picture is that, in cephalopod species, repeated elements cover on average 45% of the genome.

Here we sequence the *O. vulgaris* neural transcriptome to gain insights into the molecular composition underpinning its neural complexity. We identify a full-length LINE element and show that it is actively

106 expressed in specific areas of the brain related to known forms of behavioral plasticity including learning.
107 We also provide evidence for the transcription of thousands of long non-coding RNAs and the pervasive
108 inclusion of TE fragments in coding and non-coding transcripts in both *O. vulgaris* and *O. bimaculoides*.

Results

Thousands of putative lncRNAs are expressed in the Octopus vulgaris nervous system

We generated a *de-novo* assembly of the *Octopus vulgaris* neural transcriptome identifying and evaluating its functional annotations, potential lncRNAs and repeats composition (Fig. 1; Supplemental Tables S1, S2; Supplemental Fig. S1, S2; Supplemental Files S1, S2, S3). From each of three different octopus individuals we collected four parts, three of them as representative of the central brain: 1) supra-esophageal mass (SEM), 2) sub-esophageal mass (SUB) and 3) optic lobe (OL); and one representing the peripheral nervous system: 4) a piece of the second left arm (ARM) including the arm nerve cord. We used these parts as source of RNA for the sequencing (see Methods, Supplemental Table S1). The sequencing generated approximately 850 million paired-end reads accounting for 85 Gbp of sequence data that, following assembly and filtering, produced 64,477 unique transcripts (Supplemental File S1). The sequences showed a N50 value of 2,087 bp, average length of 1,308 and about 38% average GC-content (Supplemental Table S2). The transcriptome resulted to be more than 98% complete and we functionally annotated 21,030 (32.6%) protein coding transcripts (Supplemental Fig. S1 and Supplemental File S3). By performing stringent annotation analysis a high proportion of transcripts (7,806; 12.1%) resulted to be putative lncRNAs (see Methods and Supplemental Fig. S2). The non-coding portion of the *O. vulgaris* transcriptome shows a lower level of expression when compared to protein coding genes (Fig. 1A) and a significantly higher number of lncRNAs results to be expressed in central brain (~10%) with respect to the arm (~7%, p-value < 1e-40) in analogy to what is known about lncRNAs expression in mammals (Qureshi et al. 2010) (Fig. 1B). Functional enrichments highlight differences between transcripts expressed in the central brain and in the periphery. Specifically, transcripts expressed in the brain result enriched in functions associated with neuronal cell adhesion and reverse transcription, while the ones expressed in the arm are enriched for functions associated with signal transduction and translation (Fig. 1C and Supplemental Fig. S3).

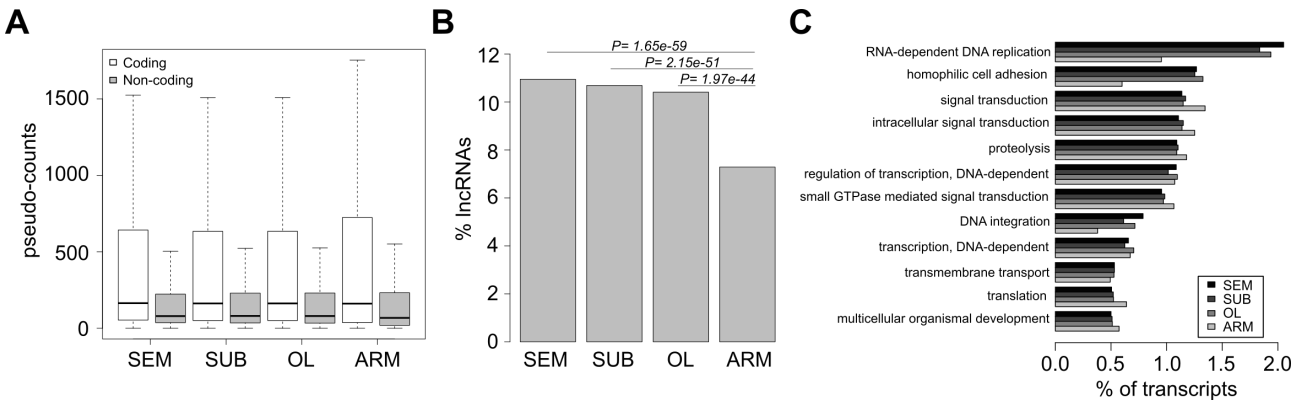

Figure 1. Features of the *Octopus vulgaris* brain and arm transcriptome. We sequenced the supra-esophageal (SEM) and sub-esophageal (SUB) masses and optic lobe (OL) as representatives of the brain, and the medial segment of an arm (ARM), including the arm nerve cord, as representative of peripheral system. (A) Expression levels for coding and non-coding transcripts. Non-coding transcripts are on average expressed at lower levels than coding. (B) Percentage of expressed non-coding transcripts. Brain samples result enriched for non-coding. (C) Percentages of transcripts expressed and their relative distribution among the most represented GO biological processes. A higher percentage of transcripts belonging to classes related to transposable elements and cell adhesion are expressed in the brain. Transcripts likely to be involved in signal transduction and translation constitute a larger quota in the arm.

We validated the expression and the sequence of selected coding and non-coding transcripts by RT-PCR and Sanger sequencing (Fig. 2) selecting a group of transcripts showing a specific peak of expression in each of the collected parts. A transcript was considered to show a peak of expression in a given part when showing an expression level higher than 0.5 counts per million (CPM) in all three biological replicates of exclusively one part and below 0.5 in all the others. This resulted in the selection of ~1,800 transcripts (~1,500 coding and ~300 non-coding). Among the coding transcripts with an expression peak only 54 resulted annotated. Among them we noticed the presence of putative homologs of homeobox genes and selected 4 of them for validation through RT-PCR in 3 different individuals. The tested *Arx* putative homolog (Aristaless Related Homeobox, comp31544_c0_seq1) resulted expressed mainly in the SEM by RNAseq and the RT-PCR validated this result. The RT-PCR also confirmed the peak of expression for the *Hoxb5a* putative homolog (Homeobox B5, comp28131_c1_seq2) which resulted expressed mainly in the SUB and the *Meox2* putative homolog (Mesenchyme Homeobox 2, comp34840_c15_seq1) which resulted expressed mainly in the ARM. The *Phox2b* putative homolog (Paired-Like Homeobox 2b, comp28142_c1_seq1) peaking in the OL by RNAseq data resulted expressed in all sampled parts of the brain by RT-PCR. For what concerns the lncRNAs RT-PCRs, *Subl* (lncRNA with a peak of expression in the SUB, comp35227_c11_seq1) and *Arml* (lncRNA with a peak of expression in ARM, comp20195_c0_seq1) appeared to be tissue-specific, as they

161 were mainly expressed in the SUB and in the ARM, respectively. On the other hand, *Seml* (lncRNAs with a
162 peak of expression in SEM, comp18661_c0_seq1) resulted to be expressed in the SEM but also in the SUB
163 and the OL, while *Oll* (lncRNA with a peak of expression in the OL, comp35506_c7_seq1) resulted
164 expressed in the OL but also in the other sampled parts (Figure 2C). Both *Oll* and *Arml* show the existence of
165 two different isoforms in at least one individual. The RT-PCR results were generally in agreement with the
166 sequencing data for both coding and non-coding transcripts tested. The identifiers indicated for every
167 transcript are the same used in the Supplementary files and can be used to identify the corresponding
168 sequences and annotations.
169
170 In order to verify the presence of lncRNAs also in *O. bimaculoides* we reassembled the public RNAseq data
171 from this species with the same method used for our *O. vulgaris* data (see Methods). We assembled 92,820
172 unique transcripts (Supplemental Table S2 and Supplemental File S4) of which 84,043 (90%) map on the
173 published assembled genome with at least 90% coverage and 90% identity (Supplemental File S5). Our
174 analysis demonstrated the presence, also in this octopus species, of thousands of putative lncRNAs (File S6).
175 Indeed we were able to classify 10,030 assembled transcript as putative non-coding. They correspond to
176 more than 10% of the assembled transcriptome of which 9,132 map on the reference genome.
177

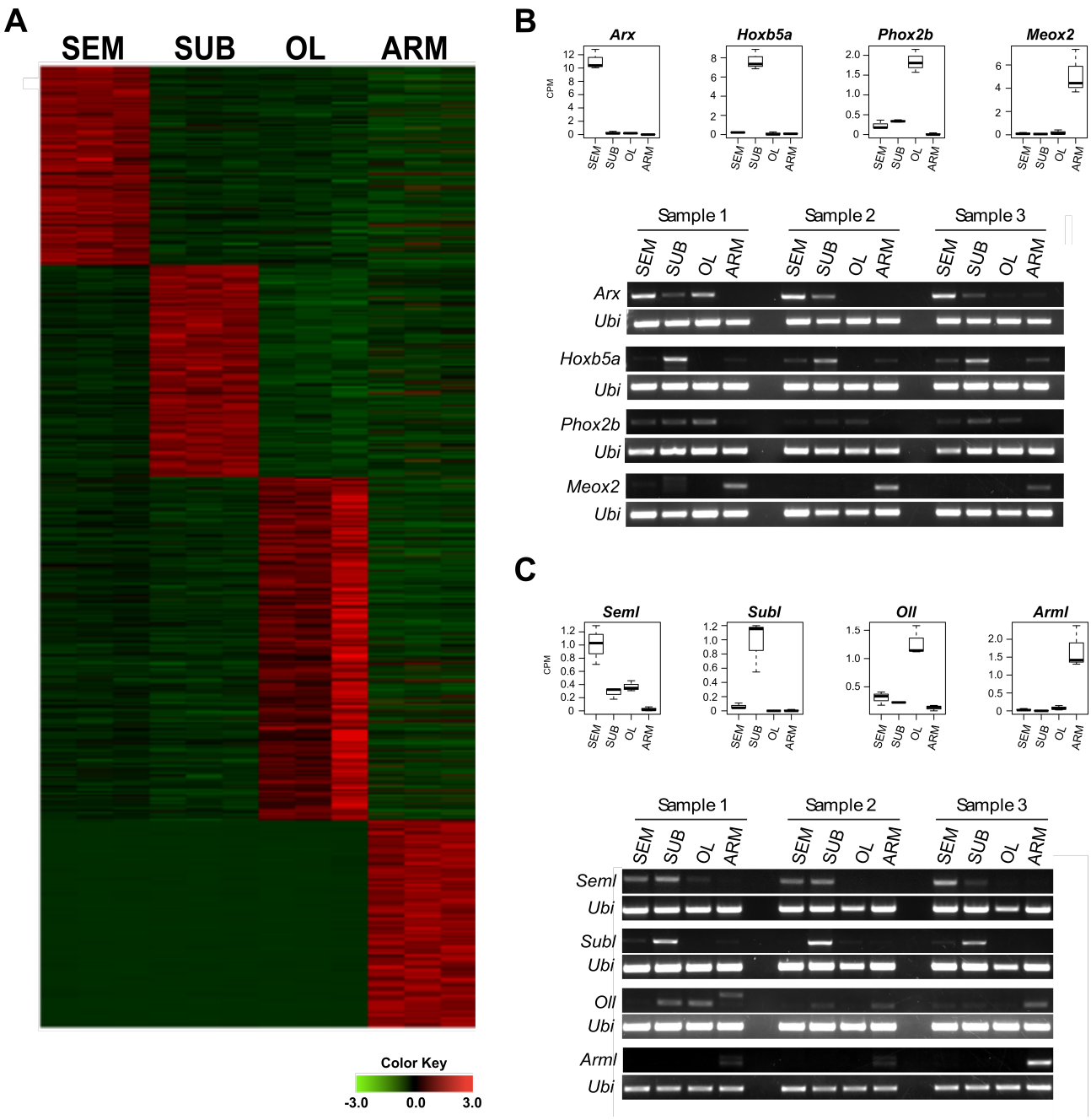

Figure 2. Transcripts with an expression peak and transcriptome validations. (A) Heatmap showing the expression levels for transcripts classified as having a peak of expression. (B) Boxplots showing RNAseq expression levels of coding transcripts selected for validation and their relative RT-PCR results from 3 different individuals. (C) Boxplots showing RNAseq expression levels of non-coding transcripts selected for validation and their relative RT-PCR results from 3 different individuals. The octopus ubiquitin transcript (Ubi) has been used as positive control in all the experiments.

A full-length LINE element is transcribed in the Octopus vulgaris brain

In order to evaluate the contribution of TEs to the *O. vulgaris* transcriptome we analyzed its content in terms of repeated elements using RepeatMasker (Supplemental File S7) and found that more than 3.5 millions of nucleotides derive from interspersed repeats (4.5% of the total transcriptome content; Supplemental Table

190 S3). More than 35% of the generated transcripts contain at least one interspersed repeat fragment. Among
191 them, retroelements represent the TE fragments more frequently embedded in transcripts (26% of transcripts
192 contain a fragment from at least one retroelement). According to the segregation of transposable elements in
193 transcripts expressed in the different sequenced parts of the organism, we observed that SINEs fragments are
194 present in a higher fraction of transcripts expressed in the brain, while LINEs, LTRs and DNA transposons
195 are present in a higher portion of transcripts expressed in the periphery (Supplemental Fig. S4A). SINE also
196 results the class of retroelements more frequently embedded in lncRNAs (Supplemental Fig. S4B).
197
198 To identify the molecular basis of the observed TEs expansion in *O. vulgaris* we searched the transcriptome
199 for putative autonomous active elements. We found a single element, a LINE mainly transcribed in the brain
200 (Fig. 3A-C) that we named RTE-2_OV following phylogenetic analysis from which the element resulted to
201 belong to the RTE clade (Fig. 3D). The identified LINE presents an ORF of 3,327 nucleotides and 5' and 3'
202 UTRs of about 600 nucleotides. The translation corresponds to a 1,109 aminoacids polypeptide chain
203 containing all the catalytic aminoacids and domains needed for retrotransposition (Fig. 3A,B): a C-terminal
204 endonuclease (EN), a reverse transcriptase (RT) and an N-terminal C2H2 zinc-finger (Znf) which is
205 relatively rare in RTE elements (Malik and Eickbush 1998). The finding that this element contains an intact
206 ORF and is transcribed indicates that it might be active and possibly drive retrotransposition. We then asked
207 whether this element might show evidence of somatic retrotransposition. Southern blots analysis did not lead
208 to conclusive results because of high background noise likely due to the high number of copies of this TE in
209 the genome of *O. vulgaris*, however the pattern observed is consistent with the existence of germinal and
210 somatic genomic variations associated to the element (Supplemental Fig. S5). We also performed
211 quantitative PCR experiments on DNA deriving from the four parts of different individuals. Results add
212 support to the presence of somatic retrotransposition in the brain (Fig. 3E) showing an average higher
213 content of RTE-2_OV DNA in the SUB with respect to the other parts. This result suggests the existence of a
214 higher amount of RTE-2_OV DNA in cells from the SUB, whereas cells from other parts show a similar
215 average content. The existence of these increased amounts of RTE-2_OV DNA in SUB is in line with its
216 higher expression in the SUB (Fig. 3F). Interestingly, the putative homolog of Piwi, known to repress the
217 translation of TEs and therefore to restrict retrotransposition, displays a lower expression in the SUB with

respect to the other portions of the central brain of (Fig. 3G) consistently with what has been observed in the fruitfly brain (Perrat et al. 2013).

In order to add further support to the existence of active LINEs in the genomes of the Octopus *genus* we searched the public genome, transcriptome and our custom transcriptome assembly (Supplementary Files S4-S6) of *O. bimaculoides* for the presence of assembled full-length LINEs. We were not able to identify any assembled full-length LINE in the transcriptome nor in the genome. However, when inspecting the repeat library consensus generated in the work by Albertin *et al.* (Albertin et al. 2015) we identified two LINEs with a full-length ORF and the complete set of domains: an RTE and a Dong (Xiong and Eickbush 1993) element. To gain insights into the possibility that the identified LINE elements might be active we generated DNAseq sequencing data from the optical lobe of an *O. bimaculoides* individual and searched for novel integration sites (IS) with respect to the reference genome using MELT (Gardner et al. 2017). The same analysis was performed using the public DNAseq data derived by the individual used to assemble the reference *O. bimaculoides* genome. Comparing the whole genome sequencing data from the two different *O. bimaculoides* individuals we obtained significant evidences of activity only for the RTE element. Indeed, the element showed a significantly higher number of non-reference ISs with respect to the reference individual (Fig. 3H, p-value 4.8e-69) which indicates that this element is at the basis of polymorphisms between the two individuals. Conversely, the Dong element resulted in a very low number of non-reference ISs which does not significantly differ from the results obtained with reads from the reference individual and therefore does not display evidences of activity.

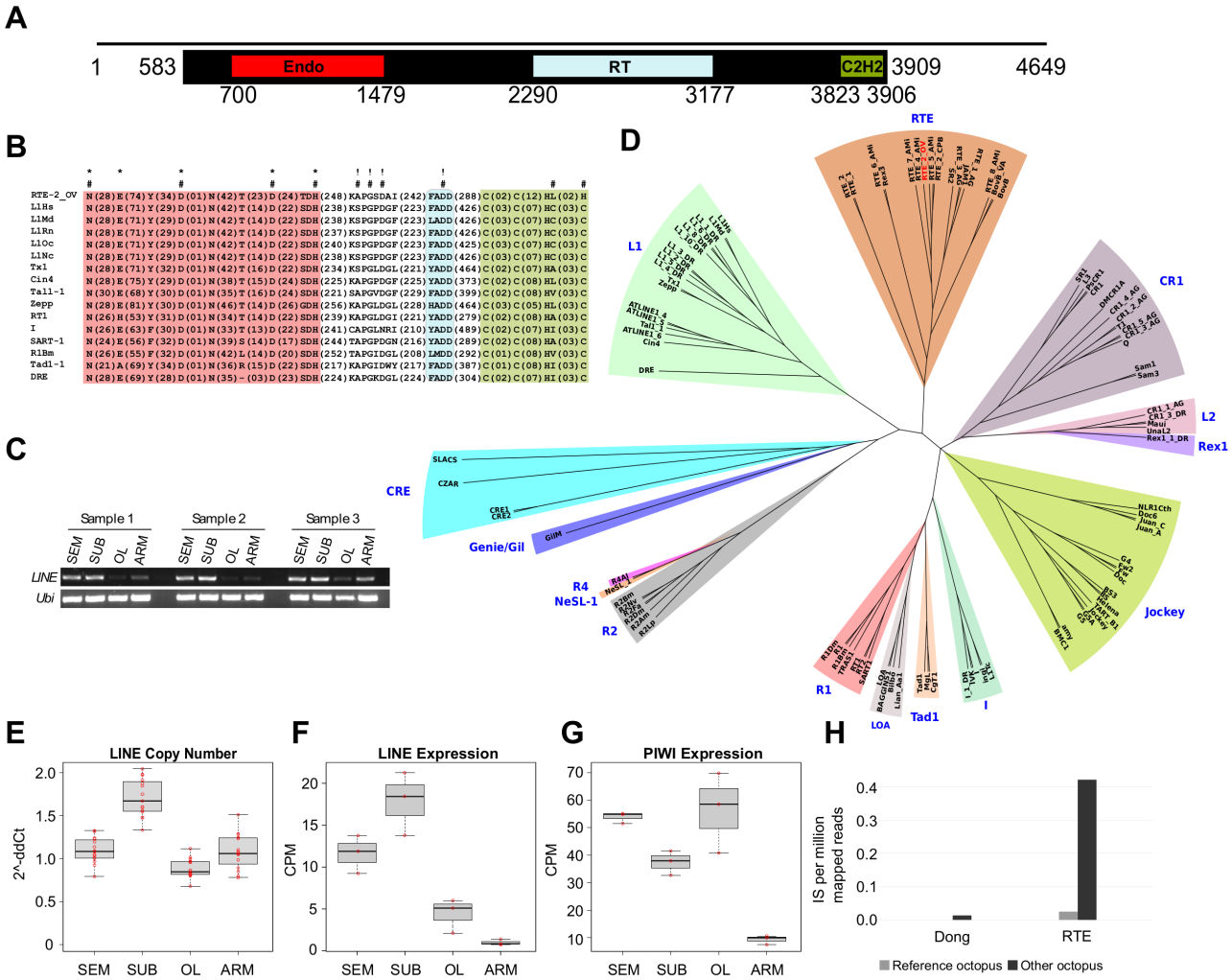

Figure 3. A full-length potentially active LINE transcribed in *Octopus vulgaris*. (A) Domain composition of the discovered LINE (comp36575_c1_seq3). Black line represents the transcript, black box the location of the ORF and colored boxes the protein domains (Endo, endonuclease; RT, reverse transcriptase; C2H2, zinc finger). Numbers are relative to nucleotide positions in the transcript. (B) Schematic alignment highlighting the conserved catalytic aminoacids in the group of LINEs adapted from (Clements and Singer 1998) plus the octopus element. Color code of the domains is the same as in A; aminoacids critical for EN (*), RT (!) and retrotrasposition (#). (C) Electrophoresis from RT-PCR of the LINE showing expression from three different animals. (D) Phylogenetic tree based on 100 LINEs from (Ohshima and Okada 2005) (see Supplementary Table 4) plus the octopus element in red. (E) LINE copy number variation analysis using quantitative real-time PCR with Taqman probes. (F) Expression levels of the LINE based on RNAseq data. (G) Expression levels of Piwi-like protein 1 (comp33731_c0_seq1) from RNAseq data. (H) Normalized number of non-reference insertions in two different *O. bimaculoides* individuals (octopus used for the reference genome and a different octopus individual) for the two LINEs identified in this species.

254 ***The RTE-2_OV element is mainly expressed in amacrine neurons***

255 To identify the domains of activity for the RTE-2_OV element we performed RNA *in-situ* hybridization and
256 immunohistochemistry analysis. Localization of the RTE-2_OV transcript in *O. vulgaris* through *in-situ*
257 hybridization (ISH) showed specific expression of the element in subgroups of neurons in brain and absence
258 of any signal in neuropils. We found most of the small cells of the sub-frontal lobe and of the five gyri of the
259 vertical lobe as the most intensely stained areas in the supra-esophageal mass (SEM; Fig 4A-E;
260 Supplemental Fig. S6B-E). Neural cells stained for RTE-2_OV create a tapestry closely matching the pattern
261 obtained by DAPI (Supplemental Fig. S6B,C; Supplemental Fig. S6D,E as control), indicating that the
262 transcript is expressed in the great majority of the cellular bodies. We also found several positive neural cells
263 in the sub-esophageal mass (SUB: anterior and posterior areas; Fig. 4F-H; Supplemental Fig. S6G) and in the
264 optic-lobe (OL; Fig. 4I-K; Supplemental Fig. S6L). In particular after ISH, positive cells (20-25 µm in
265 diameter) were observed at the level of the SUB in the pallovisceral lobe, and some larger neurons (40-50
266 µm) belonging to typical cellular-types of the motor-centre present in the area. Positivity to the mRNA of the
267 RTE-2_OV was also observed in cells belonging to the ventral side of the anterior pedal lobe (SUB) and in
268 some larger cells (up to 50 µm) at the level of the dorsal brachial lobe (anterior part of the SUB). The small
269 cells pertaining to these areas do not show any positive signal. In the optic lobe, the outer layer appeared rich
270 of intensely positive cells (small amacrine cells, < 5 µm) and the inner medulla presented scattered cell
271 bodies (up to 10 µm) expressing the element. Cell bodies of the peduncle complex at the level of the median
272 and posterior lobules of the olfactory lobe also revealed a positive signal after ISH. We finally found isolated
273 sparse large motor neurons positive at the level of the nerve cord in the arm (Supplemental Fig. S6S). No
274 positive cells were observed in any muscle fiber in the arm or any other structure.

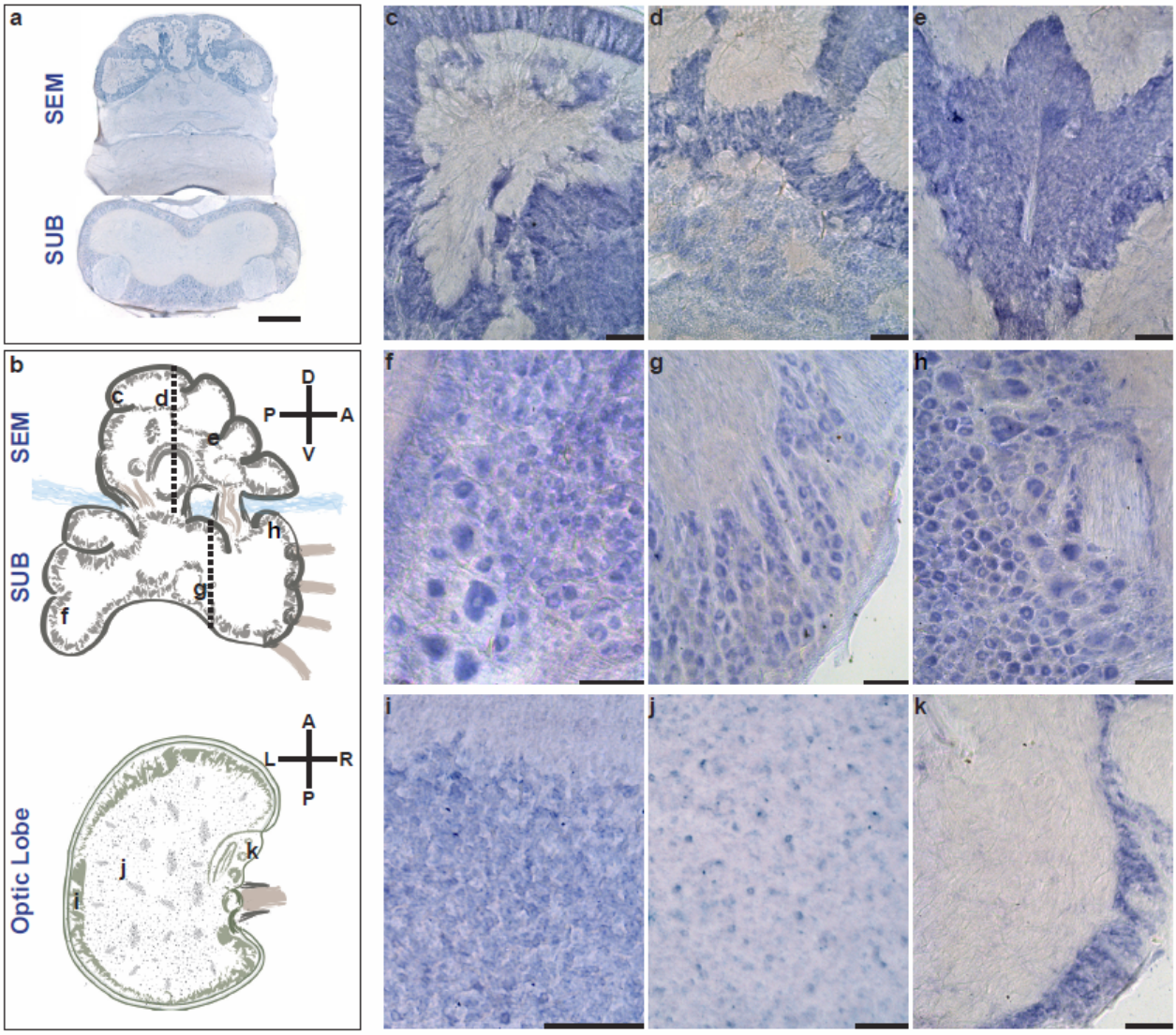

Figure 4. Localization of RTE-2_OV mRNA by in situ hybridization. (A) Bright-field micrographs of coronal sections of the supraesophageal (SEM) and subesophageal (SUB) masses hybridized to digoxigenin-labelled LINE antisense (refer to the plane indicated by the dashed-lines in B). The five gyri of the vertical lobe (dorsal in SEM) appear positively marked by the RTE-2_OV mRNA. Several cells in the cortical layers of the SUB appear also stained after in situ hybridization. Neuropil of SEM and SUB do not show signal. (B) Schematic outline of the parts of the octopus brain for SEM and SUB (sagittal plane), and the optic lobe (OL; horizontal plane). Axes illustrating dorso-ventral and antero-posterior (SEM and SUB), and antero-posterior and left-right (OL) orientations with respect to the octopus body plan. Black letters indicate approximate levels of the sections provided in the other panels of the figure. (C) Detail of a gyrus of the vertical lobe (SEM) with densely packed amacrine cells showing positive signal. (D) A similar signal in the gyri of the vertical lobe and some scattered positive cells in the sub-vertical lobe. (E) Section at the level of the sub-frontal lobe with densely packed amacrine small cells showing strong positive signal. In the SUB we observed in (F) positive cells (20-25 µm in diameter) in the pallovisceral lobe and some larger neurons (40-50 µm) belonging to typical motor-centre cellular types. (G) Cells (20-25 µm) belonging to the ventral side of the anterior pedal lobe, and at the level of the dorsal brachial lobe (H) where some larger cells (up to 50 µm) are also marked after ISH. The small cells pertaining to these areas do not show positivity. Details of horizontal sections of the *O. vulgaris* optic lobe (I, J, K; in areas indicated in B): (I) Outer layer rich of intensely positive cells (small amacrine cells, <5 µm); (J) Inner medulla with scattered LINE mRNA-expressing cell bodies (up to 10 µm); (K) cell bodies of the peduncle complex at the level of the median and posterior lobules of the olfactory lobe (cells of about 10 µm). Scale bars: 100 µm; 500 µm in a. Schematic drawings in (B) not to scale.

297 Immunohistochemistry (IHC) with RTE-2_OV custom antibodies identified in the vertical lobe a number of
298 large cells organized in trunks and positive fibers in the neuropil of the gyri. We also observed large positive
299 cells organized in chain at the level of the cellular wall of the sub-vertical lobe (Fig 5B) and few positive
300 neurons in the great majority of the islands of cells in the posterior wall of the sub-vertical lobe. A distinct
301 pattern of positive fibers was found in neuropils of the SEM (Fig. 5A-C) together with a scattered number of
302 cell bodies (see areas belonging to superior- and inferior frontal lobes; Fig. 5C-D). We did not identify
303 positive fibers in the SUB, but only a distinct population of a small number of neurons in the vasomotor
304 lobes (Fig 5E) and in discrete areas of the anterior and the lateral posterior wall of the pedal lobe (Fig. 5F-H).
305 Immuno-reactivity was also evident in several scattered amacrine cells of the external granular layer of the
306 OL (Fig 5I). We found several RTE-2_OV positive fibers dispersed in discrete areas of the neuropil of the
307 OL and a distinct pattern of positive cells and fibers in the peduncle lobe, mostly towards the internal layer of
308 cells of the neural wall of the olfactory lobe and the spine (Fig. 5L). A small number of large cells dispersed
309 in the arm nerve cord were also found to be stained (data not shown).

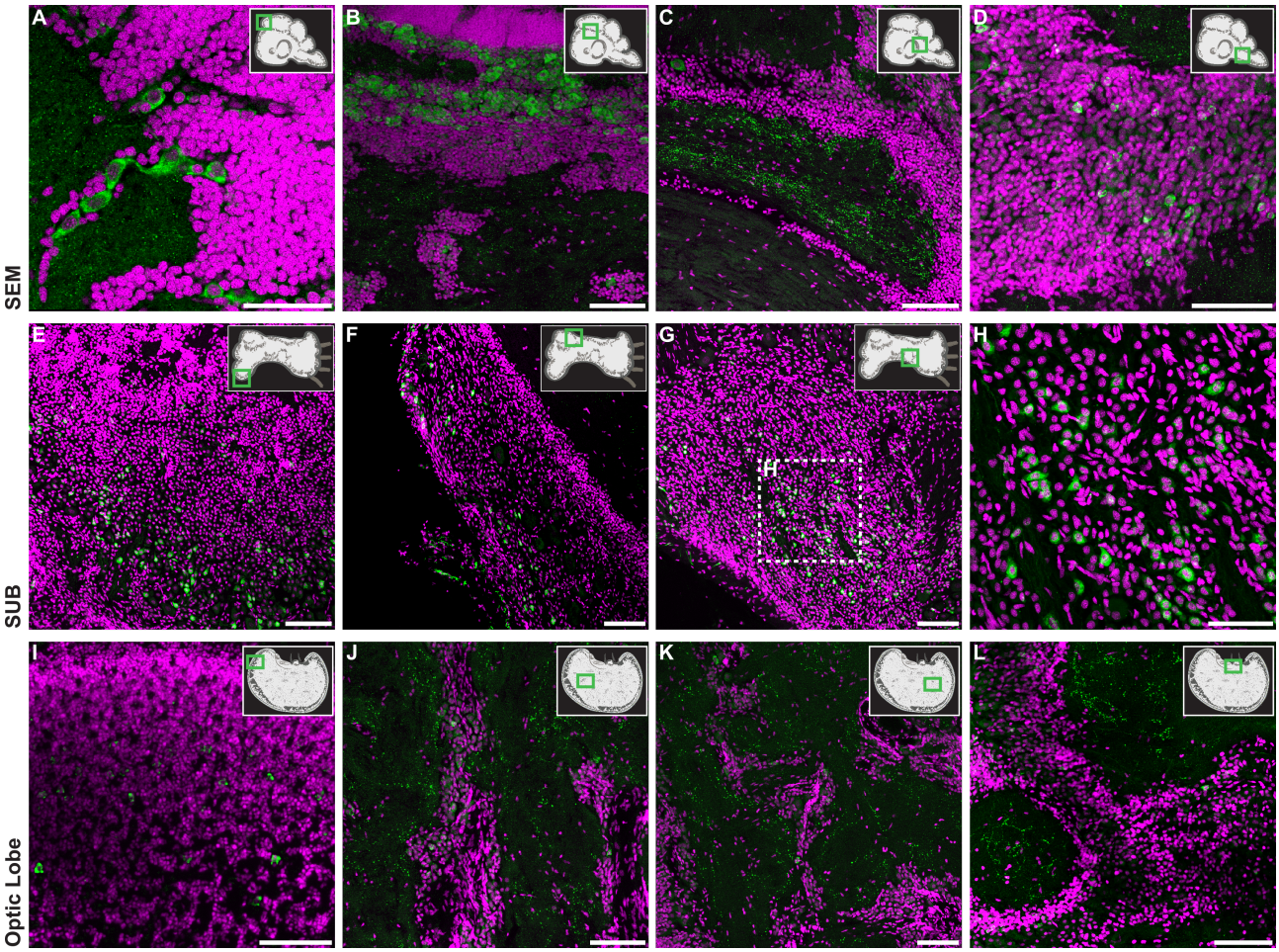

Figure 5. RTE-2_OV immunostaining in different areas of the brain. Coronal sections of the supra- (SEM; A-D) and sub-oesophageal (SUB; E-H) masses, and horizontal sections for the optic lobe (OL; I-L) following fluorescent-IHC (RTE-2_OV signal in green, DAPI used as nuclear stain in magenta) highlight a differential pattern of positive cells and fibers in *O. vulgaris* brain. A schematic drawing of the brain parts is provided with areas of interest indicated in the green square. (A) Large positive cells are found in the vertical lobe (VL). These appear organized in trunks and clearly distinguishable from the population of numerous amacrine cells constituting the VL (DAPI stained layer). (B) Large cells in the sub-vertical lobe (cellular wall), and a part of the bundle of fibers present at the beginning of the sub-frontal lobe (C). (D) Scattered positive cells are also identified in the posterior buccal lobe. Several positive cells are identified in the SUB in the cellular walls of the vasomotor lobe (E) and in discrete areas of the pedal lobe (F). A similar pattern of positive cells is recognized at the level of the anterior part of the pedal lobe (G). A detail in the higher magnification (H; square in G) of the cellular layer of the lobe serves to highlight the population of positive cells. In the OL several amacrine cells are found positive in the external granular layer (I). The OL-medulla is populated by few immune-reactive neurons found in the cellular islands (J, K), and positive fibers dispersed in the surrounding neuropil (J, K). Positive cells and fibres are also identified in the peduncle lobe. The internal layer of cells of the neural wall of the olfactory lobe and the spine (L) are shown. Scale bars: 100µm, with the exception of A and H (50µm).

328 ## Discussion

329 The analysis of the coding part of the *O. bimaculoides* genome provided interesting parallelisms with the
330 transcriptional output of mammalian genomes (Albertin et al. 2015). Our results in *O. vulgaris* support and
331 expand this view suggesting that the molecular organization of the octopus neural transcriptome resembles
332 the mammalian one also for what concerns the transcription of thousands of lncRNAs and their TEs content.
333
334 Here we report the assembly of the most complete transcriptome for the *Octopus vulgaris*. The sequencing of
335 different portion of the organism allowed us to identify transcripts whose expression showed a peak in only
336 one part. For qPCR validation we chose putative homologs of homeobox genes. *Arx* is an homeobox gene
337 associated with mental retardation. In mammalian it has been shown to be expressed in fetal brain at various
338 developmental stages, mainly in neuronal precursors in the telencephalon were it results involved in
339 corticogenesis and in the differentiation and maintenance of specific neuronal cell types (Ohira et al. 2002;
340 Friocourt et al. 2006). In our *O. vulgaris* transcriptome assembly its putative homolog results to be expressed
341 mainly in the SEM consistently with an important function in the central brain also in this species. As
342 another example, the expression of the putative homolog of *Meox2* appears to be restricted to the arm which
343 also appears consistent with the known function of this gene in vertebrates were it controls limb muscle
344 development (Mankoo et al. 1999). The selection and validation of these expression peaking transcripts
345 together with the functional information of the putative homologs added support to the quality of the
346 assembled transcriptome and allowed us to perform more specific analyses. We observed that more than 10%
347 of the assembled transcriptome for *O. vulgaris* and *O. bimaculoides* is composed of putative lncRNAs and
348 that more than 35% of the assembled transcripts contain at least one fragment derived from a transposable
349 element. Despite the fact that different classes of TEs appear to be differentially segregated between coding
350 and non-coding transcripts we did not find an overall difference nor a strong enrichment for lncRNAs to
351 contain embedded TE fragments with respect to coding ones.
352
353 The observation reported from us and others about TEs expansion in octopus drove our interest toward the
354 proportion of the transcriptome deriving from transposable elements. This lead us to the discovery of a novel
355 RTE LINE retrotransposon transcribed in the central nervous system of *O. vulgaris*. Re-analysis of the *O.*

356 *bimaculoides* data allowed us to identify, also in this species, a LINE member of the RTE family showing
357 evidences of activity. We could not find any full length copy of the *O. vulgaris* element in the genome survey
358 (Zarrella et al. 2019) nor could we find the *O. bimaculoides* element full length sequence in its reference
359 assembly. This likely resulted from the difficulties to assemble a reference genome using exclusively short
360 reads in regions containing long repeated sequences as those containing full-length LINE elements. Despite
361 this technical gap, the complementary use of transcriptomic, repeats reconstruction and annotation data
362 allowed us to identify the full length autonomous elements likely to be active in both the octopus species.
363 The identification of potential integration polymorphisms of the RTE element in *O. bimaculoides* is a strong
364 sign of its activity. On the other hand, the results from quantitative PCR and Southern blot analysis relative
365 to the element discovered in *O. vulgaris*, although not conclusive, support the existence of somatic
366 retrotransposition in the central brain for this species. It is here important to consider that results from
367 quantitative PCR for LINEs genomic copy number variations should be interpreted with care, especially in
368 case of absence of additional supporting evidences. Indeed, in addition to a potential variation in genomic
369 copy number, they might also reflect a different amount of cytoplasmic DNA which, at least in human, is
370 composed also by a fraction of LINEs DNA (Thomas et al. 2017; De Cecco et al. 2019). More specific and
371 careful inspections are needed to validate, beyond any reasonable doubt, the activity of the elements
372 presented in this work. Nevertheless our results represent the first evidences of LINE activity conserved in
373 the Octopus genus.
374
375 Through ISH we localized RTE-2_OV transcripts in specific groups of neural cells and areas of the *O.*
376 *vulgaris* nervous system, with no signal in neuropils. In particular, positive signals were observed at the level
377 of the majority of amacrine cells (about 3 micron diameter) which are suggested to have very limited neural
378 processes (Gray 1970). The most intriguing signals are those present at the level of the sub-frontal lobe and
379 of the five gyri of the vertical lobe (VL). These areas appeared to be the most intensely positive to RTE-
380 2_OV transcript found in the supra-esophageal mass. VL represents about 14% of the overall volume of the
381 SEM in the adult *O. vulgaris* (Nixon and Young 2003) and counts about 25 million neural cells, equivalent to
382 more than 65% of the total number of nerve cells estimated for the SEM in the common octopus (Young
383 1963). The largest quota of this impressive number of closely packed nerve cells is represented by the
384 amacrine cells, the smallest in the octopus brain. Lying below the vertical lobe is the sub-vertical lobe. Its

385 neural architecture is characterized by the presence of a wall that in several regions is folded to form islands
386 of cells. Many cells have a diameter less than 5 μm, and very few are larger than 10 μm. At both sides of the
387 sub-vertical lobe there are cells with diameter of 5–10 μm, but also the largest neurons of the whole SEM
388 reaching 25 μm. Through IHC in these areas we identified a number of large cells positive to RTE-2_OV
389 antibodies appearing organized in trunks as typically described by Young in 1971 (Young 1971). We also
390 observed large positive cells organized in chains at the level of the cellular wall of the sub-vertical lobe, and
391 few positive neurons in the great majority of the islands of cells (posterior wall of the sub-vertical lobe).
392 Finally, immuno-reactivity was also evident in several scattered amacrine cells belonging to the external
393 granular layer of the OL suggested to correspond to the neural cells described to project in the underlying
394 plexiform zone (Young 1991).
395
396 Remarkably, *O. vulgaris* posterior frontal and vertical lobes belong to a neural circuit considered to be
397 functionally analogous to the mammalian hippocampus and limbic lobe (Shigeno et al. 2018; Young 1991;
398 Abbott et al. 1995; Young 1995) where retrotransposons are also reported to be active (Perrat et al. 2013;
399 Upton et al. 2015). The VL is also suggested to be analogous to the mushroom bodies of insects (Hochner
400 2010). All other octopus brain areas where the mRNA and/or the protein were identified are centers of active
401 neural plasticity for visual sensory-motor processing and memory (SEM and OL), and motor processing
402 (SUB) (Young 1971; Marini et al. 2017; Shigeno et al. 2018). The positive cells overlap with the intricate
403 neural network made by amacrine cells known to act as amplifying matrices functioning as read-in read-out
404 memory of both sensory-motor visual and chemo-tactile processing (Maldonado 1963, 1965; Young 1991,
405 1995).
406
407 The reported imaging experiments allow also to made observations at the cellular level. The signals resulting
408 from IHC and ISH are principally cytoplasmic. An higher number of cells appear to contain the element
409 mRNAs with respect to the cells in which the protein can be identified. We only observed a partial overlap
410 between cells expressing the mRNA and cells containing the protein. These observations might reflect the
411 complexity of LINEs involvement in cellular activities. Such activities might be resulting by the strategies
412 employed in the cell to contrast retrotransposition but they can also reflect the basis of specific LINE
413 functions that we currently understand only with limitations. The most of our knowledge on LINE elements

414 comes from studies on human L1. Its transcript is a bicistronic mRNA which codifies for two open reading
415 frames (ORF1 and ORF2). Among them, the ORF2 protein contains the endonuclease, reverse transcriptase
416 and zinc fingers domains and therefore it can be considered the ORF codifying for the retrotranspositional
417 machinery and the functional homolog of the single ORF of the RTE elements that we found in the two
418 octopus species. Of note, the two human L1 ORFs do not result to be translated at the same levels. ORF1 is
419 translated at a higher level than ORF2, and different proteins which repress ORF2 translation have recently
420 been identified (Mita et al. 2020). Consequently, the ORF2 protein has been observed to be absent in a
421 portion of cells where the mRNA and ORF1 protein were present (Mita et al. 2018). The limited efficiency
422 for the translation of ORF2 has been proposed to be one of the mechanisms that the cell puts in action to
423 limit and/or regulate the retrotranspositional potential of the L1 mRNAs. It is important to stress here that,
424 for the retrotransposition to take place, all the L1 components, ORF1 protein, ORF2 protein and L1 mRNA
425 must be present in the same ribonucleoprotein complex (RBP) within the nucleus. The lack of any of these
426 components in the RBP does not allow retrotransposition. Reasoning along this line, another mechanism
427 used to restrict retrotransposition is the confinement of the RBP complex within the cytoplasm, where it is
428 assembled. As long as the complex is located in the cytoplasm it cannot access the genome and therefore
429 cannot retrotranspose. For this reason, cells subjected to retrotransposition have been identified mainly in
430 cells which underwent duplication. The mechanism proposed is that the RBP can get in contact with nuclear
431 DNA only after the nuclear membrane dissolves during mitosis. If the RBP remains in contact with the DNA
432 it is then found into the nucleus after the nuclear membrane of the new cell has been rebuilt thus allowing
433 retrotransposition (Mita et al. 2018; Taylor et al. 2018).
434
435 Spatial uncoupling between L1 mRNA and proteins can also be observed independently from cellular
436 defense mechanisms. Indeed, L1 elements also are also active at the mRNA level as a lncRNA. During early
437 development and in the regulation of cell potency, L1 transcription and its mRNA result to play a crucial role
438 in regulating global chromatin accessibility and the L1 mRNA can act as a nuclear RNA scaffold to recruit
439 specific regulatory proteins to the genome (Jachowicz et al. 2017; Percharde et al. 2018). Another molecular
440 event resulting in spatial uncoupling of LINEs transcription/translation can be observed when LINE mRNA
441 is used from a given cell type to target a different cell type with retrotranspositional activity. In fruitfly the
442 existence of a mechanism producing mRNAs of the I element (another member of the LINE family) by the

443 nurse cells has recently been observed. Once produced, these mRNAs are transported to the transcriptionally
444 inactive oocytes were they are translated and can cause retrotransposition (Wang et al. 2018).
445
446 Our study suffers from the limitations due to the difficulties of working with complex multicellular
447 organisms which are not established models for molecular studies and do not generally breed in captivity.
448 Nevertheless, our data corroborate the hypothesis that LINE elements, potentially capable to determine
449 genomic and functional diversity among individual cells, might be active and functionally important in the
450 central nervous system of highly intelligent organisms such as octopus (Singer et al. 2010). The localization
451 of RTE-2_OV mRNA into different neural cells and not in the neuropil or muscle, complemented by the
452 distinctive pattern of expression of its protein product and the identification of a potentially active element of
453 the same family from a congeneric species, support the view that RTE-2_OV might be functional in neural
454 centers in *O. vulgaris* and make it an important candidate to be further studied for its contribution to neural
455 plasticity in this fascinating organism.

456 # Methods

457

458 ***Animals, sampling and ethical statement***

459 Octopus vulgaris were collected by local fishermen from the Bay of Naples (Southern Tyrrhenian Sea, Italy)
460 in early summer 2012. Animals were transported to the Stazione Zoologica Anton Dohrn in Napoli and
461 maintained according to a standardized acclimatization protocol (Fiorito et al. 1990; Amodio et al. 2014;
462 Fiorito et al. 2015). Samples were taken from local fishermen, by applying humane killing following
463 principles detailed in Annex IV of Directive 2010/63/EU as described in the Guidelines on the Care and
464 Welfare of Cephalopods (Fiorito et al. 2015) and following protocol for collection of tissues described by
465 Baldascino and coworkers (Baldascino et al. 2017). Death was confirmed by transection of dorsal aorta. All
466 dissections were carried out on seawater ice bed. During surgery optic lobes (OL), supra- (SEM) and sub-
467 esophageal (SUB) masses were dissected out from the animal, a piece of an arm (ARM), usually the second
468 left arm, was also taken. The complete dissection lasted less than 10 min. Sampling from live animals
469 occurred before the entry into force of the Directive 2010/63/EU in Member States and therefore no
470 legislation was in place in Italy regulating research involving cephalopods. However, the care and welfare of
471 animals has been consistent with best practice (Fiorito et al. 2015; Andrews et al. 2013; Fiorito et al. 2014)
472 and in compliance with the requirements of the Directive 2010/63/EU that includes cephalopods within the
473 list of species regulated for scientific research involving living animals. In addition, animals killed solely for
474 tissue removal do not require authorization from the National Competent Authority under Directive 2010/63/
475 EU and its transposition into national legislation.

476

477 ***RNA extraction and sequencing***

478 For each octopus (N=3) total RNA was isolated from central nervous tissues (SEM, SUB and OL) and ARM,
479 a part of the body including the largest quota of the neuronal population belonging to the peripheral nervous
480 system (Young 1963; Hochner 2013) and thus constituted by muscle and peripheral nervous tissue. SV total
481 RNA isolation kit (Promega, #Z3100) was utilized according to the manufacturer's protocol. DNA was
482 degraded by treating samples with Turbo DNase Kit (Ambion) according to the manual. Quality and quantity
483 of RNA was assessed by NanoDrop (Thermo-Fisher) and RNA BioAnalyzer (Agilent Technologies, Santa

484 Clara, CA, USA). Paired-end libraries were prepared using the Illumina TruSeq RNA sample library
485 preparation kit (Illumina, San Diego, CA, USA). Each sample was barcoded, samples pooled and sequenced
486 in two lanes on the Illumina HiSeq 2000 platform (paired-end, non-strand specific, 2x50bp read length
487 protocol).
488
489 ***Raw reads quality filtering***
490 Quality of raw reads was assessed using FastQC (release 0.10.1). Raw reads were filtered and trimmed based
491 on quality and adapter inclusion using Trimmomatic (Bolger et al. 2014) (release 0.22; parameters: -threads
492 24, -phred 64, ILLUMINACLIP:illumina_adapters.fa:2:40:15, LEADING:3, TRAILING:3,
493 SLIDINGWINDOW:3:20, MINLEN:25). Read pairs with both reads passing the filters were considered for
494 the transcriptome assembly. Trimmed and filtered reads were normalized to remove duplicates using the
495 normalize_by_kmer_coverage.pl script from Trinity (Haas et al. 2013) (release r2013_08_14; parameters: --
496 seqType fq, --JM 240G, --max_cov 30, --JELLY_CPU 24).
497
498 ***De novo assembly and quantification of transcript abundances***
499 Transcriptome was assembled using Trinity (release r2013_08_14) on the trimmed, filtered and normalized
500 reads exploiting the Jaccard clip to limit assembly of chimeras. Assembly was performed using the following
501 parameters: --seqType fq, --JM 240G, --inchworm_cpu 24, --bflyHeapSpaceInit 24G, --bflyHeapSpaceMax
502 240G, --bflyCalculateCPU, --CPU 24, --jaccard_clip, --min_kmer_cov 2. To measure expression levels, raw
503 reads were mapped on the assembled transcriptome using Bowtie (version 1; parameters: -p 24, --chunkmbs
504 10240, --maxins 500, --trim5 2, --trim3 2, --seedlen 15, --tryhard -a -S). SAM outputs from Bowtie
505 (Langmead et al. 2009) were converted into BAM, sorted, indexed and counted using the view, sort, index
506 and idxstats programs from Samtools (Li et al. 2009). All transcripts not showing at least 0.5 reads mapping
507 per million mapped reads (CPM) in at least 2 samples were discarded from the transcriptome as being
508 expressed at too low levels and therefore likely deriving by noise or assembly artifacts.
509
510 ***Annotation of the assembled transcriptome***
511 CEGMA (Core Eukaryotic Genes Mapping Approach; release 2.5) (Parra et al. 2007) was used to measure
512 the completeness of the assembled transcriptomes using the set of 248 Core Eukaryotic Genes (CEGs).

513 Transcripts annotation was performed using the Annocript pipeline (Musacchia et al. 2015) (release 0.2) with
514 the combination of tool, parameters and databases described below and using BLAST+ (release 2.2.27)
515 (Camacho et al. 2009). To annotate proteins we used BLASTX against the UniRef90 and Swiss-Prot
516 databases from UniProt (release 2013_09) (Consortium 2015) with parameters: -word_size 4, -evalue 10e-5,
517 -num_descriptions 5, -num_alignments 5, -threshold 18. To annotate protein domains we used RPSBLAST
518 against the Conserved Domains Database (CDD v3.10) (Marchler-Bauer et al. 2013) with parameters: -
519 evalue 10e-5, -num_descriptions 20, -num_alignments 20). Ribosomal and small non-coding RNAs were
520 identified using BLASTN against a custom database made by Rfam (realease 11.0) (Burge et al. 2013) and
521 ribosomal RNA sequences from GenBank (parameters: -evalue = 10e-5, -num_descriptions 1, -
522 num_alignments 1). Each transcript was associated to Gene Ontology (GO) (Ashburner et al. 2000), Enzyme
523 Commission (EC) (Bairoch 2000) and UniPathway (Morgat et al. 2012) through cross-mapping of the best
524 match from UniRef90 or Swiss-Prot using the annotation mapping tables from UniProt. For each transcript,
525 we used the Virtual Ribosome (Dna2pep release 1.1) (Wernersson 2006) to predict the length of the longest
526 ORF searching across all reading frames without the constraint to begin translation from a methionine start
527 codon (parameters: -o none, -r all). The non-coding potential for each transcript was calculated using Portrait
528 (release 1.1) (Arrial et al. 2009).
529
530 ***Non-coding annotation of the assembled transcriptome***
531 Putative lncRNAs were classified based on a heuristic process considering the annotation results. The
532 constraints used to identify potential lncRNAs have to be considered very stringent (Extended Data Fig. 2).
533 In published studies, different combinations of analyses have been used to identify lncRNAs (Arrial et al.
534 2009; Pauli et al. 2012; Ulitsky et al. 2011; Cabili et al. 2011): 1) lack of similarity with proteins, 2) lack of
535 similarity with domain profiles, 3) lack of similarity with other RNAs (ribosomal, snoRNA, etc.), 4)
536 transcript and longest ORF lengths, 5) non-coding potential. We put all these together and classified as
537 lncRNA only those transcripts satisfying all the following conditions: a. length >= 200 nucleotides, b. lack of
538 similarity with any of the following: protein from Swiss-Prot and UniRef90, domains from CDD, rRNA from
539 GenBank and other small ncRNA from Rfam, c. longest ORF < 100 aminoacids, d. non-coding potential
540 score >= 0.95. Using these stringent constraints, we were able to predict in the *O. vulgaris* transcriptome
541 7,806 (~12%) transcripts as putative lncRNAs.

542

543 ***Assembly, mapping and annotation of the Octopus bimaculoides RNAseq data***

544 *O. bimaculoides* RNAseq raw data from Albertin et al. (Albertin et al. 2015) were downloaded from NCBI
545 SRA in October 2015 using the SRA Toolkit. Raw reads were filtered and trimmed based on quality and
546 adapter inclusion using Trimmomatic (release 0.33; parameters: -threads 32,
547 ILLUMINACLIP:illumina_adapters.fa:2:40:15:10:true LEADING:3 TRAILING:3
548 SLIDINGWINDOW:3:20 MINLEN:50). Read pairs with both reads passing the filters were considered for
549 the assembly. Trimmed and filtered reads were assembled with Trinity (release 2.1.0; parameters: --seqType
550 fq --SS_lib_type RF --CPU 32 --max_memory 240G --inchworm_cpu 32 --bflyHeapSpaceInit 24G --
551 bflyHeapSpaceMax 240G --bflyCalculateCPU --normalize_reads --min_kmer_cov 2 --jaccard_clip) using
552 digital normalization, strand information, the Jaccard clip and assuring that every kmer used in the assembly
553 was present in at least 2 reads to reduce noise. Redundancy of assembled transcripts was reduced using Cd-
554 hit (Li and Godzik 2006) (version: 4.6, parameters: -c 0.90 -n 8 -r 0 -M 0 -T 0). To measure the expression
555 levels, raw reads were mapped on the transcriptome using Bowtie (version 1, parameters: -t -q -p 32 --
556 chunkmbs 10240 --maxins 500 --trim5 2 --trim3 2 --seedlen 28 --tryhard -a -S). SAM outputs from Bowtie
557 were converted into BAM, sorted, indexed and counted using the view, sort, index and idxstats programs
558 from the Samtools software collection. All transcripts not showing at least 1 reads mapping per million
559 mapped reads (CPM) in at least 1 sample were discarded from the transcriptome. *Octopus bimaculoides*
560 genome was downloaded on August 2015 from
561 http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=Metazome.
562 Assembled and filtered unique transcripts were mapped on the genome using gmap (Wu and Watanabe 2005)
563 (version: 2015-09-28, parameters: --suboptimal-score 0 -f gff3_gene --gff3-add-separators 0 -t 32 --min-
564 trimmed-coverage 0.9 --min-identity 0.9) considering only transcripts aligning at least 90% of their length
565 with 90% minimum identity. We were able to map 84,043 (~90%) transcripts. Annotations and all the
566 remaining analysis were executed as for *O. vulgaris*.

567

568 ***Identification and classification of repetitive elements***

569 Repetitive elements for each transcriptome were annotated using RepeatMasker (A.F.A. Smit, R. Hubley &
570 P. Green RepeatMasker at http://repeatmasker.org; release 4.0.5) searching against the Repbase database

571 (Jurka et al. 2005) (release 20140131) with parameters: -species bilateria, -s, -gff. We counted from
572 RepeatMasker output the repeat fragments present at least once in each transcript and built a matrix
573 containing the percentage of transcripts containing fragments related to a. retroelements, b. DNA
574 transposons, c. satellites, d. simple-repeats, e. low complexity, f. other, and g. unknown classes for each
575 transcriptome according to the RepeatMasker classification.
576
577 ***Identification of full-length, putatively active transposable elements***
578 We parsed RepeatMasker output calculating the percentage of overlap between the assembled transcripts and
579 the repeat consensus from Repbase selecting all elements with at least 80% coverage on the repeat
580 consensus. Elements showing the highest coverage were selected. On these we used Virtual Ribosome to
581 predict the longest complete ORFs by searching across all reading frames with methionine as start codon and
582 a canonical stop (parameters: -o strict, -r all). A single transcript resulted with a complete ORF. On this we
583 used InterPro (Hunter et al. 2012) to identify and classify protein domains. The potential catalytic amino-
584 acids essential for the retrotransposition were manually identified comparing the putative translation with
585 those reported in Clements and Singer (Clements and Singer 1998). The same analysis was performed on
586 both the transcriptomes of *O. bimaculoides* (assembled by Albertin and assembled by us). The analysis was
587 also performed on the RepeatMasker annotation of the genome by Albertin downloaded from
588 http://octopus.unit.oist.jp/OCTDATA/. For consistency, we also analyzed RepeatMasker annotations of the
589 genome and the transcriptomes produced using the same tool, library and parameters used for *O. vulgaris*
590 and the other species considered in this study. In no-one of the analyses we could find a full-length
591 transposable element retaining a complete ORF for *O. bimaculoides*. We then translated the main
592 RepeatScout and RepeatModeler repeat libraries consensuses assembled by Albertin et al. (main RepeatScout
593 library: http://octopus.unit.oist.jp/OCTDATA/TE_FILES/mainrepeatlib.fa.gz; RepeatModeler library:
594 http://octopus.unit.oist.jp/OCTDATA/TE_FILES/oct.rm.tar.gz) with the Virtual Ribosome tool to predict
595 longest ORF searching across all reading frames showing methionine as start codon (parameters: -o strict, -r
596 all) and a canonical stop. The InterPro tool was then used to identify and classify the LINE characteristic
597 domains. The potential catalytic amino-acids essential for the retrotransposon activity were manually
598 identified comparing the ORF sequences with those reported in Clements and Singer. This led us to the
599 identification of two potentially functional LINE retrotransposons.

600

601 ***Identification of somatic retrotrasposition events in O. bimaculoides***

602 The two candidates retrotransposons found in *O. bimaculoides* RepeatModeler libraries were analyzed to
603 search for integration sites in gonads and optic lobe using MELT (Gardner et al. 2017) (v2.0.2). Two
604 genomic DNAseq WGS libraries from gonads (SRR2010220 and SRR2005790) were downloaded from the
605 European Nucleotide Archive (ENA) at https://www.ebi.ac.uk/ena/data/view/PRJNA270931. We generated
606 two additional DNAseq WGS libraries from DNA extracted from the optic lobe (L001 and L002) of a
607 different individual. *O. bimaculoides* reference genome was filtered for scaffold shorter than 10,000 bp and
608 reads mapped on it using BWA (Li and Durbin 2009) (v0.7.15; parameters: mem, -t 32). SAM output from
609 BWA were converted into BAM, sorted, indexed and counted using the view, sort, index and idxstats
610 programs from the SAMtools software. The resulted sorted BAM files were used as input for MELT
611 (parameters: -d 10000). Since the reads length differ between the two set of libraries (150 bp for gonads and
612 260 bp for optic lobe) the optic lobe dataset was trimmed with Trimmomatic (v0.32; parameters: CROP:200,
613 HEADCROP:50) to obtain homogeneous reads of 150 bp in all the datasets. We filtered the integration sites
614 (ISs) identified by MELT for entries which passed all the MELT checks and which presented at least 3
615 discordant pairs of reads supporting both left and right sides of the breakpoints. BLAST (v2.6.0; parameters:
616 -evalue 99999) search of the candidate retrotransposons consensus sequences was performed against the
617 genome and the identified ISs were additionally filtered out when the BLAST search showed similarity in a
618 range of 260 bp around the IS breakpoint. The same analysis was performed using non-trimmed reads and
619 two additional ISs identification programs, RetroSeq (Keane et al. 2013) and an in-house developed pipeline,
620 and the significance of results was maintained.

621

622 ***Phylogenetic tree generation***

623 Evolutionary tree in Figure 3D was generated using 100 full-length LINEs belonging to 15 LINE clades
624 (Extended Data Table 5). Protein sequences were selected from Ohshima and Okada (Ohshima and Okada
625 2005) and manually checked. InterPro was used to identify endonuclease and reverse transcriptase domains
626 in all the LINEs. Multiple sequence alignments were performed using MAFFT (Katoh and Standley 2013)
627 (v7.221; with option L-INS-i). We utilized TrimAl (Capella-Gutiérrez et al. 2009) (v1.4.rev15) to perform

automated trimming aligned sequences (parameters: -fasta -automated1). Phylogenetic relationships between LINE elements was reconstructed with MrBayes (Ronquist and Huelsenbeck 2003) (v3.2.1). Bayesian analysis was run for six million generations with twenty-two chains, sampling every 1,000 generations (6,000 samples). Convergence was attained with standard deviation of split frequencies below 0.01 and a consensus tree was generated using a burnin parameter of 1,500 (25% of 6,000 samples). The phylogenetic tree was visualized with FigTree program (release 1.4.2; http://tree.bio.ed.ac.uk/software/figtree/).

Classification of transcripts expressed in each sample, expression peaks and selection of candidates for validations

We classified subsets of transcripts according to their expression levels across the different parts. A transcript was considered expressed in a specific part if, in all the replicates of that part, it showed an expression level > 0.5 CPM. This resulted in the classification of the groups of transcripts used to perform the analysis (see: Figure 1c, Extended Data Fig. 4). We also classified transcripts having a peak of expression. These represent transcripts showing an expression level > 0.5 CPM in all three biological replicates of exclusively one part and below 0.5 in the others, which resulted in ~1,800 transcripts. They were used to draw the heatmap (Extended Data Fig. 3a) using MeV (release 4.8.1) as part of the TM4 suite (Saeed et al. 2003) with hierarchical clustering exploiting Pearson correlation. Candidates for validations were selected among the 1,800 transcripts with expression peak using the following additional criteria. For coding transcripts, we randomly selected four coding transcripts representing octopus homologs of homeobox genes. The annotation was manually verified. For lncRNAs we randomly selected four putative non-coding containing a SINE fragment. Both coding and lncRNA candidates were validated using RT-PCR and Sanger sequencing.

Polymerase chain reaction (PCR)

Octopus cDNAs were generated from 200 ng of total RNA using Superscript VILO cDNA Synthesis Kit (Life Technologies) in 20 µl reaction volume. PCR were carried out using 20 ng of cDNA, 0.25 µl of Taq DNA Polymerase (5U/µl; Roche), 1 µl of each specific forward and reverse primer (25pmol/µl), 2.5 µl of PCR reaction buffer (10x), 2.5 µl of dNTP mix (10x) and water (up to final volume: 25 µl). The Ubiquitin gene (accession number FJ617440) was used as internal control. Reactions for the coding transcripts were

656 amplified with a single step of 2 min at 94°C, 15 s at 94°C, 30s at 60°C and 1 min at 72°C for 35 cycles, and
657 7 min for 72°C. Reactions for the non-coding transcripts were amplified at annealing temperature of 58°C.
658 The following primers were utilized:
659 Arx forward 5'-TCCCTGCCTTCTCAACACAT-3'
660 Arx reverse 5'-TCCGAACTTCCACGCTTACT-3'
661 Hoxb5a forward 5'-GTGGCGAGGAATTTAGGAAG-3'
662 Hoxb5a reverse 5'-GCAACAGTCATAGTCCGAACAG-3'
663 Phox2b forward 5'-AATGGGGTGAGATCCTTTCC-3'
664 Phox2b reverse 5'-TTCATTGCAATCTCCTCTCG-3'
665 Meox2 forward 5'-TCCAGAACCGTCGGATGAAA-3'
666 Meox2 reverse 5'-TACGTAAAGGGCACACACCT-3'
667 Seml forward 5'-CACTTGTGCAAGGTACCACG-3'
668 Seml reverse 5'-AGGTCTCCTTAAATTTATTTCTGTGCA-3'
669 Subl forward 5'-ACAGAGCATCTTGAGTCTCACT-3'
670 Subl reverse 5'-CACTCCTGCGCCTTTCATTT-3'
671 Oll forward 5'-GGATTGACCCTGCAACTTGG-3'
672 Oll reverse 5'-CAGTGATGACGGACTTGCAA-3'
673 Arml forward 5'-GTACCCCACAAAATTAAATC-3'
674 Arml reverse 5'-CACTCACAAGGCTTTAGTTGGC-3'
675 Ubi forward 5'-TGTCAAGGCAAAGATTCAAGA-3'
676 Ubi reverse 5'-GGCCATAAACACACCAGCTC-3'
677
678 ***LINE copy number variation using quantitative real-time PCR***
679 Copy number variation analysis was performed on genomic DNA extracted from octopuses (N=9; SEM,
680 SUB, OL and ARM). One ARM sample was chosen as calibrator, while 18S was chosen as invariant control.
681 Purified genomic DNA concentrations were assessed by NanoDROP (ThermoFisher Scientific). According
682 to the starting concentration, DNA samples were diluted in TE buffer (10mM Tris-HCl, 1mM EDTA, pH 7.5)
683 to a concentration of 100 ng/µL and then further diluted to a concentration of 10 ng/µL. All dilutions were
684 checked by NanoDROP. Primers and multiplexing efficiencies were verified by linear regression to a

685 standard curve ranging from 50 ng to 16 pg of genomic DNA. LINE and 18S slopes were -3.3 and 3.8
686 respectively, and represent acceptable amplification efficiencies. Standard curves also confirmed that the
687 final concentration of 5ng DNA tested in qPCR was within the linear range of reaction. Reactions were
688 performed in 20 µL reaction mixture containing iQ Multiplex Powermix (Bio-Rad), Taqman primers (10
689 µM) and probes (10 µM) differentially labeled (with FAM or VIC fluorophore) and specifically designed to
690 hybridize with the target DNA sequences. LINE element was amplified using the following primers:
691 LINE forward 5'-AGCAGTGGGAATCATTCA-3'
692 LINE reverse 5'-GTCGTTTTCGTCGAACCAGT-3'
693 18S was amplified using the following primers:
694 18S forward 5'-AGTTCCGACCGTAAACGATG-3'
695 18S reverse 5'-CCCTTCCGTCAATTCCTTTA-3'
696 As probe sequences we utilized:
697 FAM: 5'-AACTCTGGGCCAAATTACGA-3'
698 VIC: 5'-GGGAAACCATAGTCGGTTCC-3'
699 qPCR was carried out for 20s at 90ºC, followed by 40 cycles of 10s at 95ºC and 30s at 59ºC using 7900HT
700 Fast Real Time PCR System (Applied Biosystems). Assays were performed for each sample in duplicate and
701 reproduced four times. Data obtained from the co-amplifications of the target DNA sequence and the internal
702 invariable control 18S were analyzed using the 2–ΔΔCt method79.
703
704 ***Southern Blotting***
705 The brains (SEM, SUB and OL) and a piece of an arm (ARM) of three *O. vulgaris* were dissected after
706 humane-killing and immediately frozen in liquid nitrogen. Pulverized samples were treated following the
707 methods utilized by Perelman and coworkers80; in brief, after phenol：chloroform (50:50) extraction DNA
708 was precipitated using cold isopropanol followed by centrifugation, suspended in TE buffer (10 mM Tris–
709 HCl pH 8.0 and 0.1 mM EDTA pH 8.0), treated with ribonuclease A (10 µg /ml) and incubated at 37°C for
710 30 min. DNA concentration was estimated using NanoDrop and quality checked by electrophoresis on 0.8%
711 agarose gel. 10 µg genomic DNA for each tissue was digested with EcoRI (New England Biolabs) overnight
712 at 37°C and resolved on a 0.9 % agarose gel for 15 hours at 1.5V. DNA was transferred to a Hybond-N+
713 nylon membrane (0.45 µm; Amersham Pharmacia Biotech) according to Sambrook and Russell81. DIG-

714 labeled LINE DNA probe was prepared by PCR DIG Probe synthesis kit (Roche). Hybridization and
715 autoradiography were performed according to the DIG Application Manual (Roche).
716
717 ***Probe synthesis for in-situ hybridization***
718 We amplified by PCR a 356-bp cDNA fragment of the assembled LINE from bp 1,512 to bp 1,868 using the
719 following primers:
720 LINE forward 5'-GGGTCAGAAAGTGACGAGGA-3'
721 LINE reverse 5'-TGCATGAGGCGGAGTTTAG-3'
722 The choice of the fragment and the design of primers have been based on manual curation steps ensuring that
723 the chosen fragment is present exclusively in the transcript of the identified LINE element and in no other
724 assembled transcripts. The amplified fragment was cloned into TOPO® TA Cloning® vector (Life
725 Technologies, CA, USA) according to manufacturer's protocol. Cloned fragment was digested using BAMHI
726 and ECORV restriction enzymes and validated by Sanger sequencing. Sense and antisense digoxigenin-
727 labelled RNA probes were generated by in vitro transcription using the DIG-RNA Labeling Kit (SP6/T7;
728 Roche Applied Sciences, QC, Canada). Labeled RNA probes were quantified by dot blot analysis.
729
730 ***In-situ hybridization experiments***
731 Brain masses (SEM, SUB, OL) and a segment of an arm from octopuses (N = 3) were fixed in
732 paraformaldehyde 4% (PFA) in phosphate buffered saline (PBS) at 4°C (3h for brain masses; overnight,
733 ARM). Samples were washed (four rinses in PBS), dehydrated in series of graded methanol/PBS (1:3, 1:1,
734 3:1 v/v) and stored at least one night in methanol (–20°C). Tissues were then rehydrated at 25°C in a series
735 of graded methanol/PBS (3:1, 1:1, 1:3 v/v) solutions and cryoprotected in 30% sucrose in PBS. After sucrose
736 infiltration, samples were embedded in tissue freezing medium (OCT; Leica Biosystems) and sectioned using
737 a cryostat (Leica CM3050 S). Sagittal and/or coronal sections (40 µm) were collected in PBST (phosphate
738 buffered saline including 0.1% Tween™ 20 and 0.2mM sodium azide). Washed free-floating sections were
739 mounted on Superfrost Plus slides (Menzel Gläser) and let dry overnight under fume hood. Hybridizations
740 were performed as described by Abler et al.82 with modifications. After rehydration in PSBT, the sections
741 were quenched at 25°C in 6% H_2O_2 (30 min) treated with proteinase K (10 min) and post-fixed with PFA-G
742 (4% paraformaldehyde and 0.2% glutaraldehyde) for 20 min. Prehybridization step was performed in

743 hybridization solution (HB: 50% formamide, 5XSSC, with 10µg/mL heparin, 10µg/mL yeast tRNA and, 1%
744 Blocking reagent) for at least 1h at 60°C and then incubated overnight in HB with the digoxigenin-labeled
745 riboprobes. Post-hybridization washes (50% formamide, 5x SSC, 1% SDS) were carried out for 2 hours at
746 60°C. Sections were washed in TNT (10mM Tris-HCL pH 7.5, 0.5M NaCl, 0.1% Tween™ 20) at 25°C and
747 incubated for 15min at 37°C with RNase (0.25µg/mL), followed by a FS (50% formamide, 5x SSC)
748 incubation of 2 hours (60°C). DIG was detected with an alkaline phosphatase labeled antibody (Roche).
749 After a saturation step in TBS pH 7.5 (10% sheep serum, 1% blocking reagent, 1% BSA, 0.1% Tween™ 20)
750 for 1h (room temperature), sections were incubated overnight at 4°C with antibodies (1:1,000; in TBS
751 containing: 5% sheep serum, 1% blocking reagent, 1% BSA). The following day, sections were washed for
752 2h in TBS (pH 7.5; 0.1% Tween™ 20 and 2mM levamisole) and then washed in alkaline phosphatase
753 solution (100mM Tris-HCL pH 9.5, 100mM NaCl, 50mM MgCl2, 0.1% Tween™ 20 and 2mM levamisole).
754 Bound antibodies were revealed using NBT-BCIP (Roche). After DIG in situ hybridization, slides were
755 counterstained with DAPI (5 µg/mL, Invitrogen) washed and mounted using aqueous mounting.
756
757 ***RTE-2_OV custom-made polyclonal antibodies***
758 Custom-made polyclonal antibodies were obtained from Primm Biotech Custom Antibody Services (Milan,
759 Italy) and raised against two peptides derived from two portions of the RTE-2_OV protein: GAA (1-100 aa),
760 and HAA (569-673 aa) resulting to be unique within the translation of the assembled transcriptome. To
761 choose the portion to select, the manufacturer also took into consideration: protein similarity (selection of
762 regions with no significant identity to the murine and rabbit proteome); low complexity and transmembrane
763 regions (exclusion of such regions); distribution of predicted antigenic peptides (selection of regions with a
764 high number of predicted antigenic peptide). The selected synthetic peptides were injected into two rabbits
765 and boosted three times within 38 days (at day: 21, 28, 35) after the first injection. The final bleeding was
766 conducted 3 days after the last injection, and the crude sera were purified on Sepharose columns by
767 immunoaffinity.
768
769 ***Immunohistochemistry and antibody validation***
770 SEM, SUB and OL dissected from *O. vulgaris* (N = 3) were immediately immersed 4% paraformaldehyde
771 (PFA) in seawater (4°C for 3h). After fixation samples were washed several times in PBS, cryoprotected in

772 sucrose 30% and embedded in OCT compound (OCT; Leica Biosystems). The embedded brain parts were
773 then sectioned at 20 μm using a cryostat (Leica CM3050 S). No antigen retrieval was required. Tissue
774 sections were rehydrated in three successive baths of 0.1 M PBS, and incubated for 90 min (at RT) in 5%
775 goat serum (Vector Laboratories Ltd) diluted in 0.1 M PBS containing 0.05% Tween (PBTw). The slices
776 were subsequently incubated at 4°C with custom-made polyclonal antibodies raised against LINE (G and H;
777 see RTE-2_OV custom-made polyclonal antibodies for details). The next day, slices were again washed by
778 several changes of PBTw and incubation (at RT for 90 min) with secondary antibodies was carried out using
779 Alexa Fluor 488 or 546 goat anti-rabbit IgG both diluted 1:200 in PBTw. Subsequently, sections were rinsed
780 and cell nuclei were counterstained with DAPI (Molecular Probes, Eugene, OR). Finally, after further
781 extensive washes, sections were mounted with fluorescent mounting medium (Fluoromount, Sigma). For all
782 antisera tested, omission of the primary antiserum and/or secondary antiserum resulted in negative staining.
783 In addition, specificity was assured by pre-incubating (4°C, overnight) the antibodies with 1mg/mL of
784 synthetic epitope (HAA and GAA, see RTE-2_OV custom-made polyclonal antibodies for details) before
785 staining. Again, no immunostaining was observed. The two custom-made polyclonal antibodies raised
786 against two different peptides of the RTE-2_OV protein stained the same spatial arrangement in the octopus
787 brain tissue.
788
789 ***Imaging***
790 Sections were observed under microscopes depending on the techniqueImage acquisition and processing
791 were performed using Leica Application Suite software (Leica Microsystems). For IHS we utilized a Leica
792 DMI6000 B inverted microscope, and for IHC a Leica TCS SP8 X confocal microscope (Leica
793 Microsystems, Germany). Tile Z-stacks were performed using a 0.2μm step size. IHC figures have been
794 assembled following guidelines for color blindness provided by Wong (Wong 2011).
795
796 ***Data availability***
797 *Octopus vulgaris* RNAseq raw reads have been submitted to the ArrayExpress database
798 (www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-3957. The data will be made public upon
799 acceptance of this article. Login credential for referees are indicated in the first page of the supplementary
800 pdf.

Acknowledgments

The work has been supported by: Progetto Premiale MolEcOC (Italian Ministry of Education, University and Research, MIUR); Flagship project RITMARE (MIUR and Stazione Zoologica); BIOforIU PON Project (MIUR and European Regional Development Fund, FESR). Giuseppe Petrosino, Swaraj Basu, Massimiliano Volpe and Giulia Di Cristina have been supported by a SZN PhD fellowship.

Authors Contributions

RS and GF conceived, designed and led the project; GPe, MV and RS performed bioinformatics analysis and interpreted the results with significant contributions from FM and SB; GPo carried out the in-situ hybridization and IHC experiments and choice sectioning and sampling of different animals; IZ cared of biological sampling and performed RNA extractions; SF and SG carried out the quantitative PCR experiments; DP and VB performed the RNAseq; OS and CA carried out extraction and sequencing of O. bimaculoides samples; IZ, CL, GDC, MTR, MIF and FR carried out the remaining molecular biology experiments and validations. RS and GF wrote the manuscript. All the Authors discussed the results and contributed to the manuscript.

Disclosure declaration

The authors declare no competing financial interests.

Current addresses

Giuseppe Petrosino: Institute of Molecular Biology (IMB), Mainz, Germany; Francesco Musacchia: Central RNA Laboratory, Istituto Italiano di Tecnologia (IIT), Genova, Italy; Swaray Basu: Strand Life Sciences, Bengaluru, India; Giulia Di Cristina: Institute of Zoology, University of Cologne, Cologne, Germany; Dinko Pavlinic: Institute of Molecular and Clinical Ophthalmology, Basel, Switzerland; Massimiliano Volpe: Department of Biomedical and Clinical Sciences, Linköping University, Linköping, Sweden; Oleg Simakov: Department of Molecular Evolution and Development, Wien University, Wien, Austria.

828 # References

829

Abbott NJ, Williamson R, Maddock L, eds. 1995. *Cephalopod Neurobiology: Neuroscience Studies in Squid, Octopus and Cuttlefish*. Oxford University Press, Oxford, New York.

Albertin CB, Simakov O, Mitros T, Wang ZY, Pungor JR, Edsinger-Gonzales E, Brenner S, Ragsdale CW, Rokhsar DS. 2015. The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature* **524**: 220–224.

Amodio P, Andrews P, Salemme M, Ponte G, Fiorito G. 2014. The use of artificial crabs for testing predatory behavior and health in the octopus. *ALTEX - Alternatives to animal experimentation* **31**: 494–499.

Andrews PLR, Darmaillacq A-S, Dennison N, Gleadall IG, Hawkins P, Messenger JB, Osorio D, Smith VJ, Smith JA. 2013. The identification and management of pain, suffering and distress in cephalopods, including anaesthesia, analgesia and humane killing. *Journal of Experimental Marine Biology and Ecology* **447**: 46–64.

Arrial RT, Togawa RC, Brigido M de M. 2009. Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus Paracoccidioides brasiliensis. *BMC Bioinformatics* **10**: 239.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**: 25–29.

Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, Brennan PM, Rizzu P, Smith S, Fell M, et al. 2011. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* **479**: 534–537.

Bairoch A. 2000. The ENZYME database in 2000. *Nucleic Acids Res* **28**: 304–305.

Baldascino E, Di Cristina G, Tedesco P, Hobbs C, Shaw TJ, Ponte G, Andrews PLR. 2017. The Gastric Ganglion of Octopus vulgaris: Preliminary Characterization of Gene- and Putative Neurochemical-Complexity, and the Effect of Aggregata octopiana Digestive Tract Infection on Gene Expression. *Frontiers in Physiology* **8**. https://www.frontiersin.org/article/10.3389/fphys.2017.01001.

Belcaid M, Casaburi G, McAnulty SJ, Schmidbaur H, Suria AM, Moriano-Gutierrez S, Pankey MS, Oakley TH, Kremer N, Koch EJ, et al. 2019. Symbiotic organs shaped by distinct modes of genome evolution in cephalopods. *PNAS* **116**: 3030–3035.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics* btu170.

Borrelli L, Fiorito G. 2008. 1.31 - Behavioral Analysis of Learning and Memory in Cephalopods. In *Learning and Memory: A Comprehensive Reference* (ed. J.H. Byrne), pp. 605–627, Academic Press, Oxford http://www.sciencedirect.com/science/article/pii/B9780123705099000693 (Accessed December 30, 2020).

Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, Eddy SR, Gardner PP, Bateman A. 2013. Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res* **41**: D226-232.

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development* **25**: 1915–1927.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972–1973.

Clements AP, Singer MF. 1998. The human LINE-1 reverse transcriptase:effect of deletions outside the common reverse transcriptase domain. *Nucleic Acids Res* **26**: 3528–3535.

Consortium TU. 2015. UniProt: a hub for protein information. *Nucl Acids Res* **43**: D204–D212.

da Fonseca RR, Couto A, Machado AM, Brejova B, Albertin CB, Silva F, Gardner P, Baril T, Hayward A, Campos A, et al. 2020. A draft genome sequence of the elusive giant squid, Architeuthis dux. *GigaScience* **9**. https://doi.org/10.1093/gigascience/giz152 (Accessed December 30, 2020).

De Cecco M, Ito T, Petrashen AP, Elias AE, Skvir NJ, Criscione SW, Caligiana A, Brocculi G, Adney EM, Boeke JD, et al. 2019. L1 drives IFN in senescent cells and promotes age-associated inflammation. *Nature* **566**: 73–78.

Edelman DB, Seth AK. 2009. Animal consciousness: a synthetic approach. *Trends Neurosci* **32**: 476–484.

Fiorito G, Affuso A, Anderson DB, Basil J, Bonnaud L, Botta G, Cole A, D'Angelo L, De Girolamo P, Dennison N, et al. 2014. Cephalopods in neuroscience: regulations, research and the 3Rs. *Invert Neurosci* **14**: 13–36.

Fiorito G, Affuso A, Basil J, Cole A, de Girolamo P, D'Angelo L, Dickel L, Gestal C, Grasso F, Kuba M, et al. 2015. Guidelines for the Care and Welfare of Cephalopods in Research - A consensus based on an initiative by CephRes, FELASA and the Boyd Group. *Lab Anim* **49**: 1–90.

Fiorito G, von Planta C, Scotto P. 1990. Problem solving ability of Octopus vulgaris Lamarck (Mollusca, Cephalopoda). *Behavioral and Neural Biology* **53**: 217–230.

Friocourt G, Poirier K, Rakić S, Parnavelas JG, Chelly J. 2006. The role of ARX in cortical development. *Eur J Neurosci* **23**: 869–876.

Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, Mills RE, Consortium T 1000 GP, Devine SE. 2017. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res* **27**: 1916–1929.

Gray EG. 1970. The fine structure of the vertical lobe of octopus brain. *Philos Trans R Soc Lond, B, Biol Sci* **258**: 379–394.

Gray EG, Young JZ. 1964. Electron Microscopy of Synaptic Structure of Octopus Brain. *J Cell Biol* **21**: 87–103.

Guan Z, Cai T, Liu Z, Dou Y, Hu X, Zhang P, Sun X, Li H, Kuang Y, Zhai Q, et al. 2017. Origin of the Reflectin Gene and Hierarchical Assembly of Its Protein. *Current Biology* **27**: 2833-2842.e6.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protocols* **8**: 1494–1512.

Hochner B. 2010. Functional and comparative assessments of the octopus learning and memory system. *Front Biosci (Schol Ed)* **2**: 764–771.

Hochner B. 2013. How nervous systems evolve in relation to their embodiment: what we can learn from octopuses and other molluscs. *Brain, behavior and evolution* **82**: 19–30.

Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, et al. 2012. InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res* **40**: D306-312.

Jachowicz JW, Bing X, Pontabry J, Bošković A, Rando OJ, Torres-Padilla M-E. 2017. LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. *Nat Genet* **49**: 1502–1510.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**: 462–467.

Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte C. 2013. Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. *PLoS Genet* **9**: e1003470.

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

Kazazian HH. 2004. Mobile Elements: Drivers of Genome Evolution. *Science* **303**: 1626–1632.

Keane TM, Wong K, Adams DJ. 2013. RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics* **29**: 389–390.

Kelley D, Rinn J. 2012. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biology* **13**: R107.

Kim B-M, Kang S, Ahn D-H, Jung S-H, Rhee H, Yoo JS, Lee J-E, Lee S, Han Y-H, Ryu K-B, et al. 2018. The genome of common long-arm octopus Octopus minor. *GigaScience* **7**. https://doi.org/10.1093/gigascience/giy119 (Accessed December 30, 2020).

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

Lee EJ, Banerjee S, Zhou H, Jammalamadaka A, Arcila M, Manjunath BS, Kosik KS. 2011. Identification of piRNAs in the central nervous system. *RNA* **17**: 1090–1099.

Lee PN, Callaerts P, de Couet HG, Martindale MQ. 2003. Cephalopod Hox genes and the origin of morphological novelties. *Nature* **424**: 1061–1065.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658–1659.

Liscovitch-Brauer N, Alon S, Porath HT, Elstein B, Unger R, Ziv T, Admon A, Levanon EY, Rosenthal JJC, Eisenberg E. 2017. Trade-off between Transcriptome Plasticity and Genome Evolution in Cephalopods. *Cell* **169**: 191-202.e11.

Maldonado H. 1965. The positive and negative learning process in Octopus vulgaris Lamarck. Influence of the vertical and median superior frontal lobes. *Z Vergl Physiol* **51**: 185–203.

Maldonado H. 1963. The visual attack learning system in Octopus vulgaris. *J Theor Biol* **5**: 470–488.

Malik HS, Eickbush TH. 1998. The RTE class of non-LTR retrotransposons is widely distributed in animals and is the origin of many SINEs. *Mol Biol Evol* **15**: 1123–1134.

Mankoo BS, Collins NS, Ashby P, Grigorieva E, Pevny LH, Candia A, Wright CVE, Rigby PWJ, Pachnis V. 1999. Mox2 is a component of the genetic hierarchy controlling limb muscle development. *Nature* **400**: 69–73.

Marchler-Bauer A, Zheng C, Chitsaz F, Derbyshire MK, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Lanczycki CJ, et al. 2013. CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res* **41**: D348-352.

Marini G, De Sio F, Ponte G, Fiorito G. 2017. Behavioral Analysis of Learning and Memory in Cephalopods. In *Learning and Memory: A Comprehensive Reference (Second Edition)* (ed. J.H. Byrne), Vol. Volume 1-Learning Theory and Behavior (Menzel, Randolf-volume Editor) of, pp. 441–462, Academic Press, Elsevier, Amsterdam, The Netherlands.

Mita P, Sun X, Fenyö D, Kahler DJ, Li D, Agmon N, Wudzinska A, Keegan S, Bader JS, Yun C, et al. 2020. BRCA1 and S phase DNA repair pathways restrict LINE-1 retrotransposition in human cells. *Nature Structural & Molecular Biology* **27**: 179–191.

Mita P, Wudzinska A, Sun X, Andrade J, Nayak S, Kahler DJ, Badri S, LaCava J, Ueberheide B, Yun CY, et al. 2018. LINE-1 protein localization and functional dynamics during the cell cycle ed. S.P. Goff. *eLife* **7**: e30058.

Morgat A, Coissac E, Coudert E, Axelsen KB, Keller G, Bairoch A, Bridge A, Bougueleret L, Xenarios I, Viari A. 2012. UniPathway: a resource for the exploration and annotation of metabolic pathways. *Nucleic Acids Res* **40**: D761-769.

Muotri AR, Chu VT, Marchetto MCN, Deng W, Moran JV, Gage FH. 2005. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435**: 903–910.

Musacchia F, Basu S, Petrosino G, Salvemini M, Sanges R. 2015. Annocript: a flexible pipeline for the annotation of transcriptomes able to identify putative long noncoding RNAs. *Bioinformatics* **31**: 2199–2201.

Nixon M, Young JZ. 2003. *The Brains and Lives of Cephalopods*. Oxford University Press.

Ohira R, Zhang Y-H, Guo W, Dipple K, Shih SL, Doerr J, Huang B-L, Fu LJ, Abu-Khalil A, Geschwind D, et al. 2002. Human ARX gene: genomic characterization and expression. *Molecular Genetics and Metabolism* **77**: 179–188.

Ohshima K, Okada N. 2005. SINEs and LINEs: symbionts of eukaryotic genomes with a common tail. *Cytogenetic and Genome Research* **110**: 475–490.

Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**: 1061–1067.

Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, Fan L, Sandelin A, Rinn JL, Regev A, et al. 2012. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res* **22**: 577–591.

Percharde M, Lin C-J, Yin Y, Guan J, Peixoto GA, Bulut-Karslioglu A, Biechele S, Huang B, Shen X, Ramalho-Santos M. 2018. A LINE1-Nucleolin Partnership Regulates Early Development and ESC Identity. *Cell* **174**: 391-405.e19.

Perrat PN, DasGupta S, Wang J, Theurkauf W, Weng Z, Rosbash M, Waddell S. 2013. Transposition-Driven Genomic Heterogeneity in the Drosophila Brain. *Science* **340**: 91–95.

Qureshi IA, Mattick JS, Mehler MF. 2010. Long non-coding RNAs in nervous system function and disease. *Brain Research* **1338**: 20–35.

Rajasethupathy P, Antonov I, Sheridan R, Frey S, Sander C, Tuschl T, Kandel ER. 2012. A Role for Neuronal piRNAs in the Epigenetic Control of Memory-Related Synaptic Plasticity. *Cell* **149**: 693–707.

Ritschard Elena A., Fitak Robert R., Simakov Oleg, Johnsen Sönke. 2019. Genomic signatures of G-protein-coupled receptor expansions reveal functional transitions in the evolution of cephalopod signal transduction. *Proceedings of the Royal Society B: Biological Sciences* **286**: 20182929.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572–1574.

Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, et al. 2003. TM4: a free, open-source system for microarray data management and analysis. *BioTechniques* **34**: 374–378.

Shigeno S, Andrews PLR, Ponte G, Fiorito G. 2018. Cephalopod Brains: An Overview of Current Knowledge to Facilitate Comparison With Vertebrates. *Front Physiol* **9**: 952.

Shigeno S, Ragsdale CW. 2015. The gyri of the octopus vertical lobe have distinct neurochemical identities. *J Comp Neurol* **523**: 1297–1317.

Shomrat T, Zarrella I, Fiorito G, Hochner B. 2008. The octopus vertical lobe modulates short-term learning rate and uses LTP to acquire long-term memory. *Curr Biol* **18**: 337–342.

Singer T, McConnell MJ, Marchetto MCN, Coufal NG, Gage FH. 2010. LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes? *Trends Neurosci* **33**: 345–354.

Styfhals R, Seuntjens E, Simakov O, Sanges R, Fiorito G. 2019. In silico Identification and Expression of Protocadherin Gene Family in Octopus vulgaris. *Front Physiol* **9**. https://www.frontiersin.org/articles/10.3389/fphys.2018.01905/full (Accessed December 30, 2020).

Taylor MS, Altukhov I, Molloy KR, Mita P, Jiang H, Adney EM, Wudzinska A, Badri S, Ischenko D, Eng G, et al. 2018. Dissection of affinity captured LINE-1 macromolecular complexes ed. S.P. Goff. *eLife* **7**: e30094.

Thomas CA, Tejwani L, Trujillo CA, Negraes PD, Herai RH, Mesci P, Macia A, Crow YJ, Muotri AR. 2017. Modeling of TREX1-Dependent Autoimmune Disease using Human Stem Cells Highlights L1 Accumulation as a Source of Neuroinflammation. *Cell Stem Cell* **21**: 319-331.e8.

Turchetti-Maia A, Shomrat T, Hochner B. 2017. The Vertical Lobe of Cephalopods. A Brain Structure Ideal for Exploring the Mechanisms of Complex Forms of Learning and Memory. In *The Oxford Handbook of Invertebrate Neurobiology* (ed. J.J. Byrne), pp. 1–27, Oxford University Press, Oxford, UK.

Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. 2011. Conserved Function of lincRNAs in Vertebrate Embryonic Development despite Rapid Sequence Evolution. *Cell* **147**: 1537–1550.

Upton KR, Gerhardt DJ, Jesuadian JS, Richardson SR, Sánchez-Luque FJ, Bodea GO, Ewing AD, Salvador-Palomeque C, van der Knaap MS, Brennan PM, et al. 2015. Ubiquitous L1 mosaicism in hippocampal neurons. *Cell* **161**: 228–239.

Wang L, Dou K, Moon S, Tan FJ, Zhang ZZ. 2018. Hijacking Oogenesis Enables Massive Propagation of LINE and Retroviral Transposons. *Cell* **174**: 1082–1094.

Wang ZY, Ragsdale CW. 2017. Cadherin genes and evolutionary novelties in the octopus. *Seminars in Cell & Developmental Biology* **69**: 151–157.

Wernersson R. 2006. Virtual Ribosome--a comprehensive DNA translation tool with support for integration of sequence feature annotation. *Nucleic Acids Res* **34**: W385–W388.

Wong B. 2011. Points of view: Color blindness. *Nature Methods* **8**: 441–441.

Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**: 1859–1875.

Xiong Y, Eickbush TH. 1993. Dong, a non-long terminal repeat (non-LTR) retrotransposable element from Bombyx mori. *Nucleic Acids Res* **21**: 1318–1318.

Young JZ. 1991. Computation in the learning system of cephalopods. *Biol Bull* **180**: 200–208.

Young JZ. 1995. Multiple matrices in the memory system of Octopus. In *Cephalopod Neurobiology* (eds. J.N. Abbott, R. Williamson, and L. Maddock), pp. 431–443, Oxford University Press, Oxford.

Young JZ. 1971. *The anatomy of the nervous system of Octopus vulgaris*. Oxford University Press, London, UK.

Young JZ. 1963. The Number and Sizes of Nerve Cells in Octopus. *Proceedings of the Zoological Society of London* **140**: 229–254.

Zarrella I, Herten K, Maes GE, Tai S, Yang M, Seuntjens E, Ritschard EA, Zach M, Styfhals R, Sanges R, et al. 2019. The survey and reference assisted assembly of the Octopus vulgaris genome. *Sci Data* **6**: 1–8.

830