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

Background: 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.

Results: 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 driving germline polymorphisms among individuals and somatic polymorphisms 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 the transcriptomes of both *Octopus* species, further testifying the crucial activity of TEs in the evolution of the octopus genomes.

Conclusions: The neural transcriptome of the octopus shows the transcription of thousands of putative lncRNAs and of a full lenght LINE element belonging to the RTE class. We speculate that a convergent evolutionary process involving retrotransposons activity in the brain has been important for the evolution of sophisticated cognitive abilities in this genus.

Keywords

Mollusks; Nervous system; Transcriptome; Transposable elements

Background

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) [1–3]. 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 [4] and causes somatic mosaicism establishing genomic variability in the brain [5]. Somatic L1 insertions are suggested to alter the transcriptional output of individual neurons, eventually affecting neuronal plasticity and behavior [6]. Transposition-driven genomic heterogeneity has also been documented in invertebrates including the neurons of mushrooms bodies of *Drosophila melanogaster* [7] 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* [8], *Drosophila melanogaster* [7] and in the mouse brain [9], 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 [10, 11] and to be a component of extra-chromosomal DNA [12].

Among invertebrates, the cephalopod mollusk *Octopus vulgaris* is known for the richness of its behavioral repertoire achieving sophisticated vertebrate-like plasticity and neural control [13–16]. The remarkable innovations in the morphological and functional organization of its nervous system [14, 17, 18] are linked to the evolution of an unprecedented complexity coded both at cellular and molecular levels [17–21]. 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 [22]. 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 [21, 24–28].

The sequencing of the *Octopus bimaculoides* genome [24] 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* [29] and the other sequenced cephalopod species such as *Euprymna scolopes* [30] and the giant squid *Architeuthis dux* [31]. Performing a survey of the *O. vulgaris* genome we have confirmed the expansion of TEs also in this species [32]. 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 expressed in specific areas of the brain related to known forms of behavioral plasticity including learning. We also provide evidence for the transcription of thousands of long non-coding RNAs and the pervasive inclusion of TE fragments in coding and non-coding transcripts in both *O. vulgaris* and *O. bimaculoides*.

113 # Results

114

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

116 We generated a *de-novo* assembly of the *Octopus vulgaris* neural transcriptome identifying and evaluating its
117 functional annotations, potential lncRNAs and repeats composition (Fig. 1; Supplemental Tables S1, S2;
118 Supplemental Fig. S1, S2; Supplemental Files S1, S2, S3). From each of three different octopus individuals
119 we collected four parts, three of them as representative of the central brain: 1) supra-esophageal mass (SEM),
120 2) sub-esophageal mass (SUB) and 3) optic lobe (OL); and one representing the peripheral nervous system:
121 4) a piece of the second left arm (ARM) including the arm nerve cord. We used these parts as source of RNA
122 for the sequencing (see Methods, Supplemental Table S1). The sequencing generated approximately 850
123 million paired-end reads accounting for 85 Gbp of sequence data that, following assembly and filtering,
124 produced 64,477 unique transcripts (Supplemental File S1). The sequences showed a N50 value of 2,087 bp,
125 average length of 1,308 and about 38% average GC-content (Supplemental Table S2). The transcriptome
126 resulted to be more than 98% complete and we functionally annotated 21,030 (32.6%) protein coding
127 transcripts (Supplemental Fig. S1 and Supplemental File S3). By performing stringent annotation analysis a
128 high proportion of transcripts (7,806; 12.1%) resulted to be putative lncRNAs (see Methods and
129 Supplemental Fig. S2). In analogy to what is known about lncRNAs expression in mammals [33], the non-
130 coding portion of the *O. vulgaris* transcriptome shows a lower level of expression when compared to protein
131 coding genes (Fig. 1A) and a significantly higher number of lncRNAs results to be expressed in central brain
132 (~10%) with respect to the arm (~7%, p-value < 1e-40) (Fig. 1B). Functional enrichments highlight
133 differences between transcripts expressed in the central brain and in the periphery. Specifically, transcripts
134 expressed in the brain result enriched in functions associated with neuronal cell adhesion and reverse
135 transcription, while the ones expressed in the arm are enriched for functions associated with signal
136 transduction and translation (Fig. 1C and Supplemental Fig. S3).

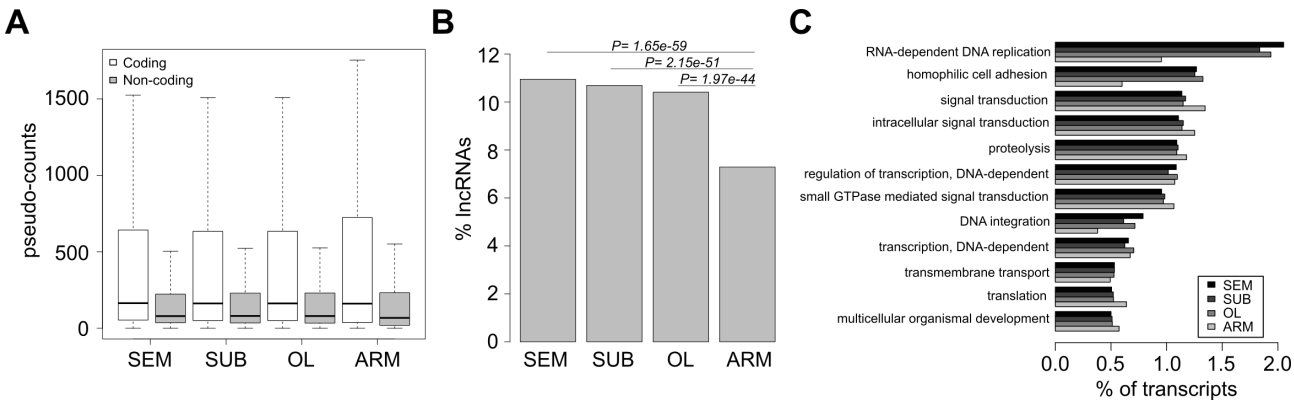

Figure 1. Features of the *Octopus vulgaris* brain and arm transcriptome. We sequenced the supra-esophageal (SEM) and sub-esophageal (SUB) masses and optic lobe (OL) as representatives of the brain, and the medial segment of an arm (ARM), including the arm nerve cord, as representative of peripheral 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 have 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 RT-PCRs concerning the lncRNAs, *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) resulted to be tissue-specific, as they were mainly expressed in

165 the SUB and in the ARM, respectively. On the other hand, *Seml* (lncRNAs with a peak of expression in
166 SEM, comp18661_c0_seq1) resulted to be expressed in the SEM but also in the SUB and the OL, while *Oll*
167 (lncRNA with a peak of expression in the OL, comp35506_c7_seq1) resulted expressed in the OL but also in
168 all the other sampled parts (Figure 2C). Both *Oll* and *Arml* show the existence of two different isoforms in at
169 least one individual. The RT-PCR results were generally in agreement with the sequencing data for both
170 coding and non-coding transcripts tested. The identifiers indicated for every transcript are the same used in
171 the Supplementary files and can be used to identify the corresponding sequences and annotations.
172
173 In order to verify the presence of lncRNAs also in *O. bimaculoides* we reassembled the public RNAseq data
174 from this species with the same method used for our *O. vulgaris* data (see Methods). We assembled 92,820
175 unique transcripts (Supplemental Table S2 and Supplemental File S4) of which 84,043 (90%) map on the
176 published assembled genome with at least 90% coverage and 90% identity (Supplemental File S5). Our
177 analysis demonstrated the presence, also in this octopus species, of thousands of putative lncRNAs (File S6).
178 Indeed we were able to classify 10,030 assembled transcript as putative non-coding. They correspond to
179 more than 10% of the assembled transcriptome of which 9,132 map on the reference genome.
180

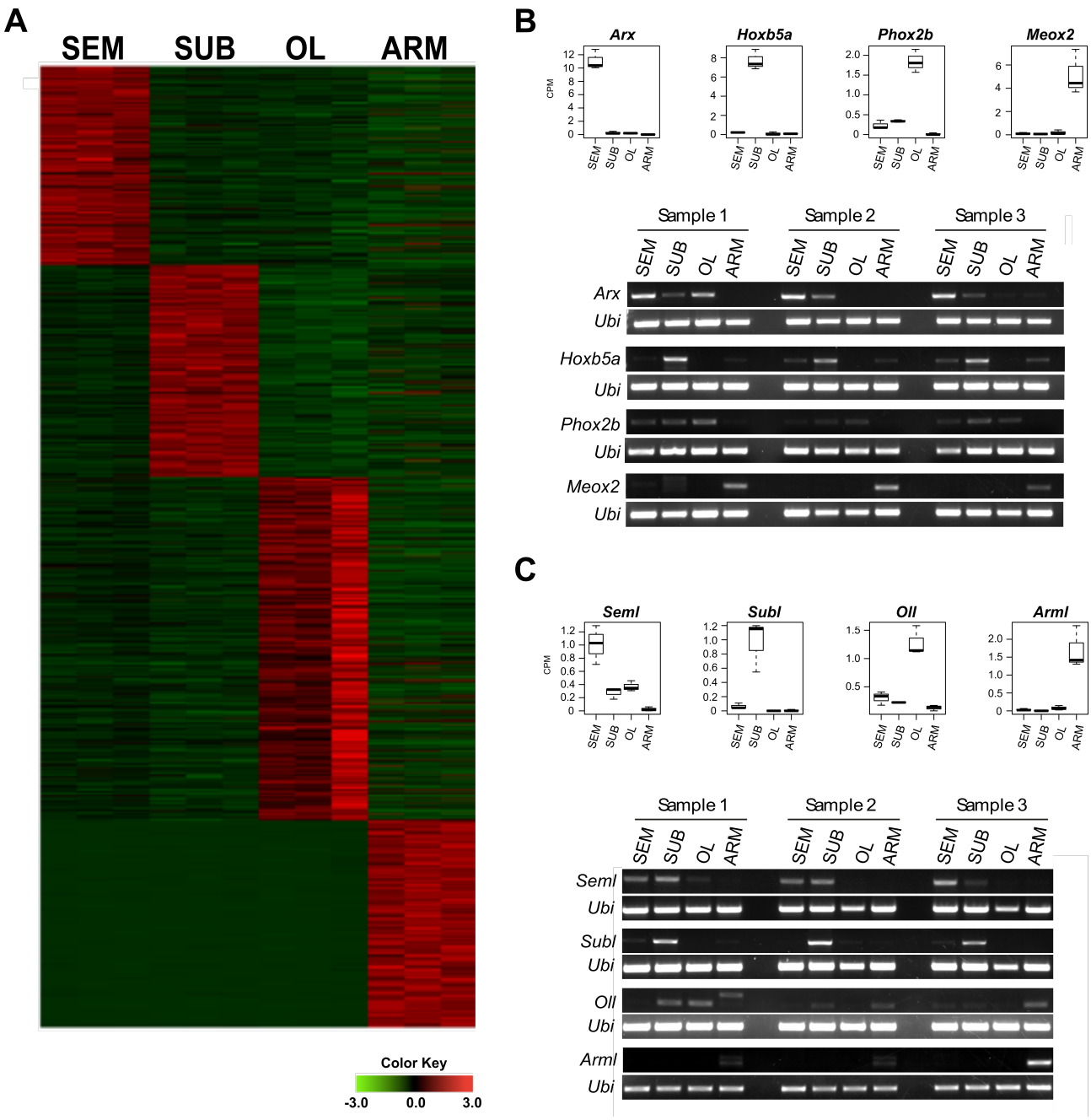

Figure 2. Transcripts with an expression peak and transcriptome validations. (A) Heatmap showing the expression levels for transcripts classified as having a peak of expression. (B) Boxplots showing RNAseq expression levels of coding transcripts selected for validation and their relative RT-PCR results 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

193 S3). More than 35% of the generated transcripts contain at least one interspersed repeat fragment. Among
194 them, retroelements represent the TE fragments more frequently embedded in transcripts (26% of transcripts
195 contain a fragment from at least one retroelement). According to the segregation of transposable elements in
196 transcripts expressed in the different sequenced parts of the organism, we observed that SINEs fragments are
197 present in a higher fraction of transcripts expressed in the brain, while LINEs, LTRs and DNA transposons
198 are present in a higher portion of transcripts expressed in the periphery (Supplemental Fig. S4A). SINE also
199 results the class of retroelements more frequently embedded in lncRNAs (Supplemental Fig. S4B).
200
201 To identify the molecular basis of the observed TEs expansion in *O. vulgaris* we searched the transcriptome
202 for putative autonomous active elements. We found a single element, a LINE mainly transcribed in the brain
203 (Fig. 3A-C) that we named RTE-2_OV following phylogenetic analysis from which the element resulted to
204 belong to the RTE clade (Fig. 3D). The identified LINE presents an ORF of 3,327 nucleotides and 5' and 3'
205 UTRs of about 600 nucleotides. The translation corresponds to a 1,109 aminoacids polypeptide chain
206 containing all the catalytic aminoacids and domains needed for retrotransposition (Fig. 3A,B): a C-terminal
207 endonuclease (EN), a reverse transcriptase (RT) and an N-terminal C2H2 zinc-finger (Znf) which is
208 relatively rare in RTE elements [34]. The finding that this element contains an intact ORF and is transcribed
209 indicates that it might be active and possibly drive retrotransposition. We then asked whether this element
210 might show evidence of somatic retrotransposition. Southern blots analysis did not lead to conclusive results
211 because of high background noise likely due to the high number of copies of this TE in the genome of *O.*
212 *vulgaris*, however the pattern observed is consistent with the existence of germinal and somatic genomic
213 variations associated to the element (Supplemental Fig. S5). We also performed quantitative PCR
214 experiments on DNA deriving from the four parts of different individuals. Results add support to the
215 presence of somatic retrotransposition in the brain (Fig. 3E) showing an average higher content of RTE-
216 2_OV DNA in the SUB with respect to the other parts. This result suggests the existence of a higher amount
217 of RTE-2_OV DNA in cells from the SUB, whereas cells from other parts show a similar average content.
218 The existence of these increased amounts of RTE-2_OV DNA in SUB is in line with its higher expression in
219 the SUB (Fig. 3F). Interestingly, the putative homolog of Piwi, known to repress the translation of TEs and
220 therefore to restrict retrotransposition, displays a lower expression in the SUB with respect to the other
221 portions of the central brain of (Fig. 3G) consistently with what has been observed in the fruitfly brain [7].

222

223 In order to add further support to the existence of active LINEs in the genomes of the Octopus *genus* we
224 searched the public genome, transcriptome and our custom transcriptome assembly (Supplementary Files S4-
225 S6) of *O. bimaculoides* for the presence of assembled full-length LINEs. We were not able to identify any
226 assembled full-length LINE in the transcriptome nor in the genome. However, when inspecting the repeat
227 library consensus generated in the work by Albertin *et al.* [24] we identified two LINEs with a full-length
228 ORF and the complete set of domains: an RTE and a Dong [35] element. To gain insights into the possibility
229 that the identified LINE elements might be active we generated DNAseq sequencing data from the optical
230 lobe of an *O. bimaculoides* individual and searched for novel integration sites (IS) with respect to the
231 reference genome using MELT [36]. The same analysis was performed using the public DNAseq data
232 derived by the gonads of the individual used to assemble the reference *O. bimaculoides* genome. Comparing
233 the whole genome sequencing data from the two different *O. bimaculoides* individuals we obtained
234 significant evidences of activity only for the RTE element. Indeed, the element showed a significantly higher
235 number of non-reference ISs with respect to the reference individual (Fig. 3H, p-value 4.8e-69) which
236 indicates that this element is at the basis of polymorphisms between the two individuals. Conversely, the
237 Dong element resulted in a very low number of non-reference ISs which does not significantly differ from
238 the results obtained with reads from the reference individual and therefore does not display evidences of
239 recent activity.

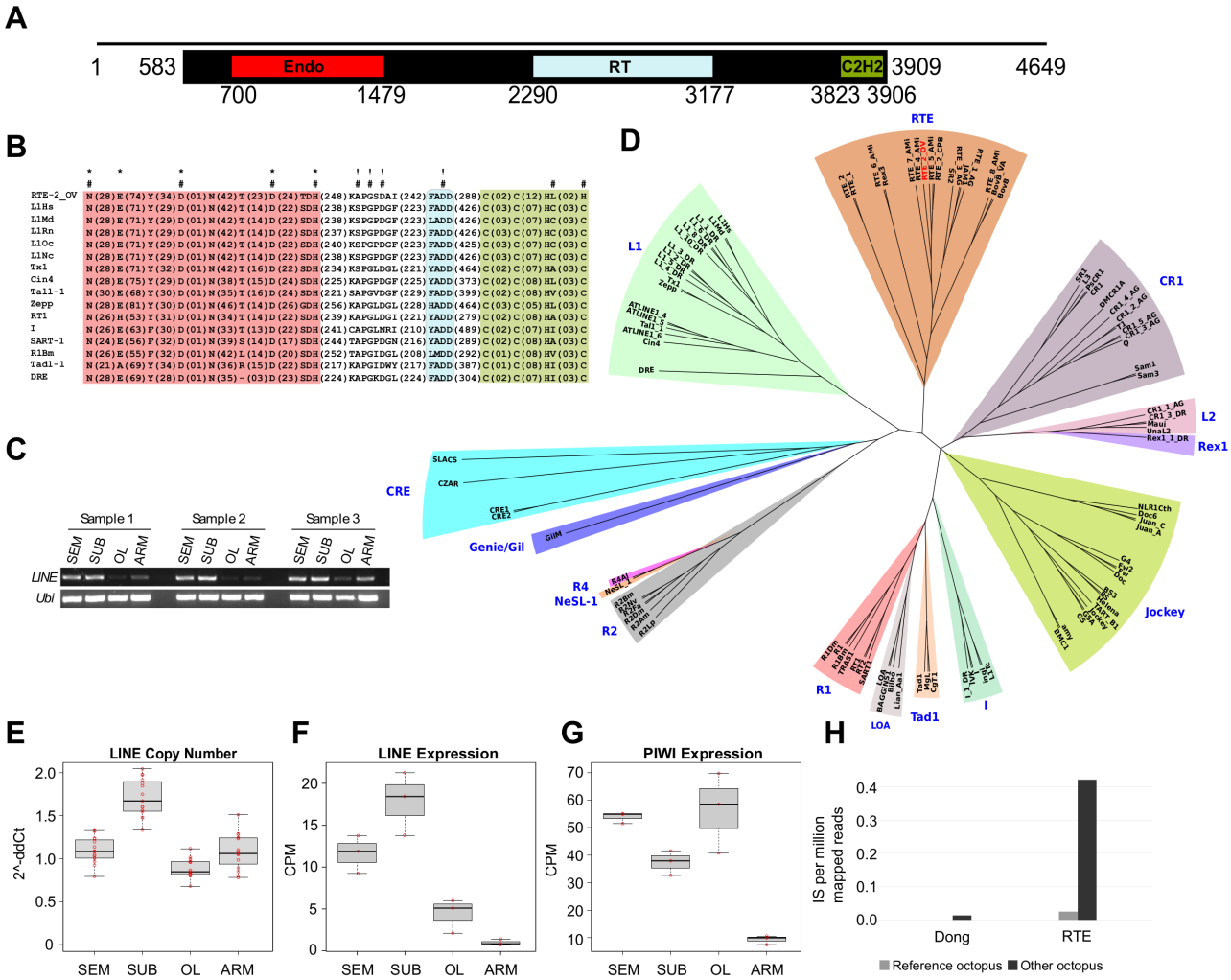

Figure 3. A full-length potentially active LINE transcribed in *Octopus vulgaris*. (A) Domain composition of the discovered LINE (comp36575_c1_seq3). Black line represents the transcript, black box the location of the ORF and colored boxes the protein domains (Endo, endonuclease; RT, reverse transcriptase; 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 [37] 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 [38] (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.

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

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

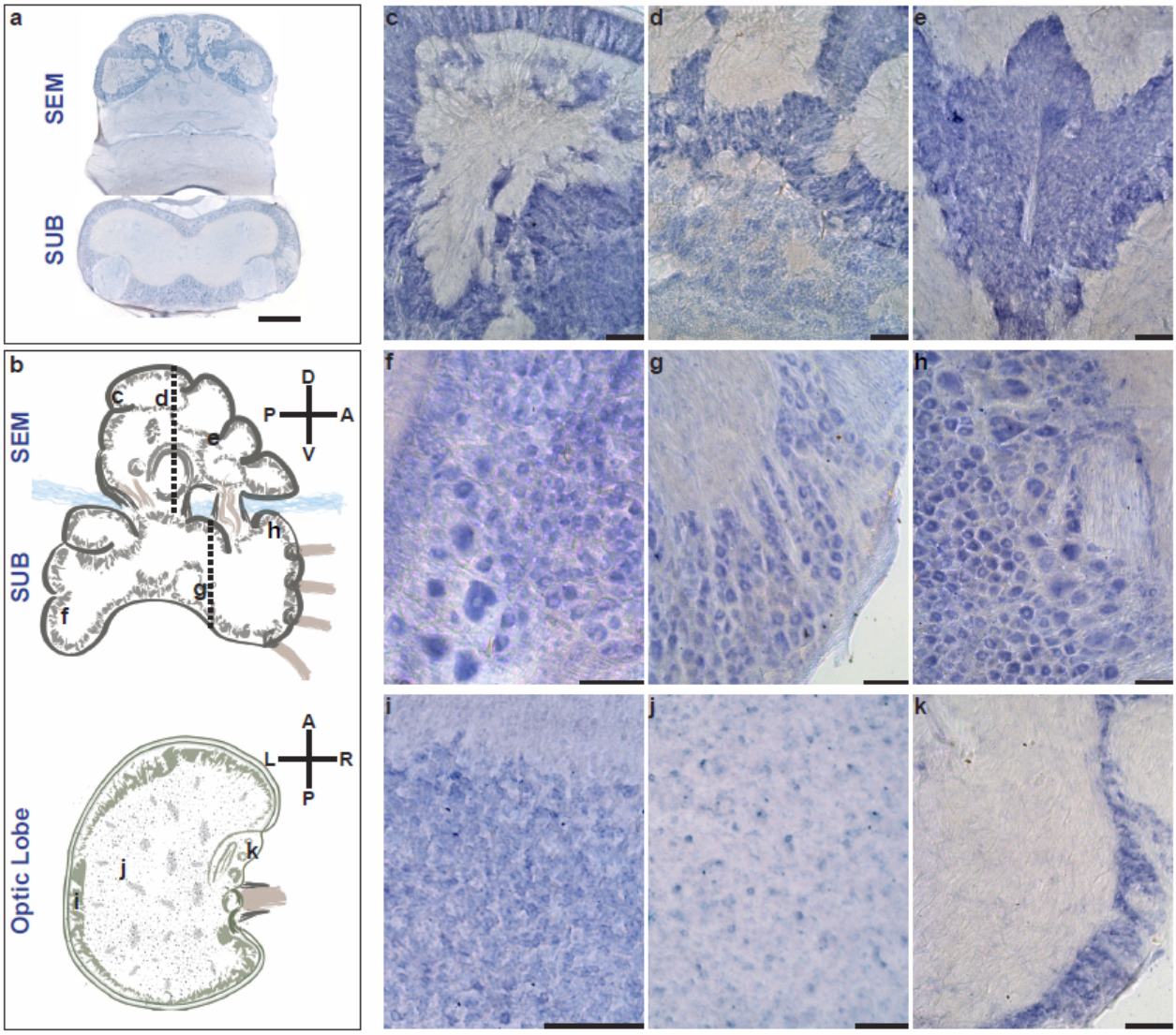

Figure 4. Localization of RTE-2_OV mRNA by in situ hybridization. (A) Bright-field micrographs of coronal sections of the supraesophageal (SEM) and subesophageal (SUB) masses hybridized to digoxigenin-labelled LINE antisense (refer to the plane indicated by the dashed-lines in B). The five gyri of the 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.

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

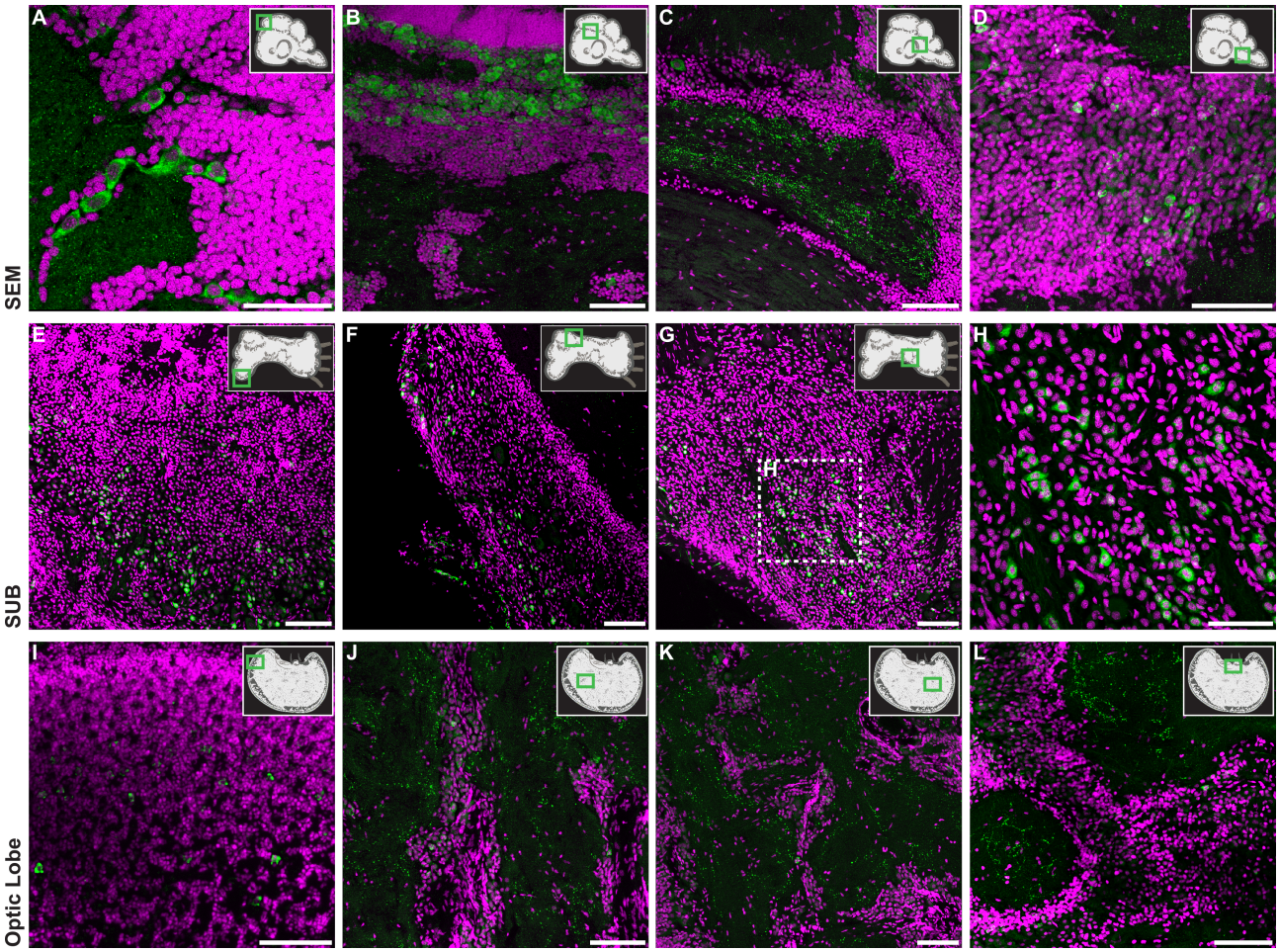

Figure 5. RTE-2_OV immunostaining in different areas of the brain. Coronal sections of the supra- (SEM; A-D) and sub-oesophageal (SUB; E-H) masses, and horizontal sections for the optic lobe (OL; I-L) following fluorescent-IHC (RTE-2_OV signal in green, DAPI used as nuclear stain in magenta) highlight 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).

330 # Discussion

331 The analysis of the coding part of the *O. bimaculoides* genome provided interesting parallelisms with the
332 transcriptional output of mammalian genomes [24]. Our results in *O. vulgaris* support and expand this view
333 suggesting that the molecular organization of the octopus neural transcriptome resembles the mammalian one
334 also for what concerns the transcription of thousands of lncRNAs and their TEs content.

335

336 Here we report the assembly of the most complete transcriptome for the *Octopus vulgaris*. The sequencing of
337 different portion of the organism allowed us to identify transcripts whose expression showed a peak in only
338 one part. For PCR validations we chose putative homologs of homeobox genes. *Arx* is an homeobox gene
339 associated with mental retardation. In mammalian it has been shown to be expressed in fetal brain at various
340 developmental stages, mainly in neuronal precursors in the telencephalon were it results involved in
341 corticogenesis and in the differentiation and maintenance of specific neuronal cell types [39, 40]. In our *O.*
342 *vulgaris* transcriptome assembly its putative homolog results to be expressed mainly in the SEM consistently
343 with an important function in the central brain also in this species. As another example, the expression of the
344 putative homolog of *Meox2* appears to be restricted to the arm in *O. vulgaris*, which also appears consistent
345 with the known function of this gene in vertebrates were it controls limb muscle development [41]. The
346 selection and validation of these expression peaking transcripts together with the functional information of
347 the putative homologs added support to the quality of the assembled transcriptome and allowed us to perform
348 more specific analyses. We observed that more than 10% of the assembled transcriptome for *O. vulgaris* and
349 *O. bimaculoides* is composed of putative lncRNAs and that more than 35% of the assembled transcripts
350 contain at least one fragment derived from a transposable element. Despite the fact that different classes of
351 TEs appear to be differentially segregated between coding and non-coding transcripts we did not find an
352 overall difference nor a strong enrichment for lncRNAs to contain embedded TE fragments with respect to
353 coding ones. This result was unexpected and warrant further investigation. One of the causes might rely in
354 our conservative strategy to call non-coding transcripts for which we might likely have underestimated the
355 fraction of lncRNAs increasing the rate of false negative.

356

357 The observation reported from us and others about TEs expansion in octopus drove our interest toward the
358 proportion of the transcriptome deriving from transposable elements. This lead us to the discovery of a novel
359 RTE LINE retrotransposon transcribed in the central nervous system of *O. vulgaris*. Reanalysis of the *O.*
360 *bimaculoides* data allowed us to identify, also in this species, a LINE member of the RTE class showing
361 evidences of activity. We could not find any full length copy of the *O. vulgaris* element in the genome survey
362 [32] nor could we find the *O. bimaculoides* element full length sequence in its reference assembly. This
363 likely resulted from the difficulties to assemble a reference genome using exclusively short reads in regions
364 containing long repeated sequences as those containing full-length LINE elements. Despite this technical
365 gap, the complementary use of transcriptomic, repeats reconstruction, annotation data and manual curation
366 allowed us to identify the full length autonomous elements likely to be active in both the octopus species.
367 The identification of potential integration polymorphisms of the RTE element in *O. bimaculoides* is a strong
368 sign of its activity. On the other hand, the results from quantitative PCR and Southern blot analysis relative
369 to the element discovered in *O. vulgaris*, although not conclusive, support the existence of somatic
370 retrotransposition in the central brain for this species. It is here important to consider that results from
371 quantitative PCR for LINEs genomic copy number variations should be interpreted with care, especially in
372 case of absence of additional supporting evidences. Indeed, in addition to a potential variation in genomic
373 copy number, they might also reflect a different amount of extra-chromosomal DNA which, at least in
374 human, is composed by a consistent fraction of LINEs DNA [12, 42]. More specific and careful inspections
375 are needed to validate, beyond any reasonable doubt, the activity of the elements presented in this work.
376 Nevertheless our results represent the first evidences of LINE activity conserved in the Octopus genus.
377
378 Through ISH we localized RTE-2_OV transcripts in specific groups of neural cells and areas of the *O.*
379 *vulgaris* nervous system, with no signal in neuropils. In particular, positive signals were observed at the level
380 of the majority of amacrine cells (about 3 micron diameter) which are suggested to have very limited neural
381 processes [43]. The most intriguing signals are those present at the level of the sub-frontal lobe and of the
382 five gyri of the vertical lobe (VL). These areas appeared to be the most intensely positive to RTE-2_OV
383 transcript found in the supra-esophageal mass. VL represents about 14% of the overall volume of the SEM in
384 the adult *O. vulgaris* [44] and counts about 25 million neural cells, equivalent to more than 65% of the total
385 number of nerve cells estimated for the SEM in the common octopus [22]. The largest quota of this

386 impressive number of closely packed nerve cells is represented by the amacrine cells, the smallest in the
387 octopus brain. Lying below the vertical lobe is the sub-vertical lobe. Its neural architecture is characterized
388 by the presence of a wall that in several regions is folded to form islands of cells. Many cells have a diameter
389 less than 5 μm, and very few are larger than 10 μm. At both sides of the sub-vertical lobe there are cells with
390 diameter of 5-10 μm, but also the largest neurons of the whole SEM reaching 25 μm. Through IHC in these
391 areas we identified a number of large cells positive to RTE-2_OV antibodies appearing organized in trunks
392 as typically described by Young in 1971 [45]. We also observed large positive cells organized in chains at the
393 level of the cellular wall of the sub-vertical lobe, and few positive neurons in the great majority of the islands
394 of cells (posterior wall of the sub-vertical lobe). Finally, immuno-reactivity was also evident in several
395 scattered amacrine cells belonging to the external granular layer of the OL suggested to correspond to the
396 neural cells described to project in the underlying plexiform zone [15].
397
398 Remarkably, *O. vulgaris* posterior frontal and vertical lobes belong to a neural circuit considered to be
399 functionally analogous to the mammalian hippocampus and limbic lobe [15, 17, 46, 47] where
400 retrotransposons are also reported to be active [7, 48]. The VL is also suggested to be analogous to the
401 mushroom bodies of insects [49]. All other octopus brain areas where the mRNA and/or the protein were
402 identified are centers of active neural plasticity for visual sensory-motor processing and memory (SEM and
403 OL), and motor processing (SUB) [17, 45, 50]. The positive cells overlap with the intricate neural network
404 made by amacrine cells known to act as amplifying matrices functioning as read-in read-out memory of both
405 sensory-motor visual and chemo-tactile processing [15, 47, 51, 52].
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 finger 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 [53]. Consequently, the ORF2 protein has been observed to be absent in a portion of cells
421 where the mRNA and ORF1 protein were present [54]. The limited efficiency for the translation of ORF2 has
422 been proposed to be one of the mechanisms that the cell puts in action to limit and/or regulate the
423 retrotranspositional potential of the L1 mRNAs. Indeed, for the retrotransposition to take place, all the L1
424 components, ORF1 protein, ORF2 protein and L1 mRNA must be present in the same ribonucleoprotein
425 complex (RBP) within the nucleus. The lack of any of these components in the RBP does not allow
426 retrotransposition. Reasoning along this line, another mechanism used to restrict retrotransposition is the
427 confinement of the RBP complex within the cytoplasm, where it is assembled. As long as the complex is
428 located in the cytoplasm it cannot access the genome and therefore cannot retrotranspose. For this reason,
429 cells subjected to retrotransposition have been identified mainly in cells which underwent duplication. The
430 mechanism proposed is that the RBP can get in contact with nuclear DNA only after the nuclear membrane
431 dissolves during mitosis. If the RBP remains in contact with the DNA it is then found into the nucleus after
432 the nuclear membrane of the new cell has been rebuilt thus allowing retrotransposition [54, 55].
433
434 Spatial uncoupling between L1 mRNA and proteins can also be observed independently from cellular
435 defense mechanisms. Indeed, L1 elements are also active at the mRNA level as lncRNAs. During early
436 development and in the regulation of cell potency, L1 transcription and its mRNA result to play a crucial role
437 in regulating global chromatin accessibility and the L1 mRNA can act as a nuclear RNA scaffold to recruit
438 specific regulatory proteins to the genome [10, 11]. Another molecular event resulting in spatial uncoupling
439 of LINEs transcription/translation can be observed when LINE mRNA is used from a given cell type to
440 target a different cell type with retrotranspositional activity. In fruitfly the existence of a mechanism
441 producing mRNAs of the I element (another member of the LINE class) by the nurse cells has recently been
442 observed. Once produced, these mRNAs are transported to the transcriptionally inactive oocytes were they
443 are translated and can cause retrotransposition (Wang et al. 2018).

Conclusions

Our data corroborate the hypothesis that LINE elements, potentially capable to determine genomic and functional diversity among individual cells, might be active and functionally important in the central nervous system of highly intelligent organisms such as octopus [6]. The localization of RTE-2_OV mRNA into different neural cells and not in the neuropil or muscle, complemented by the distinctive pattern of expression of its protein product and the identification of a potentially active element of the same class from a congeneric species, support the view that RTE-2_OV might be functional in neural centers in *O. vulgaris* and make it an important candidate to be further studied for its contribution to neural plasticity in this fascinating organism.

454 # Methods

455

456 ***Animals, sampling and ethical statement***

457 *Octopus vulgaris* were collected by local fishermen from the Bay of Naples (Southern Tyrrhenian Sea, Italy)
458 in early summer 2012. Animals were transported to the Stazione Zoologica Anton Dohrn in Napoli and
459 maintained according to a standardized acclimatization protocol [57–59]. Samples were taken from local
460 fishermen, by applying humane killing following principles detailed in Annex IV of Directive 2010/63/EU as
461 described in the Guidelines on the Care and Welfare of Cephalopods [59] and following protocol for
462 collection of tissues described by Baldascino and coworkers [60]. Death was confirmed by transection of
463 dorsal aorta. All dissections were carried out on seawater ice bed. During surgery optic lobes (OL), supra-
464 (SEM) and sub-esophageal (SUB) masses were dissected out from the animal, a piece of an arm (ARM),
465 usually the second left arm, was also taken. The complete dissection lasted less than 10 min. Sampling from
466 live animals occurred before the entry into force of the Directive 2010/63/EU in Member States and
467 therefore no legislation was in place in Italy regulating research involving cephalopods. However, the care
468 and welfare of animals has been consistent with best practice [59, 61, 62] and in compliance with the
469 requirements of the Directive 2010/63/EU that includes cephalopods within the list of species regulated for
470 scientific research involving living animals. In addition, animals killed solely for tissue removal do not
471 require authorization from the National Competent Authority under Directive 2010/63/EU and its
472 transposition into national legislation.

473

474 ***RNA extraction and sequencing***

475 For each octopus (N=3) total RNA was isolated from central nervous tissues (SEM, SUB and OL) and ARM,
476 a part of the body including the largest quota of the neuronal population belonging to the peripheral nervous
477 system [22, 63] and thus constituted by muscle and peripheral nervous tissue. SV total RNA isolation kit
478 (Promega, #Z3100) was utilized according to the manufacturer's protocol. DNA was degraded by treating
479 samples with Turbo DNase Kit (Ambion) according to the manual. Quality and quantity of RNA was
480 assessed by NanoDrop (Thermo-Fisher) and RNA BioAnalyzer (Agilent Technologies, Santa Clara, CA,
481 USA). Paired-end libraries were prepared using the Illumina TruSeq RNA sample library preparation kit

482 (Illumina, San Diego, CA, USA). Each sample was barcoded, samples pooled and sequenced in two lanes on
483 the Illumina HiSeq 2000 platform (paired-end, non-strand specific, 2x50bp read length protocol).
484
485 ***Raw reads quality filtering***
486 Quality of raw reads was assessed using FastQC (release 0.10.1). Raw reads were filtered and trimmed based
487 on quality and adapter inclusion using Trimmomatic [64] (release 0.22; parameters: -threads 24, -phred 64,
488 ILLUMINACLIP:illumina_adapters.fa:2:40:15, LEADING:3, TRAILING:3, SLIDINGWINDOW:3:20,
489 MINLEN:25). Read pairs with both reads passing the filters were considered for the transcriptome assembly.
490 Trimmed and filtered reads were normalized to remove duplicates using the normalize_by_kmer_coverage.pl
491 script from Trinity [65] (release r2013_08_14; parameters: --seqType fq, --JM 240G, --max_cov 30, --
492 JELLY_CPU 24).
493
494 ***De novo assembly and quantification of transcript abundances***
495 Transcriptome was assembled using Trinity (release r2013_08_14) on the trimmed, filtered and normalized
496 reads exploiting the Jaccard clip to limit assembly of chimeras. Assembly was performed using the following
497 parameters: --seqType fq, --JM 240G, --inchworm_cpu 24, --bflyHeapSpaceInit 24G, --bflyHeapSpaceMax
498 240G, --bflyCalculateCPU, --CPU 24, --jaccard_clip, --min_kmer_cov 2. To measure expression levels, raw
499 reads were mapped on the assembled transcriptome using Bowtie (version 1; parameters: -p 24, --chunkmbs
500 10240, --maxins 500, --trim5 2, --trim3 2, --seedlen 15, --tryhard -a -S). SAM outputs from Bowtie [66] were
501 converted into BAM, sorted, indexed and counted using the view, sort, index and idxstats programs from
502 Samtools [67]. All transcripts not showing at least 0.5 reads mapping per million mapped reads (CPM) in at
503 least 2 samples were discarded from the transcriptome as being expressed at too low levels and therefore
504 likely deriving by noise or assembly artifacts.
505
506 ***Annotation of the assembled transcriptome***
507 CEGMA (Core Eukaryotic Genes Mapping Approach; release 2.5) [68] was used to measure the
508 completeness of the assembled transcriptomes using the set of 248 Core Eukaryotic Genes (CEGs).
509 Transcripts annotation was performed using the Annocript pipeline [69] (release 0.2) with the combination of
510 tool, parameters and databases described below and using BLAST+ (release 2.2.27) [70]. To annotate

511 proteins we used BLASTX against the UniRef90 and Swiss-Prot databases from UniProt (release 2013_09)
512 [71] with parameters: -word_size 4, -evalue 10e-5, -num_descriptions 5, -num_alignments 5, -threshold 18.
513 To annotate protein domains we used RPSBLAST against the Conserved Domains Database (CDD v3.10)
514 [72] with parameters: -evalue 10e-5, -num_descriptions 20, -num_alignments 20). Ribosomal and small non-
515 coding RNAs were identified using BLASTN against a custom database made by Rfam (realease 11.0) [73]
516 and ribosomal RNA sequences from GenBank (parameters: -evalue = 10e-5, -num_descriptions 1, -
517 num_alignments 1). Each transcript was associated to Gene Ontology (GO) [74], Enzyme Commission (EC)
518 [75] and UniPathway [76] through cross-mapping of the best match from UniRef90 or Swiss-Prot using the
519 annotation mapping tables from UniProt. For each transcript, we used the Virtual Ribosome (Dna2pep
520 release 1.1) [77] to predict the length of the longest ORF searching across all reading frames without the
521 constraint to begin translation from a methionine start codon (parameters: -o none, -r all). The non-coding
522 potential for each transcript was calculated using Portrait (release 1.1) [78].
523
524 ***Non-coding annotation of the assembled transcriptome***
525 Putative lncRNAs were classified based on a heuristic process considering the annotation results. The
526 constraints used to identify potential lncRNAs have to be considered very stringent (Extended Data Fig. 2).
527 In published studies, different combinations of analyses have been used to identify lncRNAs [78–81]: 1) lack
528 of similarity with proteins, 2) lack of similarity with domain profiles, 3) lack of similarity with other RNAs
529 (ribosomal, snoRNA, etc.), 4) transcript and longest ORF lengths, 5) non-coding potential. We put all these
530 together and classified as lncRNA only those transcripts satisfying all the following conditions: a. length >=
531 200 nucleotides, b. lack of similarity with any of the following: protein from Swiss-Prot and UniRef90,
532 domains from CDD, rRNA from GenBank and other small ncRNA from Rfam, c. longest ORF < 100
533 aminoacids, d. non-coding potential score >= 0.95. Using these stringent constraints, we were able to predict
534 in the *O. vulgaris* transcriptome 7,806 (~12%) transcripts as putative lncRNAs.
535
536 ***Assembly, mapping and annotation of the Octopus bimaculoides public RNAseq data***
537 *O. bimaculoides* RNAseq raw data from Albertin et al. [24] were downloaded from NCBI SRA in October
538 2015 using the SRA Toolkit. Raw reads were filtered and trimmed based on quality and adapter inclusion
539 using Trimmomatic (release 0.33; parameters: -threads 32,

540 ILLUMINACLIP:illumina_adapters.fa:2:40:15:10:true LEADING:3 TRAILING:3
541 SLIDINGWINDOW:3:20 MINLEN:50). Read pairs with both reads passing the filters were considered for
542 the assembly. Trimmed and filtered reads were assembled with Trinity (release 2.1.0; parameters: --seqType
543 fq --SS_lib_type RF --CPU 32 --max_memory 240G --inchworm_cpu 32 --bflyHeapSpaceInit 24G --
544 bflyHeapSpaceMax 240G --bflyCalculateCPU --normalize_reads --min_kmer_cov 2 --jaccard_clip) using
545 digital normalization, strand information, the Jaccard clip and assuring that every kmer used in the assembly
546 was present in at least 2 reads to reduce noise. Redundancy of assembled transcripts was reduced using Cd-
547 hit [82] (version: 4.6, parameters: -c 0.90 -n 8 -r 0 -M 0 -T 0). To measure the expression levels, raw reads
548 were mapped on the transcriptome using Bowtie (version 1, parameters: -t -q -p 32 --chunkmbs 10240 --
549 maxins 500 --trim5 2 --trim3 2 --seedlen 28 --tryhard -a -S). SAM outputs from Bowtie were converted into
550 BAM, sorted, indexed and counted using the view, sort, index and idxstats programs from the Samtools
551 software collection. All transcripts not showing at least 1 reads mapping per million mapped reads (CPM) in
552 at least 1 sample were discarded from the transcriptome. *Octopus bimaculoides* genome was downloaded on
553 August 2015 from http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=Metazome.
554 Assembled and filtered unique transcripts were mapped on the genome using gmap [83] (version: 2015-09-
555 28, parameters: --suboptimal-score 0 -f gff3_gene --gff3-add-separators 0 -t 32 --min-trimmed-coverage 0.9
556 --min-identity 0.9) considering only transcripts aligning at least 90% of their length with 90% minimum
557 identity. We were able to map 84,043 (~90%) transcripts. Annotations and all the remaining analysis were
558 executed as for *O. vulgaris*.
559
560 ***Identification and classification of repetitive elements***
561 Repetitive elements for each transcriptome were annotated using RepeatMasker (A.F.A. Smit, R. Hubley &
562 P. Green RepeatMasker at http://repeatmasker.org; release 4.0.5) searching against the Repbase database [84]
563 (release 20140131) with parameters: -species bilateria, -s, -gff. We counted from RepeatMasker output the
564 repeat fragments present at least once in each transcript and built a matrix containing the percentage of
565 transcripts containing fragments related to a. retroelements, b. DNA transposons, c. satellites, d. simple-
566 repeats, e. low complexity, f. other, and g. unknown classes for each transcriptome according to the
567 RepeatMasker classification.
568

569 ***Identification of full-length, putatively active transposable elements***

570 We parsed RepeatMasker output calculating the percentage of overlap between the assembled transcripts and
571 the repeat consensus from Repbase selecting all elements with at least 80% coverage on the repeat
572 consensus. Elements showing the highest coverage were selected. On these we used Virtual Ribosome to
573 predict the longest complete ORFs by searching across all reading frames with methionine as start codon and
574 a canonical stop (parameters: -o strict, -r all). A single transcript resulted with a complete ORF. On this we
575 used InterPro [85] to identify and classify protein domains. The potential catalytic amino-acids essential for
576 the retrotransposition were manually identified comparing the putative translation with those reported in
577 Clements and Singer [37]. The same analysis was performed on both the transcriptomes of *O. bimaculoides*
578 (assembled by Albertin and assembled by us). The analysis was also performed on the RepeatMasker
579 annotation of the genome by Albertin downloaded from http://octopus.unit.oist.jp/OCTDATA/. For
580 consistency, we also analyzed RepeatMasker annotations of the genome and the transcriptomes produced
581 using the same tool, library and parameters used for *O. vulgaris* and the other species considered in this
582 study. In no-one of the analyses we could find a full-length transposable element retaining a complete ORF
583 for *O. bimaculoides*. We then translated the main RepeatScout and RepeatModeler repeat libraries
584 consensuses assembled by Albertin et al. (main RepeatScout library:
585 http://octopus.unit.oist.jp/OCTDATA/TE_FILES/mainrepeatlib.fa.gz; RepeatModeler library:
586 http://octopus.unit.oist.jp/OCTDATA/TE_FILES/oct.rm.tar.gz) with the Virtual Ribosome tool to predict
587 longest ORF searching across all reading frames showing methionine as start codon (parameters: -o strict, -r
588 all) and a canonical stop. The InterPro tool was then used to identify and classify the LINE characteristic
589 domains. The potential catalytic amino-acids essential for the retrotransposon activity were manually
590 identified comparing the ORF sequences with those reported in Clements and Singer. This led us to the
591 identification of two potentially functional LINE retrotransposons.

592

593 ***Identification of retrotrasposition events in O. bimaculoides***

594 The two candidates retrotransposons found in *O. bimaculoides* RepeatModeler libraries were analyzed to
595 search for integration sites in gonads and optic lobe using MELT [36] (v2.0.2). Two genomic DNAseq WGS
596 libraries from gonads (SRR2010220 and SRR2005790) were downloaded from the European Nucleotide
597 Archive (ENA) at https://www.ebi.ac.uk/ena/data/view/PRJNA270931. We generated two additional

598 DNAseq WGS libraries from DNA extracted from the optic lobe (L001 and L002) of a different individual.
599 *O. bimaculoides* reference genome was filtered for scaffold shorter than 10,000 bp and reads mapped on it
600 using BWA [86] (v0.7.15; parameters: mem, -t 32). SAM output from BWA were converted into BAM,
601 sorted, indexed and counted using the view, sort, index and idxstats programs from the SAMtools software.
602 The resulted sorted BAM files were used as input for MELT (parameters: -d 10000). Since the reads length
603 differ between the two set of libraries (150 bp for gonads and 260 bp for optic lobe) the optic lobe dataset
604 was trimmed with Trimmomatic (v0.32; parameters: CROP:200, HEADCROP:50) to obtain homogeneous
605 reads of 150 bp in all the datasets. We filtered the integration sites (ISs) identified by MELT for entries which
606 passed all the MELT checks and which presented at least 3 discordant pairs of reads supporting both left and
607 right sides of the breakpoints. BLAST (v2.6.0; parameters: -evalue 99999) search of the candidate
608 retrotransposons consensus sequences was performed against the genome and the identified ISs were
609 additionally filtered out when the BLAST search showed similarity in a range of 260 bp around the IS
610 breakpoint. The same analysis was performed using non-trimmed reads and two additional ISs identification
611 programs, RetroSeq [87] and an in-house developed pipeline, and the significance of results was maintained
612 (data not shown).
613
614 ***Phylogenetic tree generation***
615 Evolutionary tree in Figure 3D was generated using 100 full-length LINEs belonging to 15 LINE clades
616 (Extended Data Table 5). Protein sequences were selected from Ohshima and Okada [38] and manually
617 checked. InterPro was used to identify endonuclease and reverse transcriptase domains in all the LINEs.
618 Multiple sequence alignments were performed using MAFFT [88] (v7.221; with option L-INS-i). We utilized
619 TrimAl [89] (v1.4.rev15) to perform automated trimming aligned sequences (parameters: -fasta -
620 automated1). Phylogenetic relationships between LINE elements was reconstructed with MrBayes [90]
621 (v3.2.1). Bayesian analysis was run for six million generations with twenty-two chains, sampling every 1,000
622 generations (6,000 samples). Convergence was attained with standard deviation of split frequencies below
623 0.01 and a consensus tree was generated using a burnin parameter of 1,500 (25% of 6,000 samples). The
624 phylogenetic tree was visualized with FigTree program (release 1.4.2;
625 http://tree.bio.ed.ac.uk/software/figtree/).

626

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

629 We classified subsets of transcripts according to their expression levels across the different parts. A transcript
630 was considered expressed in a specific part if, in all the replicates of that part, it showed an expression level
631 > 0.5 CPM. This resulted in the classification of the groups of transcripts used to perform the analysis (see:
632 Figure 1c, Extended Data Fig. 4). We also classified transcripts having a peak of expression. These represent
633 transcripts showing an expression level > 0.5 CPM in all three biological replicates of exclusively one part
634 and below 0.5 in the others, which resulted in ~1,800 transcripts. They were used to draw the heatmap
635 (Extended Data Fig. 3a) using MeV (release 4.8.1) as part of the TM4 suite [91] with hierarchical clustering
636 exploiting Pearson correlation. Candidates for validations were selected among the 1,800 transcripts with
637 expression peak using the following additional criteria. For coding transcripts, we randomly selected four
638 coding transcripts representing octopus homologs of homeobox genes. The annotation was manually
639 verified. For lncRNAs we randomly selected four putative non-coding containing a SINE fragment. Both
640 coding and lncRNA candidates were validated using RT-PCR and Sanger sequencing.

641

642 ***Polymerase chain reaction (PCR)***

643 Octopus cDNAs were generated from 200 ng of total RNA using Superscript VILO cDNA Synthesis Kit
644 (Life Technologies) in 20 µl reaction volume. PCR were carried out using 20 ng of cDNA, 0.25 µl of Taq
645 DNA Polymerase (5U/µl; Roche), 1 µl of each specific forward and reverse primer (25pmol/µl), 2.5 µl of
646 PCR reaction buffer (10x), 2.5 µl of dNTP mix (10x) and water (up to final volume: 25 µl). The Ubiquitin
647 gene (accession number FJ617440) was used as internal control. Reactions for the coding transcripts were
648 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
649 7 min for 72°C. Reactions for the non-coding transcripts were amplified at annealing temperature of 58°C.
650 The following primers were utilized:

651 Arx forward 5'-TCCCTGCCTTCTCAACACAT-3'

652 Arx reverse 5'-TCCGAACTTCCACGCTTACT-3'

653 Hoxb5a forward 5'-GTGGCGAGGAATTTAGGAAG-3'

654 Hoxb5a reverse 5'-GCAACAGTCATAGTCCGAACAG-3'

655 Phox2b forward 5'-AATGGGGTGAGATCCTTTCC-3'

656 Phox2b reverse 5'-TTCATTGCAATCTCCTCTCG-3'

657 Meox2 forward 5'-TCCAGAACCGTCGGATGAAA-3'

658 Meox2 reverse 5'-TACGTAAAGGGCACACACCT-3'

659 Seml forward 5'-CACTTGTGCAAGGTACCACG-3'

660 Seml reverse 5'-AGGTCTCCTTAAATTTATTTCTGTGCA-3'

661 Subl forward 5'-ACAGAGCATCTTGAGTCTCACT-3'

662 Subl reverse 5'-CACTCCTGCGCCTTTCATTT-3'

663 Oll forward 5'-GGATTGACCCTGCAACTTGG-3'

664 Oll reverse 5'-CAGTGATGACGGACTTGCAA-3'

665 Arml forward 5'-GTACCCCACAAAATTAAATC-3'

666 Arml reverse 5'-CACTCACAAGGCTTTAGTTGGC-3'

667 Ubi forward 5'-TGTCAAGGCAAAGATTCAAGA-3'

668 Ubi reverse 5'-GGCCATAAACACACCAGCTC-3'

669

670 ***LINE copy number variation using quantitative real-time PCR***

671 Copy number variation analysis was performed on genomic DNA extracted from octopuses (N=9; SEM,
672 SUB, OL and ARM). One ARM sample was chosen as calibrator, while 18S was chosen as invariant control.
673 Purified genomic DNA concentrations were assessed by NanoDROP (ThermoFisher Scientific). According
674 to the starting concentration, DNA samples were diluted in TE buffer (10mM Tris-HCl, 1mM EDTA, pH 7.5)
675 to a concentration of 100 ng/µL and then further diluted to a concentration of 10 ng/µL. All dilutions were
676 checked by NanoDROP. Primers and multiplexing efficiencies were verified by linear regression to a
677 standard curve ranging from 50 ng to 16 pg of genomic DNA. LINE and 18S slopes were -3.3 and 3.8
678 respectively, and represent acceptable amplification efficiencies. Standard curves also confirmed that the
679 final concentration of 5ng DNA tested in qPCR was within the linear range of reaction. Reactions were
680 performed in 20 µL reaction mixture containing iQ Multiplex Powermix (Bio-Rad), Taqman primers (10
681 µM) and probes (10 µM) differentially labeled (with FAM or VIC fluorophore) and specifically designed to
682 hybridize with the target DNA sequences. LINE element was amplified using the following primers:

683 LINE forward 5'-AGCAGTGGGAATCATTCA-3'

684 LINE reverse 5'-GTCGTTTTCGTCGAACCAGT-3'

685 18S was amplified using the following primers:

686 18S forward 5'-AGTTCCGACCGTAAACGATG-3'

687 18S reverse 5'-CCCTTCCGTCAATTCCTTTA-3'

688 As probe sequences we utilized:

689 FAM: 5'-AACTCTGGGCCAAATTACGA-3'

690 VIC: 5'-GGGAAACCATAGTCGGTTCC-3'

691 qPCR was carried out for 20s at 90ºC, followed by 40 cycles of 10s at 95ºC and 30s at 59ºC using 7900HT
692 Fast Real Time PCR System (Applied Biosystems). Assays were performed for each sample in duplicate and
693 reproduced four times. Data obtained from the co-amplifications of the target DNA sequence and the internal
694 invariable control 18S were analyzed using the 2–ΔΔCt method79.

695

696 ***Southern Blotting***

697 The brains (SEM, SUB and OL) and a piece of an arm (ARM) of three *O. vulgaris* were dissected after
698 humane-killing and immediately frozen in liquid nitrogen. Pulverized samples were treated following the
699 methods utilized by Perelman and coworkers80; in brief, after phenol：chloroform (50:50) extraction DNA
700 was precipitated using cold isopropanol followed by centrifugation, suspended in TE buffer (10 mM Tris–
701 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
702 30 min. DNA concentration was estimated using NanoDrop and quality checked by electrophoresis on 0.8%
703 agarose gel. 10 µg genomic DNA for each tissue was digested with EcoRI (New England Biolabs) overnight
704 at 37°C and resolved on a 0.9 % agarose gel for 15 hours at 1.5V. DNA was transferred to a Hybond-N+
705 nylon membrane (0.45 µm; Amersham Pharmacia Biotech) according to Sambrook and Russell81. DIG-
706 labeled LINE DNA probe was prepared by PCR DIG Probe synthesis kit (Roche). Hybridization and
707 autoradiography were performed according to the DIG Application Manual (Roche).

708

709 ***Probe synthesis for in-situ hybridization***

710 We amplified by PCR a 356-bp cDNA fragment of the assembled LINE from bp 1,512 to bp 1,868 using the
711 following primers:

712 LINE forward 5'-GGGTCAGAAAGTGACGAGGA-3'

713 LINE reverse 5'-TGCATGAGGCGGAGTTTAG-3'

714 The choice of the fragment and the design of primers have been based on manual curation steps ensuring that
715 the chosen fragment is present exclusively in the transcript of the identified LINE element and in no other
716 assembled transcripts. The amplified fragment was cloned into TOPO® TA Cloning® vector (Life
717 Technologies, CA, USA) according to manufacturer's protocol. Cloned fragment was digested using BAMHI
718 and ECORV restriction enzymes and validated by Sanger sequencing. Sense and antisense digoxigenin-
719 labelled RNA probes were generated by in vitro transcription using the DIG-RNA Labeling Kit (SP6/T7;
720 Roche Applied Sciences, QC, Canada). Labeled RNA probes were quantified by dot blot analysis.

721

722 ***In-situ hybridization experiments***

723 Brain masses (SEM, SUB, OL) and a segment of an arm from octopuses (N = 3) were fixed in
724 paraformaldehyde 4% (PFA) in phosphate buffered saline (PBS) at 4°C (3h for brain masses; overnight,
725 ARM). Samples were washed (four rinses in PBS), dehydrated in series of graded methanol/PBS (1:3, 1:1,
726 3:1 v/v) and stored at least one night in methanol (–20°C). Tissues were then rehydrated at 25°C in a series
727 of graded methanol/PBS (3:1, 1:1, 1:3 v/v) solutions and cryoprotected in 30% sucrose in PBS. After sucrose
728 infiltration, samples were embedded in tissue freezing medium (OCT; Leica Biosystems) and sectioned using
729 a cryostat (Leica CM3050 S). Sagittal and/or coronal sections (40 µm) were collected in PBST (phosphate
730 buffered saline including 0.1% Tween™ 20 and 0.2mM sodium azide). Washed free-floating sections were
731 mounted on Superfrost Plus slides (Menzel Gläser) and let dry overnight under fume hood. Hybridizations
732 were performed as described by Abler et al.82 with modifications. After rehydration in PSBT, the sections
733 were quenched at 25°C in 6% H_2O_2 (30 min) treated with proteinase K (10 min) and post-fixed with PFA-G
734 (4% paraformaldehyde and 0.2% glutaraldehyde) for 20 min. Prehybridization step was performed in
735 hybridization solution (HB: 50% formamide, 5XSSC, with 10µg/mL heparin, 10µg/mL yeast tRNA and, 1%
736 Blocking reagent) for at least 1h at 60°C and then incubated overnight in HB with the digoxigenin-labeled
737 riboprobes. Post-hybridization washes (50% formamide, 5x SSC, 1% SDS) were carried out for 2 hours at
738 60°C. Sections were washed in TNT (10mM Tris-HCL pH 7.5, 0.5M NaCl, 0.1% Tween™ 20) at 25°C and
739 incubated for 15min at 37°C with RNase (0.25µg/mL), followed by a FS (50% formamide, 5x SSC)
740 incubation of 2 hours (60°C). DIG was detected with an alkaline phosphatase labeled antibody (Roche).

741 After a saturation step in TBS pH 7.5 (10% sheep serum, 1% blocking reagent, 1% BSA, 0.1% Tween™ 20)
742 for 1h (room temperature), sections were incubated overnight at 4°C with antibodies (1:1,000; in TBS
743 containing: 5% sheep serum, 1% blocking reagent, 1% BSA). The following day, sections were washed for
744 2h in TBS (pH 7.5; 0.1% Tween™ 20 and 2mM levamisole) and then washed in alkaline phosphatase
745 solution (100mM Tris-HCL pH 9.5, 100mM NaCl, 50mM MgCl2, 0.1% Tween™ 20 and 2mM levamisole).
746 Bound antibodies were revealed using NBT-BCIP (Roche). After DIG in situ hybridization, slides were
747 counterstained with DAPI (5 µg/mL, Invitrogen) washed and mounted using aqueous mounting.
748
749 ***RTE-2_OV custom-made polyclonal antibodies***
750 Custom-made polyclonal antibodies were obtained from Primm Biotech Custom Antibody Services (Milan,
751 Italy) and raised against two peptides derived from two portions of the RTE-2_OV protein: GAA (1-100 aa),
752 and HAA (569-673 aa) resulting to be unique within the translation of the assembled transcriptome. To
753 choose the portion to select, the manufacturer also took into consideration: protein similarity (selection of
754 regions with no significant identity to the murine and rabbit proteome); low complexity and transmembrane
755 regions (exclusion of such regions); distribution of predicted antigenic peptides (selection of regions with a
756 high number of predicted antigenic peptide). The selected synthetic peptides were injected into two rabbits
757 and boosted three times within 38 days (at day: 21, 28, 35) after the first injection. The final bleeding was
758 conducted 3 days after the last injection, and the crude sera were purified on Sepharose columns by
759 immunoaffinity.
760
761 ***Immunohistochemistry and antibody validation***
762 SEM, SUB and OL dissected from *O. vulgaris* (N = 3) were immediately immersed 4% paraformaldehyde
763 (PFA) in seawater (4°C for 3h). After fixation samples were washed several times in PBS, cryoprotected in
764 sucrose 30% and embedded in OCT compound (OCT; Leica Biosystems). The embedded brain parts were
765 then sectioned at 20 µm using a cryostat (Leica CM3050 S). No antigen retrieval was required. Tissue
766 sections were rehydrated in three successive baths of 0.1 M PBS, and incubated for 90 min (at RT) in 5%
767 goat serum (Vector Laboratories Ltd) diluted in 0.1 M PBS containing 0.05% Tween (PBTw). The slices
768 were subsequently incubated at 4°C with custom-made polyclonal antibodies raised against LINE (G and H;
769 see RTE-2_OV custom-made polyclonal antibodies for details). The next day, slices were again washed by

770 several changes of PBTw and incubation (at RT for 90 min) with secondary antibodies was carried out using
771 Alexa Fluor 488 or 546 goat anti-rabbit IgG both diluted 1:200 in PBTw. Subsequently, sections were rinsed
772 and cell nuclei were counterstained with DAPI (Molecular Probes, Eugene, OR). Finally, after further
773 extensive washes, sections were mounted with fluorescent mounting medium (Fluoromount, Sigma). For all
774 antisera tested, omission of the primary antiserum and/or secondary antiserum resulted in negative staining.
775 In addition, specificity was assured by pre-incubating (4°C, overnight) the antibodies with 1mg/mL of
776 synthetic epitope (HAA and GAA, see RTE-2_OV custom-made polyclonal antibodies for details) before
777 staining. Again, no immunostaining was observed. The two custom-made polyclonal antibodies raised
778 against two different peptides of the RTE-2_OV protein stained the same spatial arrangement in the octopus
779 brain tissue.
780
781 ***Imaging***
782 Sections were observed under microscopes depending on the techniqueImage acquisition and processing
783 were performed using Leica Application Suite software (Leica Microsystems). For IHS we utilized a Leica
784 DMI6000 B inverted microscope, and for IHC a Leica TCS SP8 X confocal microscope (Leica
785 Microsystems, Germany). Tile Z-stacks were performed using a 0.2µm step size. IHC figures have been
786 assembled following guidelines for color blindness provided by Wong [92].
787

788 # List of abbreviations

789 TEs Transposable elements
790 LINE Long interspersed element
791 SINE Short interspersed element
792 LTR Long terminal repeat
793 LncRNAs Long non-coding RNAs
794 RTE Retrotransposon element
795 SEM Supra-esophageal mass
796 SUB Sub-esophageal mass
797 OL Optic lobe

ARM	Arm
GO	Gene ontology
CPM	Counts per million
RT-PCR	Reverse transcriptase-polymerase chain reaction
qPCR	Quantitative polymerase chain reaction
MELT	Mobile element locator tool
ISH	In-situ hybridization
IHC	Immunohistochemistry
RBP	Ribonucleoprotein

Declarations

Ethics approval and consent to participate

Samples were taken from local fishermen, by applying humane killing following principles detailed in Annex IV of Directive 2010/63/EU as described in the Guidelines on the Care and Welfare of Cephalopods (Fiorito et al. 2015) and following protocol for collection of tissues described by Baldascino and coworkers (Baldascino et al. 2017).

Consent for publication

Not applicable.

Availability of data and materials

Octopus vulgaris RNAseq raw reads have been submitted to the ArrayExpress database (www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-3957. The data will be made public upon acceptance of this article. Login credential for referees are indicated in the first page of the supplementary pdf. All the remaining data are present in the supplementary files archive.

Competing interests

The authors declare that they have no competing interests.

Funding

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

826 (MIUR and European Regional Development Fund, FESR). Giuseppe Petrosino, Swaraj Basu, Massimiliano
827 Volpe and Giulia Di Cristina have been supported by a SZN PhD fellowship.

828 ***Authors Contributions***

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

837 ***Acknowledgements***

838 The authors would like to thank Paul Andrews for critical reading of the manuscript; Paolo Vatta and Marta
839 Maurutto for assistance in CNV experiments; Marco Miralto and the BIOINforMA service at SZN for
840 informatics support.

841 ***Authors' information: Current addresses***

842 Giuseppe Petrosino: Institute of Molecular Biology (IMB), Mainz, Germany.

843 Francesco Musacchia: Central RNA Laboratory, Istituto Italiano di Tecnologia (IIT), Genova, Italy.

844 Swaray Basu: Strand Life Sciences, Bengaluru, India.

845 Giulia Di Cristina: Institute of Zoology, University of Cologne, Cologne, Germany.

846 Dinko Pavlinic: Institute of Molecular and Clinical Ophthalmology, Basel, Switzerland.

847 Massimiliano Volpe: Department of Biomedical and Clinical Sciences, Linköping University, Linköping,
848 Sweden.

849 Oleg Simakov: Department of Molecular Evolution and Development, Wien University, Wien, Austria.

850 # References

851

1. Kazazian HH. Mobile Elements: Drivers of Genome Evolution. Science. 2004;303:1626–32.

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

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

4. Muotri AR, Chu VT, Marchetto MCN, Deng W, Moran JV, Gage FH. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature. 2005;435:903–10.

5. Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, et al. Somatic retrotransposition alters the genetic landscape of the human brain. Nature. 2011;479:534–7.

6. Singer T, McConnell MJ, Marchetto MCN, Coufal NG, Gage FH. LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes? Trends Neurosci. 2010;33:345–54.

7. Perrat PN, DasGupta S, Wang J, Theurkauf W, Weng Z, Rosbash M, et al. Transposition-Driven Genomic Heterogeneity in the Drosophila Brain. Science. 2013;340:91–5.

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

9. Lee EJ, Banerjee S, Zhou H, Jammalamadaka A, Arcila M, Manjunath BS, et al. Identification of piRNAs in the central nervous system. RNA. 2011;17:1090–9.

10. Jachowicz JW, Bing X, Pontabry J, Bošković A, Rando OJ, Torres-Padilla M-E. LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. Nat Genet. 2017;49:1502–10.

11. Percharde M, Lin C-J, Yin Y, Guan J, Peixoto GA, Bulut-Karslioglu A, et al. A LINE1-Nucleolin Partnership Regulates Early Development and ESC Identity. Cell. 2018;174:391-405.e19.

12. Thomas CA, Tejwani L, Trujillo CA, Negraes PD, Herai RH, Mesci P, et al. Modeling of TREX1-Dependent Autoimmune Disease using Human Stem Cells Highlights L1 Accumulation as a Source of Neuroinflammation. Cell Stem Cell. 2017;21:319-331.e8.

13. Shomrat T, Zarrella I, Fiorito G, Hochner B. The octopus vertical lobe modulates short-term learning rate and uses LTP to acquire long-term memory. Curr Biol. 2008;18:337–42.

14. Edelman DB, Seth AK. Animal consciousness: a synthetic approach. Trends Neurosci. 2009;32:476–84.

15. Young JZ. Computation in the learning system of cephalopods. Biol Bull. 1991;180:200–8.

16. Turchetti-Maia A, Shomrat T, Hochner B. The Vertical Lobe of Cephalopods. A Brain Structure Ideal for Exploring the Mechanisms of Complex Forms of Learning and Memory. In: Byrne JJ, editor. The Oxford Handbook of Invertebrate Neurobiology. Oxford, UK: Oxford University Press; 2017. p. 1–27.

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

18. Styfhals R, Seuntjens E, Simakov O, Sanges R, Fiorito G. In silico Identification and Expression of Protocadherin Gene Family in *Octopus vulgaris*. Front Physiol. 2019;9. doi:10.3389/fphys.2018.01905.

19. Shigeno S, Ragsdale CW. The gyri of the octopus vertical lobe have distinct neurochemical identities. J Comp Neurol. 2015;523:1297–317.

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

21. Wang ZY, Ragsdale CW. Cadherin genes and evolutionary novelties in the octopus. Seminars in Cell & Developmental Biology. 2017;69:151–7.

22. Young JZ. The Number and Sizes of Nerve Cells in Octopus. Proceedings of the Zoological Society of London. 1963;140:229–54.

23. Borrelli L, Fiorito G. 1.31 - Behavioral Analysis of Learning and Memory in Cephalopods. In: Byrne JH, editor. Learning and Memory: A Comprehensive Reference. Oxford: Academic Press; 2008. p. 605–27. doi:10.1016/B978-012370509-9.00069-3.

24. Albertin CB, Simakov O, Mitros T, Wang ZY, Pungor JR, Edsinger-Gonzales E, et al. The octopus genome and the evolution of cephalopod neural and morphological novelties. Nature. 2015;524:220–4.

25. Guan Z, Cai T, Liu Z, Dou Y, Hu X, Zhang P, et al. Origin of the Reflectin Gene and Hierarchical Assembly of Its Protein. Current Biology. 2017;27:2833-2842.e6.

26. Ritschard Elena A., Fitak Robert R., Simakov Oleg, Johnsen Sönke. 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. 2019;286:20182929.

27. Lee PN, Callaerts P, de Couet HG, Martindale MQ. Cephalopod Hox genes and the origin of morphological novelties. Nature. 2003;424:1061–5.

28. Liscovitch-Brauer N, Alon S, Porath HT, Elstein B, Unger R, Ziv T, et al. Trade-off between Transcriptome Plasticity and Genome Evolution in Cephalopods. Cell. 2017;169:191-202.e11.

29. Kim B-M, Kang S, Ahn D-H, Jung S-H, Rhee H, Yoo JS, et al. The genome of common long-arm octopus Octopus minor. GigaScience. 2018;7. doi:10.1093/gigascience/giy119.

30. Belcaid M, Casaburi G, McAnulty SJ, Schmidbaur H, Suria AM, Moriano-Gutierrez S, et al. Symbiotic organs shaped by distinct modes of genome evolution in cephalopods. PNAS. 2019;116:3030–5.

31. da Fonseca RR, Couto A, Machado AM, Brejova B, Albertin CB, Silva F, et al. A draft genome sequence of the elusive giant squid, Architeuthis dux. GigaScience. 2020;9. doi:10.1093/gigascience/giz152.

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

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

34. Malik HS, Eickbush TH. The RTE class of non-LTR retrotransposons is widely distributed in animals and is the origin of many SINEs. Mol Biol Evol. 1998;15:1123–34.

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

36. Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, et al. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Res. 2017;27:1916–29.

37. Clements AP, Singer MF. The human LINE-1 reverse transcriptase:effect of deletions outside the common reverse transcriptase domain. Nucleic Acids Res. 1998;26:3528–35.

38. Ohshima K, Okada N. SINEs and LINEs: symbionts of eukaryotic genomes with a common tail. Cytogenetic and Genome Research. 2005;110:475–90.

39. Ohira R, Zhang Y-H, Guo W, Dipple K, Shih SL, Doerr J, et al. Human ARX gene: genomic characterization and expression. Molecular Genetics and Metabolism. 2002;77:179–88.

40. Friocourt G, Poirier K, Rakić S, Parnavelas JG, Chelly J. The role of ARX in cortical development. Eur J Neurosci. 2006;23:869–76.

41. Mankoo BS, Collins NS, Ashby P, Grigorieva E, Pevny LH, Candia A, et al. Mox2 is a component of the genetic hierarchy controlling limb muscle development. Nature. 1999;400:69–73.

42. De Cecco M, Ito T, Petrashen AP, Elias AE, Skvir NJ, Criscione SW, et al. L1 drives IFN in senescent cells and promotes age-associated inflammation. Nature. 2019;566:73–8.

43. Gray EG. The fine structure of the vertical lobe of octopus brain. Philos Trans R Soc Lond, B, Biol Sci. 1970;258:379–94.

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

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

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

47. Young JZ. Multiple matrices in the memory system of Octopus. In: Abbott JN, Williamson R, Maddock L, editors. Cephalopod Neurobiology. Oxford: Oxford University Press; 1995. p. 431–43.

48. Upton KR, Gerhardt DJ, Jesuadian JS, Richardson SR, Sánchez-Luque FJ, Bodea GO, et al. Ubiquitous L1 mosaicism in hippocampal neurons. Cell. 2015;161:228–39.

49. Hochner B. Functional and comparative assessments of the octopus learning and memory system. Front Biosci (Schol Ed). 2010;2:764–71.

50. Marini G, De Sio F, Ponte G, Fiorito G. Behavioral Analysis of Learning and Memory in Cephalopods. In: Byrne JH, editor. Learning and Memory: A Comprehensive Reference (Second Edition). Amsterdam, The Netherlands: Academic Press, Elsevier; 2017. p. 441–62.

51. Maldonado H. The visual attack learning system in *Octopus vulgaris*. J Theor Biol. 1963;5:470–88.

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

53. Mita P, Sun X, Fenyö D, Kahler DJ, Li D, Agmon N, et al. BRCA1 and S phase DNA repair pathways restrict LINE-1 retrotransposition in human cells. Nature Structural & Molecular Biology. 2020;27:179–91.

54. Mita P, Wudzinska A, Sun X, Andrade J, Nayak S, Kahler DJ, et al. LINE-1 protein localization and functional dynamics during the cell cycle. eLife. 2018;7:e30058.

55. Taylor MS, Altukhov I, Molloy KR, Mita P, Jiang H, Adney EM, et al. Dissection of affinity captured LINE-1 macromolecular complexes. eLife. 2018;7:e30094.

56. Wang L, Dou K, Moon S, Tan FJ, Zhang ZZ. Hijacking Oogenesis Enables Massive Propagation of LINE and Retroviral Transposons. Cell. 2018;174:1082–94.

57. Fiorito G, von Planta C, Scotto P. Problem solving ability of *Octopus vulgaris* Lamarck (Mollusca, Cephalopoda). Behavioral and Neural Biology. 1990;53:217–30.

58. Amodio P, Andrews P, Salemme M, Ponte G, Fiorito G. The use of artificial crabs for testing predatory behavior and health in the octopus. ALTEX - Alternatives to animal experimentation. 2014;31:494–9.

59. Fiorito G, Affuso A, Basil J, Cole A, de Girolamo P, D'Angelo L, et al. 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. 2015;49 2 Suppl:1–90.

60. Baldascino E, Di Cristina G, Tedesco P, Hobbs C, Shaw TJ, Ponte G, et al. 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. 2017;8. doi:10.3389/fphys.2017.01001.

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

62. Fiorito G, Affuso A, Anderson DB, Basil J, Bonnaud L, Botta G, et al. Cephalopods in neuroscience: regulations, research and the 3Rs. Invert Neurosci. 2014;14:13–36.

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

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

65. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protocols. 2013;8:1494–512.

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

67. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

68. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;23:1061–7.

69. Musacchia F, Basu S, Petrosino G, Salvemini M, Sanges R. Annocript: a flexible pipeline for the annotation of transcriptomes able to identify putative long noncoding RNAs. Bioinformatics. 2015;31:2199–201.

70. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.

71. Consortium TU. UniProt: a hub for protein information. Nucl Acids Res. 2015;43:D204–12.

72. Marchler-Bauer A, Zheng C, Chitsaz F, Derbyshire MK, Geer LY, Geer RC, et al. CDD: conserved domains and protein three-dimensional structure. Nucleic Acids Res. 2013;41 Database issue:D348-352.

73. Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, et al. Rfam 11.0: 10 years of RNA families. Nucleic Acids Res. 2013;41 Database issue:D226-232.

74. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25:25–9.

75. Bairoch A. The ENZYME database in 2000. Nucleic Acids Res. 2000;28:304–5.

76. Morgat A, Coissac E, Coudert E, Axelsen KB, Keller G, Bairoch A, et al. UniPathway: a resource for the exploration and annotation of metabolic pathways. Nucleic Acids Res. 2012;40 Database issue:D761-769.

77. Wernersson R. Virtual Ribosome--a comprehensive DNA translation tool with support for integration of sequence feature annotation. Nucleic Acids Res. 2006;34 Web Server issue:W385–8.

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

79. Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, et al. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res. 2012;22:577–91.

80. Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. Conserved Function of lincRNAs in Vertebrate Embryonic Development despite Rapid Sequence Evolution. Cell. 2011;147:1537–50.

81. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes & Development. 2011;25:1915–27.

82. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22:1658–9.

83. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;21:1859–75.

84. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110:462–7.

85. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, et al. InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res. 2012;40 Database issue:D306-312.

86. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.

87. Keane TM, Wong K, Adams DJ. RetroSeq: transposable element discovery from next-generation sequencing data. Bioinformatics. 2013;29:389–90.

88. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.

89. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–3.

90. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 2003;19:1572–4.

91. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, et al. TM4: a free, open-source system for microarray data management and analysis. BioTechniques. 2003;34:374–8.

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

852