UNIVERSITY OF NAPOLI FEDERICO II

Doctorate Program

in

Computational Biology and Bioinformatics

XXVII cycle

The transcriptional landscape of the nervous system of *Octopus vulgaris*

Giuseppe Petrosino

Stazione Zoologica Anton Dohrn

Naples, 2015

Director of studies: Dr Graziano Fiorito
Biology and Evolution of Marine Organisms
Stazione Zoologica Anton Dohrn
Naples, Italy

Internal supervisor: Dr Remo Sanges
Biology and Evolution of Marine Organisms
Stazione Zoologica Anton Dohrn
Naples, Italy

External supervisor: Dr Sandro Banfi
Telethon Institute of Genetics and Medicine (TIGEM)
Naples, Italy

Abstract

Cephalopoda, a class of the phylum Mollusca, has undergone dramatic evolutionary changes in the body plan and in the morphology of the nervous system. The complexity of the nervous system can be recognized from the brain size, the number of neuronal cells and the neuroanatomical organization and it is comparable to that of vertebrates. In the *Octopus vulgaris*, the nervous system reaches a great level of complexity both in the Central nervous system (CNS) and Peripheral nervous system (PNS). The aim of the my PhD is to generate the reference transcriptome of the nervous system to understand, at the molecular level, its organization, development and evolution, allowing comparative analysis across both the different areas as well as multiple animal phyla.

The *Octopus* transcriptome sequencing allowed me to show, for the first time, an unexpected high frequency of retroelements embedded in transcripts, much more similar to that shown in mammals and much higher than in other invertebrate and vertebrate non-mammalian species. Interestingly, Short INterspersed Elements (SINEs) are significantly more abundant in long noncoding RNAs (lncRNAs) than in protein coding genes, and both lncRNAs and SINEs are enriched in transcripts expressed in the central nervous system (CNS). Moreover, in this study I also identified a Long INterspersed Element (LINE) that is transcribed in the CNS and seems to be fully competent for the retrotransposition of itself and its partner SINE.

These observations suggest a possible convergent evolutionary scenario involving retroelements and lncRNAs pervasive expression in the *Octopus* nervous system, contributing to develop its cognitive abilities, unique among invertebrates.

Acknowledgements

I would like to express my sincere gratitude to my supervisors Dr. Graziano Fiorito and Dr. Remo Sanges for their patient guidance, constant support and motivation. I have been extremely lucky to have them as mentors who cared so much about my work, and who responded to my questions and queries so promptly.

I would like to thank Dr. Mariella Ferrante for giving me the opportunity to attend the bioinformatic internship at the National Center for Genome Resources (NCGR) in New Mexico (USA) where I spend a productive period.

It has been a great experience sharing office space, lab meetings, and countless lunches with Swaraj Basu, Francesco Musacchia and Veerendra Parsappa Gadekar. I would like to thank all of them for helping me with important comments and suggestions during the research program. They also introduced myself to the field of Bioinformatics, as Swaraj said, the dark site of Biology.

Furthermore, I am very grateful to Giovanna Ponte, Ilaria Zarrella and Pamela Imperadore for the feedback, direction and assistance when I needed it. I have learnt so much from all of those people about the Octopuses, sharing their incredible passion in this field.

To the other members of Fiorito Lab, Giulia Di Cristina, Giovanni De Martino, Elena Baldascino and Maria Grazia Lepore, thank you for the encouragement and countless coffee prepared during the last three years.

A special recognition goes out to my family, for their support throughout my life. Finally, I want to express my deepest love and thanks to my loving partner, Viviana. I couldn't imagine how would finish this thesis if it were not for her constant love and faith in me, taking care of me during some hard times. Thank you. This work is truly also yours!

Abbreviations

(alphabetical order)

3P	3'end
5P	5'end
An	poly(A)
ANRIL	antisense non-coding RNA in the INK4 locus
ARM	arm
ARX	aristaless-related homeobox
AS	antisense
ATP	ATPase
BACE-AS	b-site APP-cleaving enzyme 1-antisense
BANCR	BRAF-regulated lncRNA 1
BBB	blood-brain barrier
CEGMA	core eukaryotic genes mapping approach
CEGs	core eukaryotic genes
CNS	central nervous system
CPM	counts per million
EC	enzyme commission
EN	endonuclease
ERV	endogenous retroviral
ERVs	endogenous retroviruses
ESTs	expressed sequence tags
FDR	false discovery rate
GAG	specific group antigen
GO	gene ontology
H3K27me3	histone H3 lysine-27 trimethylation
Hel	helicase
hERV	human endogenous retroviruses
HMMs	hidden markov models
HOXB5a	homeobox B5a
HT	horizontal transfer
IN	integrase
Jarid2	jumonji, AT rich interactive domain 2
Lhx2	LIM homeobox protein 2
LINEs	long interspersed elements
lncRNAs	long non-coding RNAs
LTRs	long terminal repeats
M	medium
Malat1	metastasis-associated lung adenocarcinoma transcript 1
MBs	mushroom bodies
MECP2	methyl-CpG-binding protein 2
MEOX2	mesenchyme homeobox 2
MeV	multi-experiment viewer
MGEs	mobile genetic elements
miRNAs	microRNAs

Mya	million years ago
NFAT	nuclear factor of activated T cells
NGS	next generation sequencing
NK-kB	nuclear factor-kB
OL	optic lobe
ONS	octopus nervous transcriptome
ORFs	open reading frames
PC1	principal component 1
PC2	principal component 2
PCA	principal component analysis
PCR	polymerase chain reaction
PEC	paired-end complex
PHOX2B	paired-like homeobox 2b
piRNA	piwi-interacting RNA
PNS	peripheral nervous system
pol	polymerase
Pol II	DNA polymerase II
POLB	DNA polymerase B
PRC2	polycomb repressive complex 2
PRO	protease
RBP	RNA binding protein
Rep	replication initiator
RH	RNase H
RNA-seq	RNA-sequencing
RT	reverse transcriptase
RUNX3	runt-related transcription factor 3
S	sense
SEC	single-end complex
SEM	supraesophageal mass
SINEs	short interspersed elements
Six3	SIX homeobox 3
sncRNAs	small non-coding RNAs
snoRNAs	small nucleolar RNAs
SUB	subesophageal mass
SVM	support vector machine
Tes	transposable elements
TIR	terminal inverted repeats
TLR	Toll-like receptors
TP	terminal protein
TPRT	target primed reverse transcription
TRs	tandem repeats
TSD	target site duplication
TSS	transcription start site
TUG1	taurine up-regulated gene 1
Uchl1-as1	ubiquitin carboxy-terminal hydrolase L1 antisense RNA 1
UTR	untranslated region

VL	vertical lobe
YY1	yin yang 1
Znf	zing-finger

Table of Contents

ABSTRACT I

ACKNOWLEDGEMENTS II

ABBREVIATIONS IV

CHAPTER 1 – AIMS AND INTRODUCTION TO THE *OCTOPUS VULGARIS* 1

1.1 Prologue and plan of this PhD Project 1

1.2 Why Cephalopods 2

1.3 Cephalopods as Bilaterians 4

1.4 Main features of Cephalopods 7

1.5 Octopus' central nervous system 13

1.6 Octopus' peripheral nervous system: arms 19

1.7 Comparison of the *Octopus* and the *Homo* Nervous System 22

CHAPTER 2 – REPETITIVE ELEMENTS WITHIN THE GENOME 26

2.1 Repetitive Elements 26
2.1.1 Classification and structure 26
2.1.2 Mobilization Mechanism of Class I transposable elements 29
2.1.3 Mobilization Mechanism of Class II transposable elements 31
2.1.4 Transposon activity across species 34
2.1.5 Transposable elements and Central Nervous System 39

2.2 Long non-coding RNAs 41
2.2.1 Identification and functions 41

2.3 Transcriptomic impact of repetitive elements and long non-coding RNAs 45

CHAPTER 3 – GENOMIC TOOLKIT FOR CEPHALOPODS 50

3.1 Genomic toolkit for Cephalopods 50

CHAPTER 4 – MATERIALS AND METHODS 54

4.1 *In-silico* studies 54
4.1.1 Collection of public transcriptomes 54

4.2 Generation of the transcriptome 56
4.2.1 RNA extraction, sequencing and quality filtering 56
4.2.2 *De novo* assembly and annotation of assembled transcripts 57

4.3 Gene Ontology enrichment analysis of the CNS and PNS expressed transcripts **60**

4.4 Classification of repetitive elements **60**

4.5 Identification of conserved transposable elements **61**

4.6 Identification of tissue-specific and candidate transcripts **62**

4.7 Polymerase chain reaction (PCR) **64**

CHAPTER 5 – RESULTS AND DISCUSSION **65**

5.1 General statistics from the *de-novo* assembly **65**

5.2 Functional annotation of the transcriptome **69**

5.3 Gene Ontology enrichment analysis of the CNS and PNS transcripts **74**

5.4 Distribution of TEs and lncRNAs in the *Octopus* neural transcriptome **77**

5.5 Comparison of the repeat and lncRNA contents among the species **80**

5.6 Structure and conservation of transposable elements **88**

5.7 Inspection of the CNS- and PNS-specific transcripts **93**

5.8 Experimental validations of candidate coding and lncRNAs **97**

5.9 Exploring the *Octopus*-resources **98**

CHAPTER 6 – CONCLUSIONS **101**

REFERENCES **105**

APPENDIX **123**

Chapter 1 – Aims and Introduction to the *Octopus vulgaris*

1.1 Prologue and plan of this PhD Project

When I started my studies in biology I learned that the genes required for the functioning of the cells and organisms, i.e. for the life on Earth, represent just a small portion of the whole genome of each individual species. I also learned that is particularly the case for human genome. Most of the DNA content present in each cell is either hidden in highly condensed DNA, or simply "junk" without known functions. In the following years, I learned that such "junk" DNA sequences play an important role in the regulation of gene activity, thus providing a significant contribution to the evolution of complex organisms. I also never imagined that during my PhD project I should be exposed to the challenging task of contributing – even at a very small "scale" – with the analysis of the role of this "junk" DNA in a marine organism, namely the cephalopod mollusc *Octopus vulgaris*, an invertebrate.

This resulted to be a particularly challenging task because for these organisms the genomic era started only recently and very little is know about genome (Albertin et al., 2012) composition and transcriptomes (DeGiorgis et al., 2011; Smith et al., 2011; Zhang et al., 2012b) of cephalopod species.

To face with this PhD Project, I applied computational tools and strategies on *O. vulgaris* transcriptome data to ***i.*** generate a reference transcriptome of the nervous system of this species, and then ***ii.*** analyze putative lncRNA populations, ***iii.*** identify and analyze repeats, and ***iv.*** identify transposition-competent elements present into the transcriptome.

In particular, I have utilized a computational pipeline (Musacchia et al., 2015) developed in the research group of Dr. Remo Sanges at the Stazione Zoologica Anton Dohrn (SZN; Napoli, Italy), to develop a reference transcriptome of *O. vulgaris* nervous system based on a set of RNA-seq experiments, and estimated the putative lncRNA populations included therein.

I have identified the specific expression patterns of both the coding and long non-coding transcripts in *O. vulgaris* nervous system. A few candidate tissue-specific coding and long noncoding RNAs have also been validated.

To contribute to the understanding of the organization of the transcriptome in an evolutionary context, I have identified repeats and explored their abundance among the central- and peripheral nervous systems of *O. vulgaris*. I have also compared the repeats content of transcriptomes of several invertebrate and vertebrate species.

Finally, I have explored the transcriptome resource generated to identify a subset of transcripts aimed to facilitate to address specific biological questions providing to decipher phenomena of biological plasticity in the cephalopod *O. vulgaris* for the research group of Dr. Graziano Fiorito.

1.2 Why Cephalopods

Cephalopods have greatly contributed to modern biology. The contribution to the emergence of neuroscience due to the discovery of the squid giant axon is one – among many – examples (Llinás, 2003; Young, 1985). Despite their importance as cases for studying biological novelties (e.g. Bonnaud et al., 2014) and the efforts

provided by researchers during different times of the last century in scientific research, the exploitation of cephalopods as 'model system' has been largely limited (De Sio, 2011; Ponte et al., 2013).

In recent years, this has been largely impeded by the difficulty of rearing cephalopods at different life-stages, and of the lack of genomic resources for these species. Current scientific attention is overwhelming these difficulties and are providing the impetus for systematic investigation and exploitation of cephalopods as 'resources' for scientific endeavours and public-utility (Albertin et al., 2012; Huffard, 2013; Ponte et al., 2013; 2014; Fiorito et al., 2014).

When I started this PhD project in 2012 cephalopods were exceptions among the species considered to be interesting enough to make their genome available to scientific community.

Cephalopods are an important component of the marine biome; they are important commercially as food resource for humanity, and are an emerging taxon for aquaculture.

In the words of Albertin and coworkers (2012), it is without doubt that the basis of neuronal function at the cellular and systems levels, the understanding of cephalopod population dynamics, the evolution of gene regulatory elements mediating body plan variation, would benefit greatly from the molecular insight that high-quality cephalopod genomics would provide.

In addition, biological research will enormously benefit from cephalopod genomics. The great majority of biological processes have been studied based on the favoured 'model organisms' among bilateral animals: deuterostomes and ecdysozoans. The first, essentially vertebrate species, have been sequenced with an expanding view towards basal chordates and selected non-chordate organisms such as sea urchins

and hemichordates. The second are highly represented by the insect *Drosophila melanogaster* and the nematode *Caenorhabditis elegans*. However, Cephalopods represent a large number of species with an elevated diversification of morphological and functional adaptations and only a very limited number of genomes have been sequenced from these species (Albertin et al., 2012).

The morphological novelties characterizing cephalopods, as compared with other molluscs, the extreme sophisticated physiological adaptations, including complex behavioural repertoire and cognitive abilities (Packard, 1972; Borrelli and Fiorito, 2008; Tricarico et al., 2011; Albertin et al., 2012; Huffard, 2013), place these species among the ideal candidates for modern genomics, and for a revisited dedicated research effort that may help our understanding on the evolution of biological processes.

1.3 Cephalopods as Bilaterians

Cephalopods are Bilateria and particularly Protostomes. In spite of the potential importance for understanding the evolution and diversification of bilaterians, the origins of cephalopods are still poorly understood. There is little consensus about the linkage between molluscs, cephalopods and other bilaterians. In recent years, emerging interests on these topics provided a revised analysis based on contributions derived from paleontology (Vinther, 2015), embryology (Shigeno et al., 2008), transcriptome (Kocot et al., 2011; Smith et al., 2011), and gene expression patterns (Yoshida et al., 2014).

Bilateria (i.e. Protostomes and Deutorostomes), are a large group of different animal taxa characterized by bilateral symmetry and three germ layers (**Figure 1.1**).

Phyla belonging to Protostomes and Deuterostomes are distinguishable on the basis of what occurrs during the embryonic development (Arendt et al., 2001; Hejnol and Martindale, 2008), that is the relative position/origin of the two openings of the digestive tract, namely the mouth and the anus. During gastrulation, the blastopore represents the unique opening to the primitive gut (Hejnol and Martindale, 2008). This will subsequently differentiate into the mouth (Protostomes), or the anus (Deuterostomes).

Members of phyla Echinodermata (e.g. starfish) and Chordata (e.g. fish, humans) belong to Deuterostomes.

Protostomes are currently distinguished into:

i. Platyzoa, including platyhelminthes (flatworms) and rotifers (Cavalier-Smith, 1998).

ii. Lophotrochozoa, including the segmented worms, molluscs, lophophorates and several other smaller phyla (Halanych et al., 1995).

iii. Ecdysozoa, including arthropods and phyla that moult periodically (Telford et al., 2008).

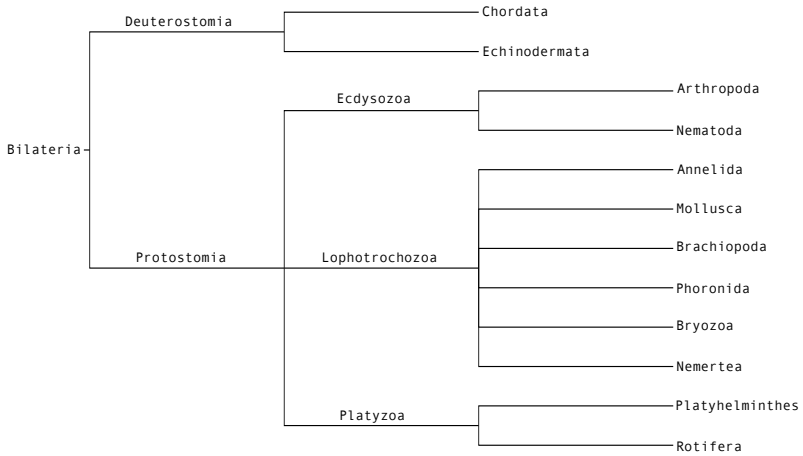

Figure 1.1 The relationship between major bilaterian clades.
(modified after Niehrs et al., 2010.)

Mollusca are therefore belonging to Lophotrocozoa. The phylum is the second most numerous in the animal kingdom, with more than one hundred thousand species described and classified in eight major lineages (Ponder and Lindberg, 2008). Morphological diversity among those animals has generated conflicting phylogenetic hypothesis. Neither traditional morphological or molecular phylogenetic approaches have solved the question (Passamaneck et al., 2004). However, recently transcriptome and genomic data greatly contributed to the analysis of the evolutionary relationships within molluscs (Kocot et al., 2011; Smith et al., 2011).

From these studies, a revisited view of the Mollusca lineage emerge, with two major groups: Aculifera and Conchifera (**Figure 1.2**). The first include Polyplacophora and Aplacophora.

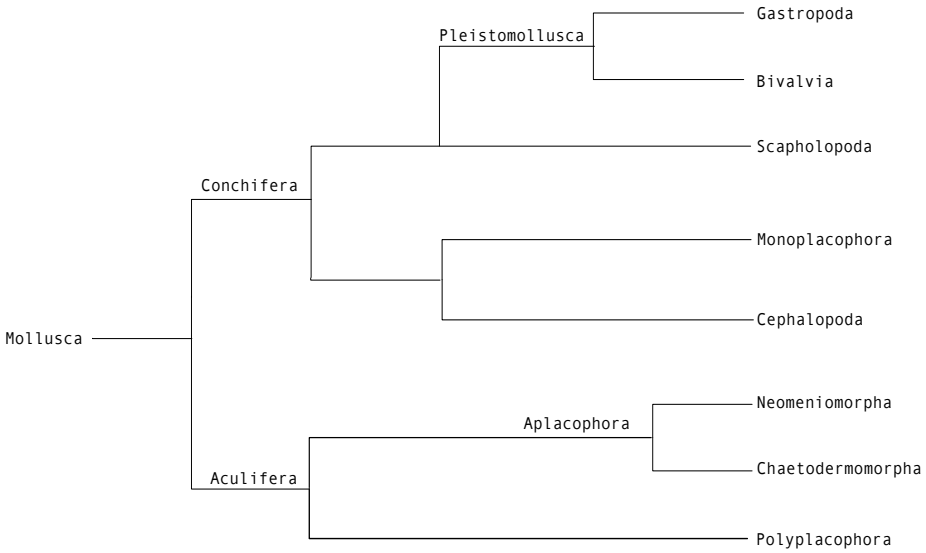

Figure 1.2 Phylogenetic relationships between major Mollusca lineages.
(deduced from: Kocot et al, 2011; Smith et al. 2011).

Within Conchifera, Gastropoda and Bivalvia appear as sister taxa within the Pleistomollusca clade; these two classes accounts for more than 95% of described

species of the entire phylum. The Cephalopoda have been proposed as sister group of Monoplacorphora (Smith et al., 2011).

1.4 Main features of Cephalopods

The Cephalopoda appeared in the Upper Cambrian period about 500 million years ago (Mya) (Smith and Caron, 2010). During the mid-Palaeozoic period (~400 Mya) cephalopods diverged into Nautiloids and forms that originated modern Coleoids, i.e. cuttlefish, squid and octopus (Packard, 1972) (**Figure 1.3**).

Among Mollusca, gastropods and cephalopods are characterized by an elevated degree of 'cephalization' when compared to other molluscs, the presence of complex eyes, and the emergence of the head from visceral mass. However, it has been suggested that gastropods and cephalopods have evolved independently (Kocot et al., 2011; Smith et al., 2011). The evolution of cephalopod body plan is considered one of the most intriguing topics to study in zoology and evo-devo. The body parts of coleoids include multiple sets of characters with striking analogy with vertebrates (Packard, 1972).

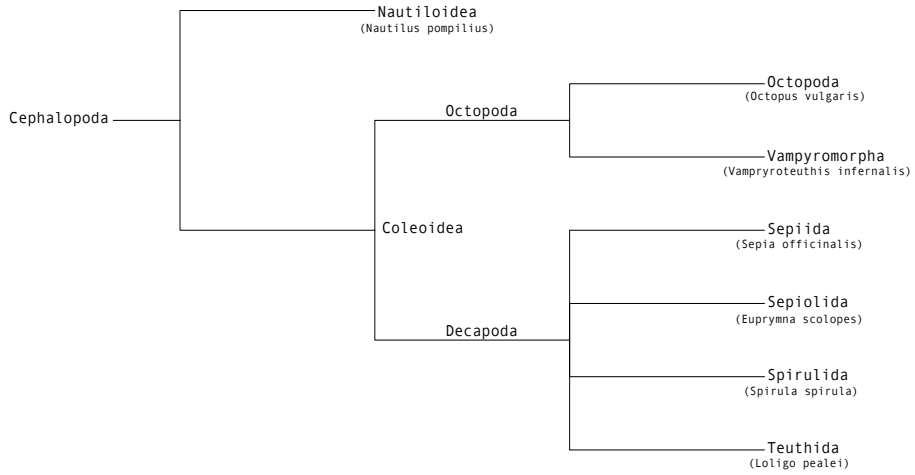

Figure 1.3 Phylogenetic relationships between major taxa of Cephalopoda
(modified after Ruppert et al., 2003).

Resemblance between cephalopods and fish is traceable at different levels of analysis: body form and locomotion, hydrostatic control and buoyancy, feeding habits, general physiology (including cardiovascular system, respiration and excretion), organization of the central nervous system and sensory organs, ontogeny and growth rates (Packard, 1972; Kröger et al., 2011). In the words of Packard, basic physiological mechanisms characteristics of molluscs have been incorporated in systems with performance comparable to vertebrates; this is also achieved at the level of behavioural repertoire.

Nautiloids are the earliest already living cephalopods forms found in the fossil record. They are all pelagic (six species are known) characterized by an external shell. In addition, they possess only tentacles (up to 90) and an esophageal nerve ring made of different ganglia, commissures and connectives. In Coleoidea the shell is internalized or lost. More than 700 living species are known inhabiting almost the great majority of marine habitats (Packard, 1972; Hanlon and Messenger, 1996) On the basis of the presence of a pair of tentacles they are distinguished into Decapoda

and Octopoda. In decapods (e.g. *Sepia*) the shell is internalized (cuttlebone), or greatly reduced (e.g. *Loligo*) to form a gladius located dorsally within the mantle. All decapods have eight arms and two tentacles. In octopods tentacles disappear and the shell is absent (with few exception due to the presence of a very small vestigial shell).

In more "functional" terms, the phylogeny of modern cephalopods can be described as the story of how the reduction and eventual loss of an external shell brought to the emergence of alternative devices adopted to regulate buoyancy (Wells, 1994). According to Packard (1972) factors that acted behind the dramatic modifications in the cephalopod body-plan ("Bauplan") are largely due to the competition of evolving vertebrates in the Mesozoic. To adequately compete for resources and as defence mechanisms from predatory pressure, cephalopods evolved unique adaptations in morphology, physiology, behaviour and ecology. The loss of a static external covering, provided by the shell, made coleoids vulnerable to sharp-toothed predators. The disappearance of the shell stimulated the capability to evolve and diversify jet propulsion (Wells and O'Dor, 1991). This resulted in a premium: the evolution of a dynamic covering, the skin, capable of altering the animal's appearance from one moment to the next due to sophisticated neuro-muscolar control (review in Messenger, 2001), thus confusing potential predators. Coleoids became master of disguise. In this, cephalopods are unique in the animal kingdom. In fish, such changes are not so fast as they are mainly under non-neural control mechanisms. In competing with vertebrates, cephalopods thus reached high levels of functional convergence with them (Packard, 1972; see also: Aronson, 1991; O'Dor and Webber, 1986, 1991).

The nervous system of cephalopods represents one of the most intriguing "features" of these animals in terms of morphological arrangements and functional adaptations. Although it shares with other members of the phylum the basic molluscan design, it reaches an elevated degree of centralization (i.e. encephalization) comparable to the one of vertebrates; similarly to vertebrates, coleoids evolved a rigid cartilagineous enclosure to protect the masses surrounding the oesophagus that are commonly named as "brain". In terms of mass, this centralization of ganglia and nervous components results in providing cephalopods with a brain-to-body weight ratio that exceeds that of fishes, amphibians and reptiles among vertebrates (Packard, 1972).

Cephalopods represent something unusual among invertebrates and this is due to 'their' huge investment in nervous machinery. In the common octopus, about 500 million nerve cells constitute the nervous system (Young, 1963). This number results to be ten thousand times higher than that found in another mollusc, *Aplysia*, and still remains two hundred times higher when compared to the number of neurons present in the brain of the honeybee (*Apis mellifera*; review in Borrelli and Fiorito, 2008). In a part of the octopus brain (i.e. the vertical lobe) accounting for less than 10% of the entire 'central' nervous system, more than 25 million of neurons are accommodated. These cells are of about 3 microns in size, and arranged in circumvolutions (lobules) thus to maximizing 'computational' surface per volume.

During evolution, the different ganglia of the putative molluscan ancestor started to fuse together by the shortening of the connectives and commissures that were clustered tightly around the anterior part of the oesophagus, with the cerebral and buccal ganglia mostly arranged above and the remaining ganglia below (for review see: Bullock and Horridge, 1965; Budelmann, 1995). As mentioned above, the

simplest organization of the central nervous system of cephalopods is that of *Nautilus,* with three broad bands, one dorsal (cerebral ganglia and commissure) and two ventral to the oesophagus (pedal anterior and palliovisceral posterior) are joined laterally (Owen, 1832). Lobes protruding from these bands are missing, and the fusion of the ventral structures in a single subesophageal mass is not observed. However, a closer examination of the assemblage of the three bands allows to identify various differentiated lobes inside the bands, that closely resemble similar structures of coleoids (Young, 1965; for review see: Nixon and Young, 2003).

During the course of its evolution, the brain of cephalopods increased its complexity. In Dibranchiates (or Coleoids), it became completely surrounded by a cartilaginous capsule and reached the maximum agglomeration of the neural masses by being fused in a supra– and sub– esophageal part and two large optic lobes (one on each side), extending laterally from the supraesophageal mass.

Cephalopods and vertebrates should be seen as two independent experiments in the evolution of large and complex nervous systems. Within vertebrates, mammals and birds have been separated for about 300 million years ago. However, vertebrates share a common body-plan, and their nervous systems show a common inheritance (Edelman and Seth, 2009). Cephalopods have an entirely different organization, both in body and brain.

The vertebrate plan is characterized by a rather centralized design featured by a head and spinal cord, with the peripheral nervous system coming off it. Molluscs, along with several other invertebrate groups, developed a 'ladder-like' nervous system; knots of neurons, or ganglia, spread along the body, linked by two kinds of connections, i.e 'vertical' (along the body) and 'horizontal' (across it), thus to

resemble a ladder. During the evolution of coleoid cephalopods expansion and partial centralization of this ladder-like nervous system has been achieved.

Differently from vertebrates, cephalopods share with most of the molluscan species the phenomenon of neural giantism (Gillette, 1991). Large neural cells are required for fast conductive properties in absence of a proper myelinisation. The counterpart of the neural giantism is that large ganglia are produced; these in cephalopods have collapsed to produce masses relatively large in size. Thus, the gross size of the cephalopod "brain" is the by-product of a relatively impressive number of cells, mostly of which relatively small in diameter (about 3 to 10 microns). Furthermore, the idea that cephalopods are exceptional invertebrates due a simple nervous system that produce complex behaviour also encounters challenges. Examples are many other arthropod species (e.g. social insects, stomatopods) in which cases of complex behavior, life-history strategies, and mini-brains (Menzel and Giurfa, 2001) appear to have formed their own 'unusual' evolutionary lineage (Cronin et al., 2006).

Despite these remarkable differences, the vertebrate and cephalopods "experiments" in evolution show fascinating comparisons (for a critical review see also Shigeno et al., 2010). In cephalopods, neurons are packed together up front between the eyes, and many of the ganglia are fused, thus to provide with a partial submerging of the invertebrate neural plan. The nervous system of cephalopods shows a series of features that are considered to be unusual to molluscan, and invertebrate, or even vertebrate standards. As reviewed by Borrelli and Fiorito (2008), these are: ***i.*** the highest degree of centralization compared to any other mollusc or invertebrate (insects excluded), achieved by the shortening of the

connectives; ***ii.*** the presence of very small neurons (3-5 micron of nuclear size) acting as local interneurons; ***iii.*** the apparent absence of correspondence of receptors in the body regions to specific functional areas of the brain ("somatotopy") (except for the chromatophore lobes) contrary to what appears to be the case for the insect or vertebrate brain (Plän, 1987); ***iv.*** a blood-brain barrier (an exception for molluscs; Abbott and Pichon, 1987); ***v.*** compound field potentials (similar to those of vertebrate brains); ***vi***. an elevated efferent innervation of the receptors (e.g. the retina, the equilibrium receptor organs); ***vii.*** peripheral first order afferent neurons; ***viii***. a large variety of putative transmitters (for review see Messenger, 1996).

Such a sophisticated central nervous system coupled with a battery of well developed sense organs is the by-product of cephalopods' life style as voracious marine predators and competition with emerging vertebrates in the sea realm (Packard, 1972; Young, 1977).

1.5 Octopus' central nervous system

As mentioned above, in cephalopods the ganglia recruited to form the central nervous system may be considered homologous to the labial, buccal, cerebral, pedal, pleural, and visceral ganglia of gastropod molluscs. Differently from the typical molluscan design, in cephalopods the ganglia are fused together and clustered around the most anterior part of the esophagus, forming three almost distinct parts: the supra- and the sub- esophageal masses, and optic lobes (a pair) lateral to the supraesophageal mass and positioned just behind the eyes (**Figures 1.4; 1.5**).

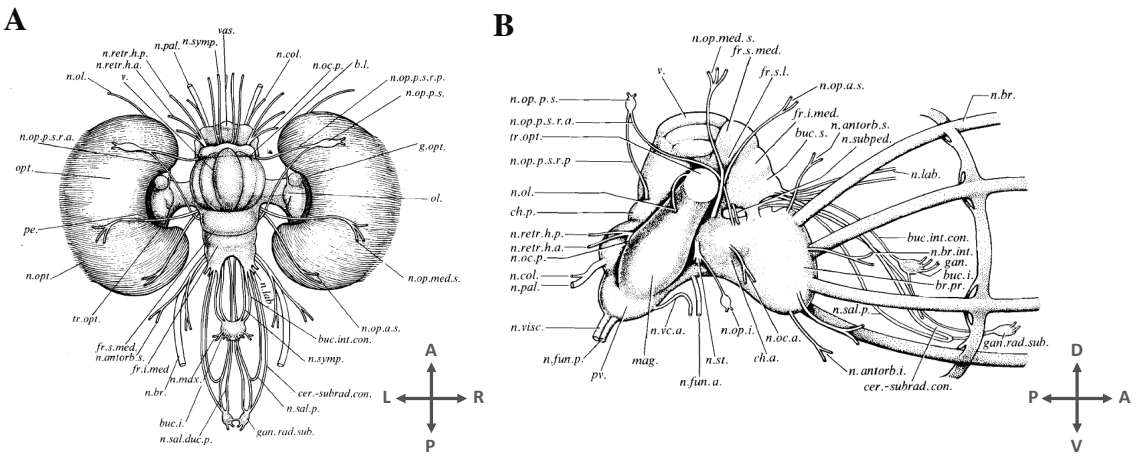

Figure 1.4 (A) Dorsal view and (B) Lateral view of the *Octopus vulgaris* nervous system. a., anterior; antorb., antorbital; b. basal; br., brachial; buc., buccal; c., commissure; ce., cell; cen., central; cer., cerebral; ch., chromatophore; col., collar; con., connective; cr., crista; d., dorsal; duc., duct; fl., fin lobe; fr., frontal; fun., funnel; g., gland; gan., ganglion; gi., giant; h., head; i., inferior; int., inter; l. lobe; lab., labial; lat., lateral; mac., macula; max., maxillary; med., median/medial/middle; mus., muscle; n., nerve; oc., oculomotor; oes., (level of) esophagus; ol. Olfactory; op., ophthalmic; opt., optic; p., posterior; pal., pallial; pe., pedal; ped., peduncle; po., post; prec. Precommissural; pv., palliovisceral; r., root; rad., radular; retr., retractor; s., superior; sal., salivary; st., static; sth., statolith; subrad., subradular; subv., subvertical; symp., sympathetic; tr., tract; v., vertical; vas. Vasomotor, ven., ventral; visc., visceral. The arrows indicate the posterior (P), anterior (A), dorsal (D), ventral (V), left (L) and right (R) lateral. The figures are modified from J.Z. Young (1971).

The “central brain” i.e. the supra- and sub-oesophageal masses, is surrounded by a cartilaginous capsule (the skull); the two masses are dorsal and ventral to the oesophagus, respectively. The two large optic lobes, one on each side, are connected to the retina of the highly developed camera eyes and to the supra- oesophageal mass.

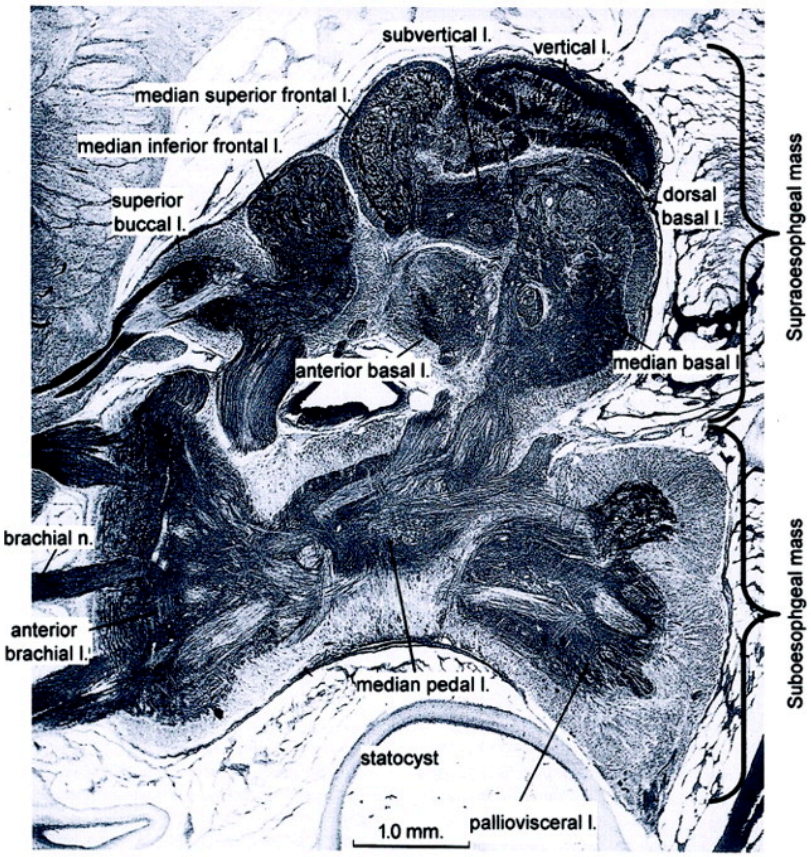

Figure 1.5 Sagittal section of the central nervous system of the *Octopus vulgaris* (taken from Hochner et al., 2006).

The ***supraesophageal mass*** (***SEM***) is constituted by numerous lobes:

- **superior buccal lobe**
- **inferior frontal lobe**
- **sub-frontal lobe**
- **superior frontal lobe**
- **vertical lobe**
- **sub-vertical lobe**
- **pre-commisural lobe**
- **basal lobe system**

The **superior buccal lobe** (the most anterior part of the SEM) is composed by 150 x 10^3 cells having a size of about 5–10 μm. According to Young (1963), it has the

peculiarity of having many cells larger than 10 μm with some reaching 20 μm. The posterior part of buccal lobe is strictly connected to the **inferior frontal lobe** (1085 x 10^3 cells), a centre of elaboration of chemo-tactile information. The lobe is classically divided into lateral inferior frontal, median inferior frontal and sub-frontal lobes. These three parts are very dissimilar. The medial part of inferior frontal lobe has a large neuropil and a thin layer of small cells (<5 μm; with only a few much larger cells: 5–20 μm). The lateral inferior frontal lobe has many small cells (<5 μm) and some large neurons (up to 20 μm). The **sub-frontal lobe** is principally composed by very small cells (nuclear diameter less than 3 μm) forming dense masses; dispersed in the wall of this lobe are cells with nuclei up to 20 μm.

In the **superior frontal lobe** (1772 x 10^3 cells) two regions are clearly distinguishable: lateral and medial superior frontal lobe. The first is composed by about half small cells (less than 5 μm in diameter), the remaining being 5–10 μm; the median superior frontal lobe has small cells (about 4 μm in diameter).

The **vertical lobe** is the most dorsal and elongated structure of the SEM. It is partitioned in *O. vulgaris* into five gyri (or lobules), which are approximately cylindrical structures running along the antero-posterior axis of the lobe. Each 'cylinder' has a cellular wall surrounding a neuropil where cellular bodies are dispersed. The lobe is composed by a total of about 25000 x 10^3 cells; the larger number of cells is represented by small ones: the amacrine cells. These are the smallest recognizable in the octopus brain (3 μm in diameter). A much smaller number of larger cells with nuclear diameter 5–10 μm are also found. In contrast, with the usual arrangement of cells in the other lobes, the larger cells are more close to the neuropil.

Lying below the vertical lobe, the **sub-vertical lobe** (810 x 10^3 cells) is extending itself between the vertical, superior frontal, the basal lobes (ventral). It is characterized by the presence of a wall that in several regions is folded to form islands of cells. Many of these cells have a diameter inferior to 5 μm and very few are larger than 10 μm. At both sides of the sub-vertical lobe there are cells 5–10 μm in diameter, but also neurons that are the largest of the whole supraesophageal mass (i.e. 25 μm).

In the center of the supraoesophageal mass and below the subvertical lobe, the **pre-commissural lobe** (78 x 10^3 cells) is a structure characterized by many fibers, considered as a "meeting-point" for many fibre systems. It is characterized by few layers of cells, mostly of small ones (<5 μm), by a similar proportion of medium sized (5–10 μm) and some larger cells (i.e. about 1000, 10–15 μm in diameter). Its neuropil is continuous with the above lying sub-vertical lobe.

More in a ventral position of the supra-oesophageal mass, are located a series of lobes belonging to the **basal lobe system**. This is considered the most 'higher' centre for motor control. The 'system' is characterized by several structures (six lobes or parts) with cells of different dimensions and either distributed in layers or dispersed in the neuropile.

Outside the cranium and at each side of the head, there are the largest lobes of the whole central nervous system, i.e. the **optic lobes**. They contains about 92,000 x 10^3 cells and represent about the 45% of the total volume of an octopus brain (Nixon and Young, 2003). The optic lobes are characterized by the presence of an outer cortex and a central medulla. The cortex is composed by outer and inner layers of granular cells separated by a plexiform layer (i.e. neuropil). These layers contain cells

smaller than 5 μm, and several other medium sized cells (5–10 μm). The smaller cells are located next to the neuropil; the larger ones lie at larger distance away from it. In the center of the lobe, islands of nerve cells and many scattered small cells are found. Description of the ultrastructure and connections within the optic lobes is available (Dilly, 1963; Saidel, 1982).

Proximal to the optic lobes and positioned on the optic tracts there are three different structures quite small in the octopus brain relative to those occurring in Decapods: the **peduncle** (142 x 10^3 cells), the **olfactory** lobes (136 x 10^3 cells) and the **optic glands**.

The **peduncle lobe** is considered to be involved in the control of attack and movements, thus a possible analogous to the cerebellum (*sensu* Messenger, 1967b). It contains principally medium-sized cells (5–10 μm) with some others smaller than 5 μm. The **olfactory lobe** has a cellular composition similar to the peduncle lobe, and it seems to play a chemoreceptor function. According to Wells and coworkers the **optic gland** has endocrine function (Wells and Wells, 1959).

According to Young (1971), in the ***subesophageal mass*** (**SUB**) it is possible to distinguish three regions: anterior, middle and posterior. The anterior region is constituted by the **prebrachial** (261 x 10^3 cells) and **postbrachial** lobes (80 x 10^3 cells), involved in the control of arm movements. The prebrachial lobe is composed by a high proportion of small cells (5 μm) and large ones (20 μm in diameter); in contrast, in the postbrachial lobe large cells (25 μm) are the main constituents.

The middle SUB region includes the **anterior chromatophore** lobes (217 x 10^3 cells) and the **pedal lobe** (243 x 10^3 cells). The higher proportion of neurons here is medium-sized (5–15 μm) or large cells (15–20 μm), and few smaller ones (<5 μm). In

contrast, the lateral parts of middle subesophageal mass contain a greater number of small cells. The posterior part of the middle SUB has pear-shaped cells (mainly medium in size).

In the posterior SUB region are identifiable: the **palliovisceral lobe** (108 x 10^3 cells), the **posterior chromatophore** lobes (309 x 10^3 cells), **vasomotor** (1307 x 10^3 cells) and **magnocellular** lobes (581 x 10^3 cells). The palliovisceral lobe contains neurons that play different roles, although they appear not separated in anatomically distinct districts. The outer cellular layer is composed by large cells (10–20 μm), in proximity of the neuropil these are very few and small (5 μm). The vasomotor lobe has thick walls of numerous small cells (5–10 μm) with few large ones towards the periphery. The chromatophore lobes (one for each side) contain very numerous cells; in the most anterior part mainly small cells (5-10 μm) are present; in contrast, the most posterior part of the lobe has 15–20 μm cells.

The magnocellular lobes surround the esophagous (one for each side) and are considered to control the basic behavioural reactions including defensive motor patterns. The most dorsal part has small and medium-sized cells (5-10 μm) with only few larger cells (10–15 μm). In the ventral part the cell layers are much thicker and there are many larger cells (between 15 and 20 μm).

1.6 Octopus' peripheral nervous system: arms

The peripheral nervous system of Cephalopoda is a complex network of nerves that has been described in great detail by Pfefferkorn (1915). The eight cords along the length of the arms and the ganglia at their bases, the two stellate (one for each

side of the mantle), two brachial, two cardiac and one gastric, are included in the peripheral nervous system (Bullock and Horridge, 1965).

The 'peripheral system', i.e. mantle (Bone et al., 1981; Milligan et al., 1997), arm (Matzner et al., 2000) and chromatophore muscles (Florey and Kriebel, 1969) are all finely innervated and post-synaptic recording techniques have been applied for the analysis of synaptic dynamics in each of these preparations.

O. vulgaris has eight arms, which are regularly arranged around the mouth. On the inner surface a double row of about 240 suckers for each arm. In each arm a complex muscular system, together with the neural components, allow complex movements. Within the arm the intrinsic muscles of the arm and of the sucker, and the acetabulo-brachial muscles that unite intrinsic muscles of arms and suckers, provide the entire setting-up of the system (**Figure 1.6**). Three muscle bundles (longitudinal, transverse, and oblique) are identifiable; each bundle is innervated separately from the surrounding units and shows a remarkable autonomy. The contraction and relaxation of different muscle bundles allows the animal to have extraordinary motor capabilities through the absence of a bony or cartilaginous skeleton.

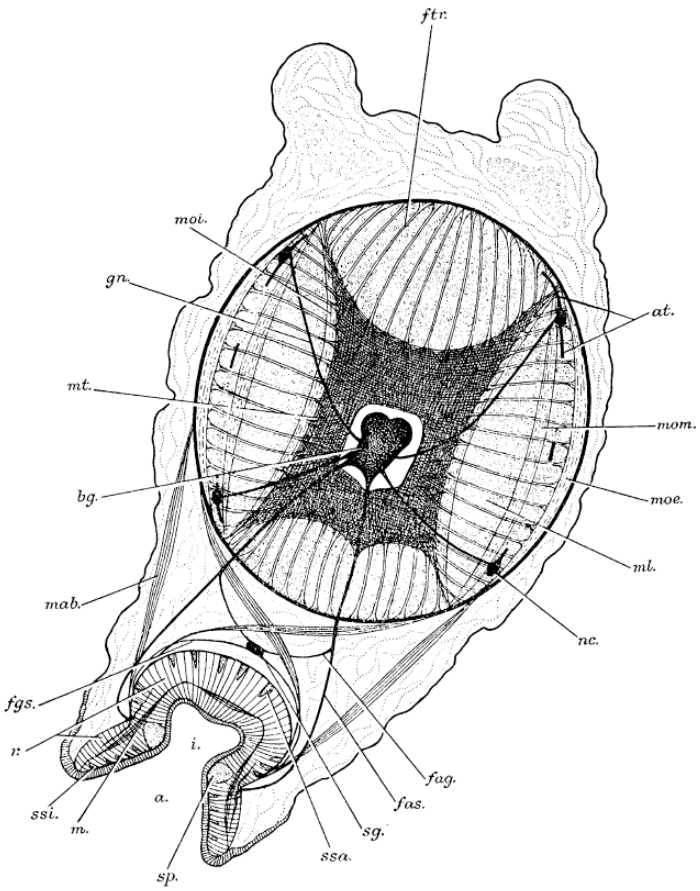

Figure 1.6 Cross-section of the arm of the *Octopus vulgaris*. Figure is reproduced from Graziadei (1964). a., acetabulum, at., anastomosis between intramuscular nerve cords, bg., brachial ganglion (axial nerve cord), fag., nerve bundle connecting the sucker's nerves to the ganglion, fas., nerve bundle connecting the brachial ganglion to the sucker, fgs., nerve bundle connecting the sucker's ganglion to the sucker itself, ftr., trabecular bundles of the transverse muscle system, gm., nerve bundle connecting the bg. with the nc, i., infundibulum, m., meridial muscles, mab., acetabulo-brachial muscles, ml., longitudinal muscles, moe., oblique external muscles, moi., oblique internal muscles, mom., oblique median muscles, mt., transverse muscles, nc., intramuscular nerve cords, r., radial muscles, sg., sucker's ganglion, sp., principal sphincter, ssa., secondary sphincters of the acetabulum, ssi., secondary sphincters of the infundibulum.

The nervous system of the arms contains about two-thirds of the total number of neurons in the octopus. It consists of a large brachial (axial) nerve cord in the centre and four smaller intramuscular nerve cords projecting in the periphery. The structure of the brachial nerve cord is similar to that of the other invertebrates (Graziadei, 1965). The central portion of the cord consists of a network of nerve fibres forming the neuropil. The nerve cells lie around this in an uninterrupted layer. The diameter

of these nerve centres is smaller near the tip of the arm. Each axial cord sends their axons to innervate the smaller intramuscular nerve cords situated in the angles of a square lying among and the ganglion of the sucker that is a small assembly of nerve cells lying below the acetabular cup of each sucker. The ganglion of sucker consists of a few hundred neurons; some of them are motor neurons that send axons to the muscles of the peduncle and the sucker, other bipolar and multipolar neurons have been observed in this ganglion, but their function is still unknown.

1.7 Comparison of the *Octopus* and the *Homo* Nervous System

The modern Cephalopods (coleoids) have changed dramatically the body structure and way of life from those of *Nautilus* and other Molluscs (Packard, 1972). These changes became possible following the development of sophisticated motor, sensory and cognitive abilities, such as excellent vision, highly efficient flexible arms and the ability to learn rapidly (Budelmann, 1996; Flash and Hochner, 2005; Hanlon and Messenger, 1998; Muntz, 1999).

The evolutionary process leading to the selection, in some cases, of unique features among the modern Cephalopods (e.g. the chromatophore system) and in other cases, of analogous systems in structure and function to those of vertebrates through convergent evolutionary processes (O'Dor and Webber, 1986; Packard, 1972). Undoubtedly the most famous cephalopod-vertebrate convergence is the camera eye (Packard, 1988). This organ is highly diverse in structure, ranging from small groups of light-sensitive cells to highly sophisticated and complex structures that register precise images in some groups of protostomes and deuterostomes (arthropods, molluscs and vertebrates). Despite some differences, there is a

structural similarity between cephalopod and vertebrate eyes. Indeed, between human and octopus eyes, each of the tissues such as eyelid, cornea, pupil, iris, ciliary muscle, lens, retina, and optic nerve/ganglion corresponds well to each other (**Figure 1.7**).

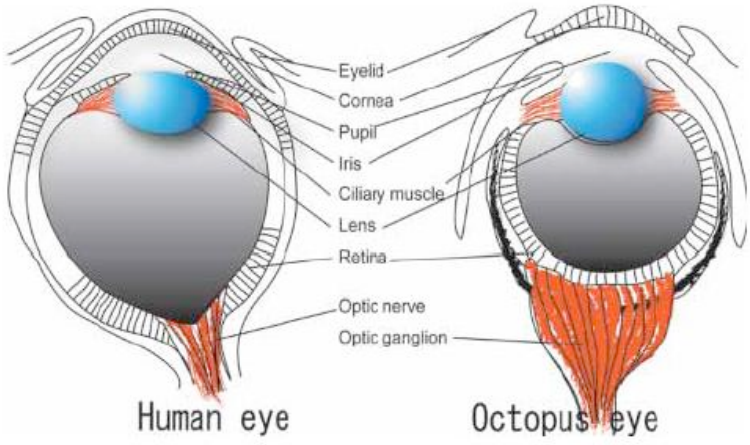

Figure 1.7. **Structural similarities between human and octopus eyes.**
(taken from Ogura et al., 2004).

More recently, a comparative study using gene-expression profiling has demonstrated that more than half of the genes examined (69%) were commonly expressed in the camera eyes of human and octopus (Ogura et al., 2004). These genes are known to be important for the formation and function of the vertebrate camera eye. For example, *Six3* (SIX homeobox 3) is necessary for the patterning of anterior neuroectoderm including the retina, and *Lhx2* (LIM homeobox protein 2) is also necessary for normal development of the eye, particularly the retina (Carl et al., 2002; Porter et al., 1997). The camera eye is not the only Cephalopod system that shows convergence to the vertebrates. Certain areas of the *Octopus* brain are particularly interesting with respect to evolutionary convergence because they show a strikingly similar morphological organization to areas of the vertebrate brain that mediate

similar functions. For example, the three cortical layers of the Cephalopod optic lobe are organized similarly to the deeper layers in the vertebrate retina (Young, 1971). This similarity in the integrational layers is all the more striking because the mechanisms of transduction and physiological responses to light are totally different (Hardie and Raghu, 2001). As in other invertebrates, the membrane potential of the *Octopus* photoreceptor cells is depolarized (positive potential change) in response to light, whereas in vertebrates it is hyperpolarized (negative potential change). These opposite response to light are mediated by two different second messenger cascades. Similarly, the structure of the peduncle lobe, in which small granular cells give rise to arrays of thin parallel fibers, resembles the arrangement in the folia of the vertebrate cerebellum (Hobbs and Young, 1973; Woodhams, 1977; Young, 1976). The peduncle, together with higher motor centers in the basal lobes, receives inputs from both the visual and gravitational (statocysts) systems and has cerebellar-type effects on motor function (Messenger, 1967a, 1967b). The parallel and linear organization of small-diameter fibers in the vertebrate and the *Octopus* system suggests the importance of this type of organization for the timing computations needed to integrate visual and gravitational information for body motor and eye coordination. Finally, the vertical lobe (VL), containing 25 million of neuronal cells, is hierarchy the highest in the *Octopus* brain. Studies show that the VL is especially involved in the long-term and more complex forms of memory (Fiorito and Chichery, 1995; Fiorito and Scotto, 1992). Its architecture and physiological connectivity resembles the vertebrate hippocampus. More recently, it has been studied the central nervous system (CNS) of the *Octopus* through an RNA-seq approach (Zhang et al., 2012b). It is certainly to be noted that the CNS of *O. vulgaris* has a large number of proteins involved in specific junctions, transporters, and enzymes (e.g. Claudins, Occludin, Junctional adhesion

molecules, Cadherins and Catenins) which are definitely indispensable to form an incredible system that may possesses most vertebrate Blood-Brain Barrier (BBB) functions. The overall body shape of an *Octopus* is most unlike that of a vertebrate, but there are actually many anatomical convergences. Despite their flexibility, when the arms grasp food, they show remarkable similarity to a human arm, being reconfigured into a stiffened structure consisting of three units that articulate via a series of "pseudo-elbows" (Sumbre et al., 2006). Cephalopods normally swim by jet propulsion. Some octopuses, *Octopus marginatus* and *Octopus* (*Abdopus*) *aculeatus*, can also stroll across the sea floor by bipedal locomotion (Huffard et al., 2005). Two of the eight arms are applied sucker-side down and unfurl along their length to provide a rolling locomotion that kinematically can be classified as walking. In male octopuses, one of the arms (known as the hectocotylus) is modified for copulation and used to transfer a spermatophore to the oviduct of the female. In the California *Octopus* (*Octopus bimaculoides*), the arm tip is erectile, probably due to transfer more sperm. It is highly unusual among invertebrates, and structurally strongly convergent to the mammals (Thompson and Voight, 2003).

Chapter 2 – Repetitive elements within the genome

2.1 Repetitive Elements

2.1.1 Classification and structure

The rapid influx of data from whole-genome sequencing has created a rich data source to study the evolution of eukaryotic genomes. Most of those genomes are largely composed by noncoding regions and repetitive elements (for review see Jurka et al., 2007). The term "repetitive elements" refers to DNA fragments that are present in multiple copies in the genome and are capable to encode enzymes for the integration of their DNA into the host genome. They were discovered and initially classified into "highly" and "middle" repetitive sequences based on renaturation time and concentration of denatured DNA (Britten and Kohne, 1968). Currently, repetitive elements can be divided in two basic types: tandem repeats and interspersed repeats (for review see Jurka et al., 2007). Tandem repeats (TRs) occur in DNA when a pattern of two or more nucleotides is repeated and the repetitions are directly adjacent to each other. Interspersed repeats, also known as transposable elements (TEs) or mobile genetic elements (MGEs), are repeated DNA sequences able to move from one chromosomal position to another within the same genome (for review see Finnegan, 2012).

Transposable elements can be classified into two major classes according to the mechanism of transposition: ***i.*** Class I elements transpose via an RNA intermediate and are generally referred to as "Retroelements", ***ii.*** Class II elements are DNA mediated and are often called "DNA transposons" (Levin and Moran, 2011). The retroelements are subdivided into two groups based on the presence or absence of

long terminal repeats (LTRs), which flank the body of the element: LTR and non-LTR retrotransposons (**Figure 2.1**).

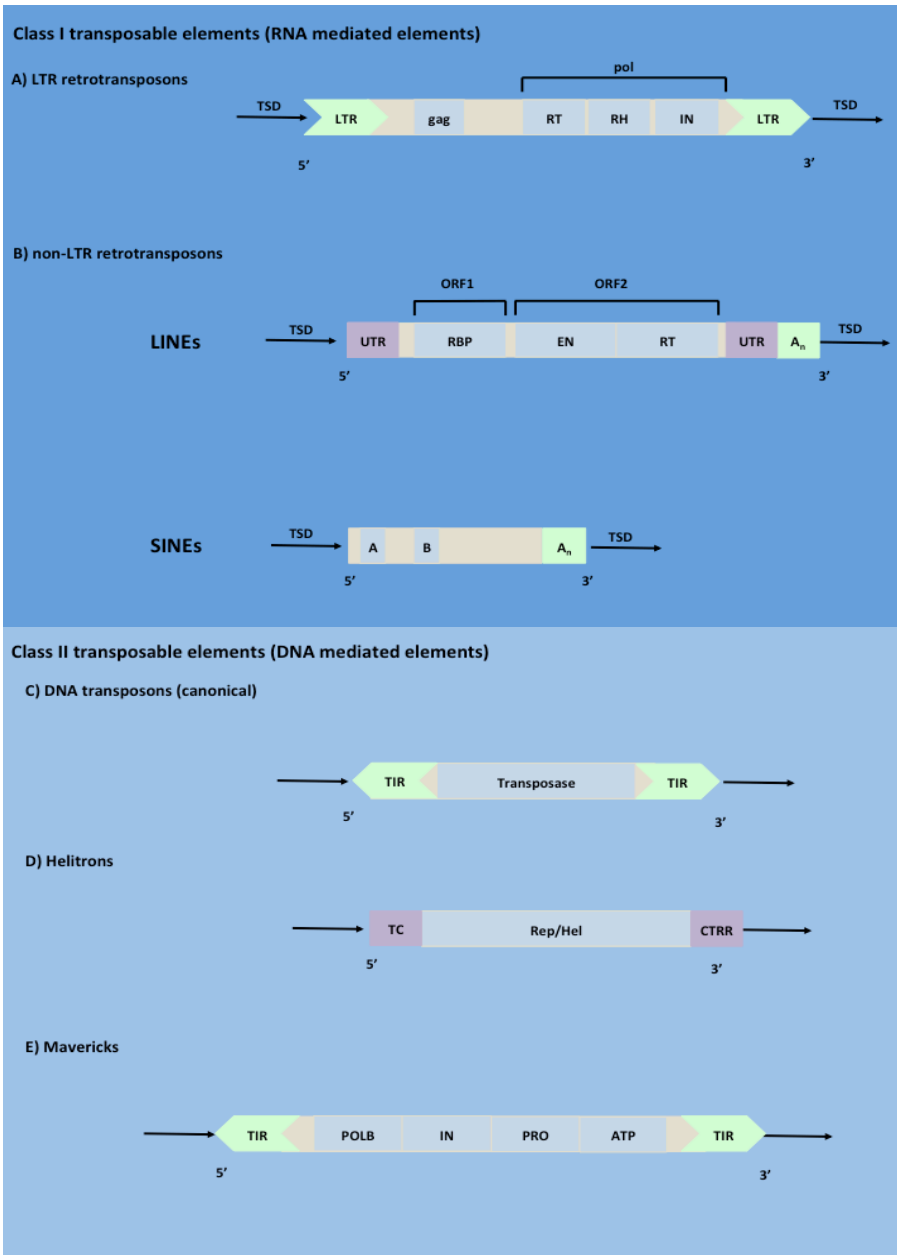

Figure 2.1 Schematic representation of major classes of transposable elements. Class I transposable elements RNA mediated (**A**) Long terminal repeat (LTR) and (**B**) non-LTR elements and Class II transposable elements DNA mediated (**C**) DNA transposons (**D**) Helitrons (**E**) Mavericks. TSD target site duplication, LTR long terminal repeats, *gag* gene encodes a capsid like protein, *pol* gene contains: RT reverse transcriptase, RH RNaseH, IN integrase; UTR untranslated region, RBP RNA binding protein, EN endonuclease, RT reverse transcriptase, A_n poly(A), TIR terminal inverted repeats, Rep replication initiator, Hel helicase, POLB DNA polymerase B, PRO protease, ATP ATPase.

The LTR retrotransposons were the first retroelements to be discovered in eukaryotes, and are similar in structure and coding capacity to endogenous retroviruses (ERVs) (Boeke and Corces, 1989). These can range from a few hundred bases up to 5kb, encoding two open reading frames (ORFs) called *gag* (specific group antigen) and *pol* (polymerase). The *gag* codifies an RNA binding protein and other proteins involved in the maturation and packaging of retrotransposons. The *pol* ORF encodes the enzymes necessary for the transposition: reverse transcriptase (RT), integrase (IN) or endonuclease (EN), and RNase H (RH) (Burke et al., 2002; Prak and Kazazian, 2000).

Non-LTR retrotransposons consist of two sub-types, long interspersed elements (LINEs) and short interspersed elements (SINEs). The LINEs contains one or two open reading frames (ORFs) (Martin, 1991). In the human genome, LINEs are about 6-kb and comprise a 5'untranslated region (UTR) sequence that contains an internal bidirectional DNA polymerase II (Pol II) promoter, followed by two ORFs, a 3'UTR and a poly(A) tail. The ORF1 encodes a RNA-binding protein, and ORF2, a protein with endonuclease and reverse transcriptase activity (Feng et al., 1996; Mathias et al., 1991). These enzymatic activities allow their mobility along the genome.

The insertion of LINEs into DNA during retrotransposition results in target site duplications (TSD), which flanks the new insertions. SINEs are short elements, approximately 80- to 400-bp (Jurka et al., 2007) typically divided in three modules: head, body and tail. The 5'-terminal head has two internal promoters (boxes A and B) for the RNA polymerase III (pol III), revealing their origin from one of the cellular RNAs synthesized by pol III: tRNA, 7SL RNA or 5S rRNA. The body is usually unique for each SINE family. The 3' part of a typical SINE body is similar to the 3' end of a partner LINE (Ohshima et al., 1996; Okada and Hamada, 1997). The 3'-terminal tail is

composed by a poly(A) with a variable length (Vassetzky and Kramerov, 2013). SINEs are non-autonomous because they do not encode proteins and use LINE protein machinery for their retrotransposition (Dewannieux and Heidmann, 2005; Dewannieux et al., 2003).

The DNA transposons are subdivided into three groups based on their mode of transposition: canonical DNA transposons, Helitrons and Mavericks (**Figure 2.1**). The first group is characterized by a transposase encoded by autonomous copies and by the long terminal inverted repeats (TIRs) at both ends (Kapitonov and Jurka, 1999). The Helitrons encode the 1500 aa, so-called Rep/Hel protein, composed of the replication initiator (Rep), and helicase (Hel) conserved domains. They have conserved 5'-TC and CTRR-3' termini (R stands for A or G) and do not have terminal inverted repeats. The Mavericks, also known as Polintons, are the third class of DNA transposons and they are most complex eukaryotic transposons known so far. The Polintons are 15-20 kb long, with 6-bp TSDs and 100-1000 bp TIRs at both ends. They code for up to 10 proteins, including DNA polymerase B (POLB), integrase (INT), protease (PRO), and putative ATPase (ATP) (Kapitonov and Jurka, 2006).

2.1.2 Mobilization Mechanism of Class I transposable elements

Class I transposable elements utilize a process known as retrotransposition and it includes long terminal repeats (LTRs) and non-LTR retrotransposons (LINEs and SINEs) (**Figure 2.1 A-B**) (Finnegan, 1989). The life cycle of retrotransposons is based on the 'copy-and-paste' mechanism. Retrotransposons synthesize a mRNA, which can be translated into proteins related to the replication and also serves as the template for reverse transcription into cDNA (Boeke et al., 1985).

Reverse transcription of LTRs begins with copying the region near the 5' end of the genomic RNA into DNA using a tRNA primer, followed by a degradation of the 5' region of the template (**Figure 2.2 A**). Then, the new DNA synthesized jumps to the 3' end of the RNA template and complete the synthesis of the first DNA strand. The RNAase H degrades most of the RNA living only a small portion near the 5' end of the DNA. Then, it is used as primer to synthesize the 3' end of the second DNA strand, that jumps to the 3' end of the first DNA strand and complete the second-strand synthesis. The result is a linear double strand DNA with LTR at the end which is attacked to the target chromosomal DNA by integrase (for review see Kazazian, 2004).

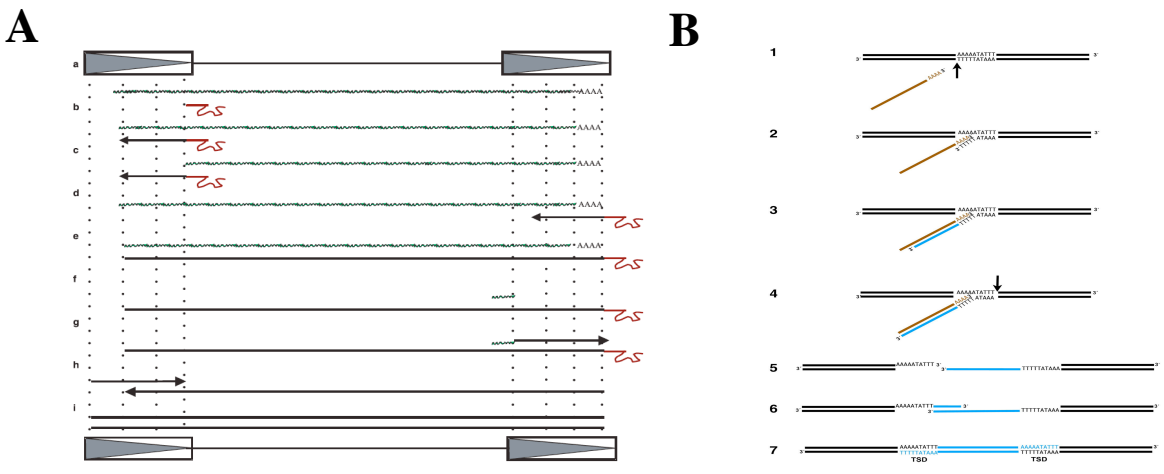

Figure 2.2 Mechanisms of retrotranspositions. (**A**) Reverse transcription of LTR retrotransposons. The donor and target locations are depicted as a bar with two boxes at the extremity. The RNA intermediate and the tRNA are shown in green and red. The DNA synthesis is depicted as a black bar with an arrow. The figure is borrowed from Kazazian, 2004. (**B**) Reverse transcription of non-LTR retrotransposons. RNA intermediate and target chromosomal DNA are shown in brown and black bars. The arrows depicted the nuclease breaks. The DNA synthesized is shown in blue and target site duplications are depicted as nucleotides. Figure is borrowed from Finnegan, 2012.

In contrast, reverse transcription of non-LTR retrotransposons occurs by a very different mechanism; this process takes place on genomic DNA through target primed reverse transcription (TPRT) (for review see Finnegan, 2012). The synthesis of LINEs

begins with a single-strand break to the 3'-OH of the target chromosomal DNA made by endonuclease (**Figure 2.2 B**). Usually, the endonuclease cuts place with an AT rich region sequence. The 3'-OH is used as a primer where the first DNA strand using the LINE RNA as a template. Then, the endonuclease makes a second break on the opposite strand of the target chromosomal DNA and the RNAase H can degrades the RNA template. The second DNA strand is synthesized using the 3'-OH at the break in the other strand of target chromosomal DNA. The result is a new integrated LINE element flanked by a target site duplication. Most copies of a LINE element are incomplete, losing the 5' end region of the coding strand, probably because reverse transcription frequently terminates before the first strand of DNA is complete. These truncated copies are clearly not able to retrotranspose.

Unlike autonomous transposons SINE elements have non coding capacity and completely rely on the cell machinery and autonomous retrotransposons for their replication. The first region contains a RNA polymerase III promoter ensuring that the element will be transcribed, while the LINE-related sequence is recognized by the proteins encoded by the corresponding LINE. Hence, SINEs are amplified using the LINE transposition machinery.

2.1.3 Mobilization Mechanism of Class II transposable elements

Class II elements can be divided into three major subclasses: ***i*** the classic 'cut-and-paste' transposons; ***ii*** those that utilize a mechanism probably related to rolling-circle replication, Helitrons; and ***iii*** Mavericks, whose mechanism of transposition is not yet well understood, but that likely replicated using a self-encoded DNA polymerase.

DNA transposons are excised (cut) from their original genomic location and inserted (pasted) into new site without involve an RNA intermediate (Jurka et al., 2007). The process starts with two transposase molecules that recognize the terminal inverted repeats (TIRs) forming the Single-End Complex (SEC) (Lipkow et al., 2004) (**Figure 2.3 A**). Both transposases cleave the 5'ends of the TIRs by hydrolysis of the phospodiester bond to liberate the non-transferred strands (5'-P extremes), which do not participate further in the transposition process. The two transposase molecules interact and bring together the transposon ends to form the Paired-End Complex (PEC) generating a transposase dimer (Richardson et al., 2009). At this point, the phosphodiester bond undergoes a hydrolysis in the 3'ends to produce the transferred strands (3'-OH extremes). The PEC binds to target DNA forming the Target Capture Complex, at which insertion takes place. The 5'end in the target DNA undergoes a nucleophilic attack from the transposon transferred strands 3'-OH. The gaps in the transposon 5'ends are filled by the host, generating canonical TSDs flanking the new transposon insertion (Muñoz-López and García-Pérez, 2010).

An other category of DNA transposon is Helitron, which transpose by rolling-circle replication mechanism (**Figure 2.3 B**). Helitron transposition starts from the cleavage of the donor and target sites by Rep proteins and the binding to the resulting 5' ends (Kapitonov and Jurka, 2001). Then the strand is displaced from its original location at the site of the break by the helicase and attached to the target break, forming a circular heteroduplex. Then, if the palindrome and 3' end of the element are correctly recognized, cleavage occurs after the CTRR-3' sequence and the one Helitron strand is transferred to the donor site where DNA replication resolves the heteroduplex. Alternatively, DNA flanking the 3' end of the element could be

transferred along with the element to the donor site. This may be how Helitrons have acquired additional coding sequences (Feschotte and Wessler, 2001).

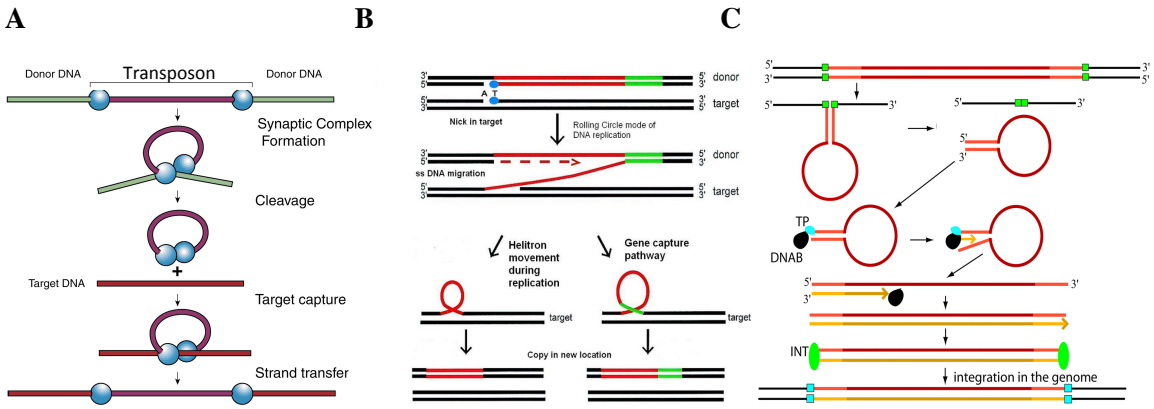

Figure 2.3. **Mechanisms of transpositions.** (**A**) Representation of the cut and paste transposition mechanism. DNA transposon is shown in purple and transposase molecules in blue circles. The donor and target locations are depicted as green and red bars. Figure is is borrowed from Muñoz-López and García-Pérez, 2010. (**B**) Model of the rolling circle transposition mechanism for the Helitron element. Helitrons is shown in red, transposase molecules in blue circles and the flanking region to the element in green. Figure is is borrowed from Feschotte and Wessler, 2001. (**C**) Model of the Polinton transposition. Polinton single-stranded DNAs are shown in red (those synthesized de novo are shown in orange); their TIRs are in light red and orange. The polymerase, terminal protein, and integrase are depicted as black, blue, and green ovals. Old and new target site duplications are marked by small green and blue rectangles. Figure is borrowed from Kapitonov and Jurka, 2006.

The Mavericks, also known as Polintons, are the third class of DNA transposons. Polinton transposition follows a completely different mechanism unseen previously in transposons through self-synthesis by their polymerase (Kapitonov and Jurka, 2006). During host genome replication, the integrase-catalyzed excision of a Polinton element from the host DNA leads to an extrachromosomal single-stranded element that forms a racket-like structure (**Figure 2.3 C**). Then, the initiation of the replication requires the terminal protein (TP) and the POLB that binds a free 5′ end and replicates the extrachromosomal Polinton, respectively. After the double-

stranded is synthesized, the integrase (INT) molecules bind its termini and catalyze its integration in the host genome.

2.1.4 Transposon activity across species

The lifecycle of a transposable element (TE) is similar to a birth-and-death process: a new TE family is born when an active copy colonizes a novel host genome and it dies when all copies in a lineage are lost or inactivated, a process which may be driven by host defence mechanisms and/or by the accumulation of disabling mutations in the TE sequence (Schaack et al., 2010). To escape extinction the TEs follow the horizontal transfer (HT) process through the transmission of genetic material between the genomes of two individuals that may belong to different species. The most extensively studied case of HT is that of DNA transposons in Drosophila and insects. The best example to demonstrate the ability and range of HT in insects is the *mariner* transposon. Maruyama and Hart found an higher similarity of mariner sequences between distantly related species (*Zaprionus tuberculatus* versus *Drosophila mauritiana*) than between closely related species within the melanogaster species group (Maruyama and Hartl, 1991). Many subsequent studies presented evidence of *mariner* elements being horizontally transferred between different insect species (Lampe et al., 2003; Lohe et al., 1995). These findings demonstrated that HT plays a crucial part in the *mariner* replication cycle, preventing its extinction by introducing it to new hosts. Cases of horizontal transfer are also found for LTR retrotransposons in plants, concerning species distantly related as palm and grapevine, tomato and bean or poplar and peach (Baidouri et al., 2014). Emerging evidence shows that HT of BovB non-LTR retrotransposons is significantly more widespread than believed. BovB is a long interspersed element (LINE) about 3.2 kb

long, originating in squamates but now found in a wide range of genomes including ruminants, marsupials, monotremes and mammals (Gentles et al., 2007; Adelson et al., 2009; Kordis, 2009; Walsh et al., 2013).

Among the Class I transposable elements, the LTR-retrotransposons constitute one of the most abundant classes of mobile genetic elements in eukaryotes; their recent activity seems highest in plants (Kumar and Bennetzen, 1999; Schnable et al., 2009). Although plants genome size may vary greatly as a consequence of polyploidy, different taxonomic groups often preserve a similar gene number, suggesting that other factors also regulate genome expansion. Indeed, LTR-retrotransposon numbers highly correlate with plant genome size and compose the largest single component of most plant genomes. Furthermore, their insertions in plants have occurred within the past few million years. For example, in *Arabidopsis thaliana*, more than 90% of LTR insertions (*Copia* family) were inserted within the past 5 million years (Pereira, 2004). The maize contains ~400 families of LTR-retrotransposons that comprise ~75% of the genome (Bowen and McDonald, 2001; SanMiguel et al., 1996; Schnable et al., 2009). All retrotransposons in maize with intact LTRs were inserted within the past 4 million years. Among the invertebrates, the majority of LTR-retrotransposons belong to the endogenous retroviral (ERV) superfamily. These make up about 8% of the human genome and approximately 10% of the mouse genome. However, the activity levels differ dramatically between rodents and humans. Indeed, active LTR-retrotransposons remain scarse in the human and mouse genomes (Gibbs et al., 2004; Zhang et al., 2008). In *Drosophila*, LTR-retrotransposons comprise ~1% of the genome including approximately 20 distinct families, while they account for only

0.4% of the *C. elegans* genome although it has been documented the activity of full-length elements (Finnegan, 1992; Ganko et al., 2003).

Non-LTR retrotransposons are widespread among eukaryotes, but have been especially prolific in mammalian genomes (Treangen and Salzberg, 2012). For example, LINE-1 (L1) elements and the non-autonomous SINEs comprise approximately 45% of human genome. L1 is the most common superfamily of autonomous retrotransposons in mammals. Mice and rats show extensive L1 activity (Akagi et al., 2008). In contrast, bats have drawn special attention due to their completely opposite transposon activity pattern compared to neighboring species in the phylogenetic tree (Cantrell et al., 2008; Grahn et al., 2005). Infact, L1s appear to have gone extinct in megabats and sigmodontine rodents. Non-LTR retrotransposon activity is minimal in plants (Huang et al., 2008; Ziolkowski et al., 2009). Comparison of subspecies pairs either in rice or Arabidopsis showed that recent transposition frequencies are two or three orders of magnitudes lower than for other transposons, including LTR retrotransposons.

Among the Class II transposable elements, the DNA transposons include the first mobile elements discovered in maize using cytogenetic methods by Barbara McClintock in the 1940s (McClintock, 1950). Since then, DNA transposons have been found to be almost ubiquitous among both prokaryotes and simpler eukaryotes. Their activity appears to be extinct in most mammals, which fuelled speculation that DNA transposons play a limited role in the ongoing evolution of mammalian genomes. However, recent studies suggest that DNA transposons, namely non-autonomous hobo/Activator/TAM (nhAT transposons and helitrons), are active in certain bat species (Ray et al., 2008). The DNA transposons comprise the 12% of the *C. elegans* genome, where the Tc1 is the most abundant element (Fischer et al., 2003). The P

element is the best-characterized active DNA transposon in *D. melanogaster* (Bingham et al., 1982). It is known to create phenotypes, and therefore is frequently used in mutagenesis screens.

Helitrons were discovered in *C. elegans*, rice and *Arabidopsis* genomes (Kapitonov and Jurka, 2007). They are well studied in maize where the first examples of de novo Helitron insertions were isolated from two mutants (Gupta et al., 2005; Lal et al., 2003). Helitrons comprise the 3% of the brown bats genomes of little brown bats (Pritham and Feschotte, 2007). They have amplified during the past 30-36 million years, therefore, most copies are substantially older than their counterparts in other eukaryotes (it has been estimated that most of the maize Helitrons transposed <250,000 years ago).

Polintons are widespread in the genome of many eukaryotes (Kapitonov and Jurka, 2006). They are comprised in protists, fungi and many animals, including trichomonas, sea urchin, frog, zebrafish, nematode, turtle, and insects but they are not found in plants. Currently, the database of eukaryotic repetitive DNA elements (Repbase) contains consensus sequences and information of more than 70 Polinton elements from 28 organisms (Haapa-Paananen et al., 2014).

Transposable elements have a great impact on genome evolution by changing the whole genomes size. Interestingly, among plants the genome size ranges from 63 Mbp in *Genlisea margaretae* (Greilhuber et al., 2006) to more than 120,000 Mbp in *Fritillaria assyriaca* (Ambrozová et al., 2011). Significant genome size variations are also common even within a single family such as *Oryza sativa* (rice) and *Zea mays* (maize) with 430 Mbp and 2300 Mbp, respectively (Goff et al., 2002; Schnable et al., 2009). In plants this variation is also due to polyploidization. Indeed, across the

Oryzia genus it has been attributed to both polyploidization and LTR-retrotransposon proliferation (Zuccolo et al., 2007). The proportion of TEs in plant genomes is variable and could make up to 85% in maize and 65% in rice (Bennetzen, 2005; Zuccolo et al., 2007).

In the animal kingdom the proportion of TEs may vary up to 77% as in the frog *Rana esculenta* which has a genome of 5.5-7.8 GB (Biémont and Vieira, 2006). The salamanders have the largest genomes in vertebrates, with ranges from 14 to 120 GB (Sun et al., 2012). The LTR retroelements contribute to genomic gigantism in both these amphibians (Sun et al., 2012; Grau et al., 2014). The human genome is about 3.6 GB and it includes around 45% of TEs, mainly represented by the non-LTR element LINEs (21%) (Treangen and Salzberg, 2012).

In addition to the global expansion of genome size due to TE colonization of the host genome, TEs affect the genome structure by facilitating chromosomal sequence rearrangements. TEs were originally discovered because of their potential to cause chromosomal mutations such as deletions, translocations and inversions in maize (McClintock, 1987). However, genome rearrangements mediated by TEs have been generalized to all organisms including plants (Bennetzen et al., 2005; Parisod et al., 2010), vertebrate and invertebrates animals (Miller et al., 1992; Bourque et al., 2008). In human, the abundance of LINEs (L1) and SINEs (Alu) elements are often found in the vicinity or even within the brakeage points of chromosomal rearrangements (Zhao and Bourque, 2009). The influence of such elements is visible in several diseases characterized by chromosomal rearrangement such as different types of cancer (Konkel and Batzer, 2010). For example, Alu sequences were found in the chromosome breakpoints producing the translocation between human chromosome 9 and 22 (called the Philadelphia chromosome) that leads to chronic myeloid

leukaemia development (Jeffs et al., 1998). Finally, homologous recombination between Alu elements causes the duplication of the MYB proto-oncogene, which encodes an essential transcription factor, and leads to acute T cell lymphoblastic leukemia (O'Neil et al., 2007).

These evidences show how the TEs can impact the evolution of genome structure and trigger chromosome rearrangements, some of which can induce disease.

2.1.5 Transposable elements and Central Nervous System

Transposition is essential for survival of transposable elements, ensuring that are not lost by chance or eroded by mutation. Previous studies have indicated that transposition can occur only in germ cells or during early embryogenesis, before the germ line becomes a distinct lineage (Ostertag et al., 2002; Prak and Kazazian, 2000). However a cultured cell retrotransposition assay has revealed that human and mouse L1 elements can retrotransposase in a variety of transformed or immortalized cell lines (Han et al., 2004; Moran et al., 1996; Morrish et al., 2002). Muotri and coworkers show L1 retrotranspositions in mouse neuronal precursor cells (Muotri et al., 2005). The L1 retrotransposition was also observed in the hippocampus and cerebellum of adult human brains when compared with the heart or liver genomic DNAs (Baillie et al., 2011). These results raised an interesting possibility that those events can also occur in the cells that cannot transmit genetic information thereby generating somatic mosaicisms.

A recent study demonstrated that such scenario is not restricted to mammals. Indeed, retrotransposition events are conserved in certain neurons (αβ neurons) of the mushroom bodies (MBs) in the *Drosophila* brain (Perrat et al., 2013). Interestingly, MB neurons and mammalian hippocampus are critical regions for learning and

memory. Therefore, a conserved mechanism could contribute to increased retrotransposition levels in brain regions involved in learning and memory processes. The activation of transposition seems to be regulated through epigenetic mechanisms, such as DNA methylation and histone modifications. When such mechanisms are repressed, retrotransposition events can be activated. The first example of control of retrotransposition comes from L1 expression in neurogenesis. The canonical L1 promoter contains binding sites for SOX2 (Tchénio et al., 2000), YY1 (Yin Yang 1) (Athanikar et al., 2004), RUNX3 (runt-related transcription factor 3) (Yang et al., 2003) and TCF and LEF (WNT signalling pathway transcription factors) (Kuwabara et al., 2009), which are transcription factors that are known to be involved in neurogenesis. In neural stem cells, L1 promoter is repressed by DNA methylation, trimethylation of lysine 9 on histone H3 (repressive histone mark), MECP2 binding (methyl-CpG-binding protein 2) and SOX2 (Muotri et al., 2010). As neural stem cells differentiate this repression is decreased. The L1 promoter undergoes an open chromatin state becoming demethylated. At the same time, the WNT transcription factors, ß-catenin and members of the TCF/LEF family, with the cooperation of YY1 lead to the activation of L1 expression (Kuwabara et al., 2009).

The transposition can also be regulated by small RNAs. In the fly brain, loss of the piwi-interacting RNA (piRNA) proteins result in an increased expression of LTR elements, LINE-like elements and DNA transposons (Li et al., 2013; Perrat et al., 2013). The piRNA proteins, Aubergine and Argonaute 3, are usually less abundant in the αβ neurons of MBs and well correlating with an increase of transposable elements. The role of piRNAs has also been described in the control of memory-related genes expression in *Aplysia* neurons (Rajasethupathy et al., 2012). These results demonstrate that transposition can generate cells with unique genome, which

might in turn generate distinct transcriptional outputs. Interestingly, the effects of increased somatic retrotransposition in neurons suggest expression changes of neighbouring genes. Indeed, brain-specific insertions into genes that are important for neural function, including those encoding dopamine receptors and neurotransmitters, have been identified in both human and *Drosophila* (Baillie et al., 2011; Perrat et al., 2013). However, those insertions do not show a clear functional impact. Thus, it has been suggested that somatic retrotranspositions occur to generate a neuronal diversity in the brain.

Although a controlled level of retrotransposition may be beneficial for neuronal genomes, upregulated transposition mechanisms may also have deleterious effects on cognitive functions. Evidence from studies in patients and model organisms suggest that transposon misregulation may occur in various neurological disorders, including neurodegeneration and mental disorders (Bundo et al., 2014; Muotri et al., 2010). It has been observed both altered expression levels and increased numbers of somatic insertions. However, whether and how transposable elements misregulation can directly cause neurological disorders is unknown so far.

2.2 Long non-coding RNAs

2.2.1 Identification and functions

Although the central role of RNA in cellular functions and organismal evolution has been advocated periodically during the past 50 years, RNA has been largely relegated to an intermediate between gene and protein, encapsulated in the central dogma 'DNA makes RNA makes protein' (Mercer et al., 2009). However, the finding

that only a small proportion of the genome codifies for proteins, whereas the vast majority of the genome is transcribed into what was previously defined as "dark matter", non-coding RNAs (ncRNAs), challenges this assumption and suggests that RNA has continued to evolve and expand alongside proteins and DNA (Carninci et al., 2005; Johnson et al., 2005; Okazaki et al., 2002). This transcriptional class can be classified in small (sncRNAs) and long (lncRNAs) non-coding RNAs. The lncRNAs were first described during the large-scale sequencing of full-length cDNA libraries in mouse (Okazaki et al., 2002). Their identification can be difficult because there is no a biological argumentation widely accepted in the community. Currently, the transcripts are arbitrarily considered lncRNAs if they are longer than 200 nucleotides but that appear to lack protein-coding potential (Cabili et al., 2011; Derrien et al., 2012). According to Taft, this size cut-off distinguishes lncRNAs from small ncRNAs such as microRNAs (miRNAs), small nucleolar RNAs (snoRNAs), piwiRNAs (piRNAs) (Taft et al., 2010). Even though the lncRNAs are observed in a large diversity of species, they are poorly conserved when compared with the well-studied coding and sncRNAs, such as mRNAs, miRNAs, snoRNAs (Baker, 2011). Although, the function of most lncRNAs is unknown, there are an increasing number of publications suggesting they play roles in negatively or positively regulating gene expression in development, differentiation and human disease (Huarte and Rinn, 2010; Taft et al., 2010; Wilusz et al., 2009). LncRNAs are enriched in the nucleus, however some significant lncRNA-mediated mechanisms of gene regulation have also been identified in the cytoplasm. Besides, it is possible to distinguish between *cis*-acting lncRNAs – that control the expression of genes in the vicinity of their transcription sites – and *trans*-lncRNAs – influencing genes at independent loci – (Batista and Chang, 2013; Guttman and Rinn, 2012; Rinn and Chang, 2012). The major evidence of *cis*-regulation in the nucleus

came from studies of X inactivation by the Xist lncRNA in female mammals during development (Wutz, 2011). Xist induces the formation of repressive heterochromatin, at least in part, by tethering Polycomb Repressive Complex 2 (PRC2) to inactive X chromosome. The interaction between Xist and chromatin may involve transcriptional repressor protein YY1 (Yin Yang 1) that is thought to function as a recruitment platform for Xist by binding its first exon (Jeon and Lee, 2011). Moreover, it has been recently shown that Xist can be regulated by its natural antisense non-coding transcript Tsix (Lee and Bartolomei, 2013). Another well-studied is the *cis*-acting lncRNA HOXA distal transcript antisense RNA (HOTTIP). It was identified in primary human fibroblasts and its down-regulation induced the transcription of several downstream HOXA genes. It is transcribed upstream to the 5' end of the HOXA cluster, where it binds adaptor proteins and sets nearby chromatin marks to drive transcription (Wang et al., 2011). Hox transcript antisense RNA (HOTAIR) was one of the first *trans*-acting lncRNAs to be identified (Rinn et al., 2007). HOTAIR is expressed from the HOXC cluster and was shown to repress transcription of the HOXD cluster, which is located on a different chromosome. HOTAIR interacts with PRC2 and is required for repressive histone H3 lysine-27 trimethylation (H3K27me3) of the HOXD cluster (Ng et al., 2012). A substantial proportion of lncRNAs has been identified in the cytoplasm where they can modulate protein localization, mRNA translation and stability. For example, NRON lncRNA regulates the trafficking of the NFAT (nuclear factor of activated T cells) transcription factor from the cytoplasm to the nucleus to activate target genes in response to calcium-dependent signals (Willingham et al., 2005). An example of lncRNA that modulate translational control is the ubiquitin carboxy-terminal hydrolase L1 antisense RNA 1 (Uchl1-as1) (Carrieri et al., 2012). It is complementary to the *UCHL1* mRNA that is under the control of

stress signaling pathways. An increasing of *UCHL1* protein is associated to the shuttling of the antisense from the nucleus to the cytoplasm. Similarly, lncRNAs can modulate mRNA stability; b-site APP-cleaving enzyme 1-antisense (BACE-AS) is transcribed from the opposite strand to *BACE1* and regulates its expression levels by increasing BACE1 mRNA stability (Faghihi et al., 2008). Other roles have been identified for lncRNAs in establishing and maintaining cell-type-specific gene expression patterns during organ development (Qureshi and Mehler, 2012). Particularly, the central nervous system (CNS) shows intense transcription of lncRNAs that have functions in neurogenesis. A set of lncRNAs was predicted to be transcribed in the developing mouse central nervous system. They show a preference to be located adjacent to transcription factor genes and thus may regulate their transcription (Ponjavic et al., 2009). Among the lncRNAs of the nervous system, metastasis-associated lung adenocarcinoma transcript 1 (Malat1) regulates synapse formation by modulating a subset of genes that have roles in nuclear and synapse function (Bernard et al., 2010). Indeed, the knockdown of Malat1 in cultured mouse hippocampal neurons produced decreased synapse density and decreased dendrite growth. Recently, it has been described a vertebrate-conserved and central nervous system-expressed lncRNA termed Pauper (Vance et al., 2014). It is transcribed from a locus upstream of the gene encoding the *Pax6*, a transcription factor required for eye and diencephalon specification. Knockdown of Paupar disrupts the normal cell cycle profile of neuroblastoma cells and induces neural differentiation. LncRNAs are observed in a large diversity of species, including animals, plants, and yeast (Brown et al., 1992; Houseley et al., 2008; Swiezewski et al., 2009). In *Danio rerio*, several hundreds of lncRNAs have been identified to be expressed during the development. Two of them were (Cyrano and Megamind) resulted to be expressed during the

embryogenesis and their knockdown induced neural and brain development defects, respectively (Ulitsky et al., 2011). In *Arabidopsis thaliana*, the COLDAIR lncRNA is required for the flowering process by blocking the expression of genes (FLC) through targeting repressive chromatin marks (Ietswaart et al., 2012). The basic function, the class and site of action of lncRNAs discussed so far are summarised in the **Table 2.1**.

Table 2.1 Summary of lncRNAs with known function and site of action.

Name	Function	Class	Site of action	Organism	Reference
Xist	X chromosome inactivation	cis-acting	Nucleus	Human, mouse	Wutz, 2011
Tsix	Xist inactivation	cis-acting	Nucleus	Human, mouse	Lee and Bartolomei, 2013
HOTTIP	Organism development and angiogenesis	cis-acting	Nucleus	Human	Wang et al., 2011
HOTAIR	Organism development	trans-acting	Nucleus	Human	Rinn et al., 2007
NRON	Immune response and developmental processes	trans-acting	Cytoplasm	Human, mouse	Willingham et al., 2005
Uchl1-as1	Brain development, neurodegenerative disease	cis-acting	Cytoplasm	Mouse	Carrieri et al., 2012
BACE-AS	Alzheimers disease	cis-acting	Cytoplasm	Human	Faghihi et al., 2008
MALAT1	Cell cycle control, cancer metestasis	trans-acting	Cytoplasm	Human, mouse	Bernard et al., 2010
Pauper	Proliferation and neuronal differentiation	cis- and trans-acting	Nucleus	Human, mouse	Vance et al., 2014
Cyrano	Organism development	Not-analyzed	Not-analyzed	Zebrafish	Ulitsky et al., 2011
Megamind	Organism development	Not-analyzed	Not-analyzed	Zebrafish	Ulitsky et al., 2011
COLDAIR	Control of flowering time	cis-acting	Nucleus	*Arabidopsis thaliana*	Ietswaart et al., 2012

2.3 Transcriptomic impact of repetitive elements and long non-coding RNAs

In a recent study Kelley and Rinn highlighted an intriguing relationship between transposable elements (TEs) and long non-coding RNAs (lncRNAs) in mammals genomes (Kelley and Rinn, 2012). A large proportion (41%) of human lncRNAs are derived from TEs, and the majority of lncRNAs (83%) contain at least one TE

fragment (**Figure 2.4**). In contrast, only a small portion (6%) of protein coding sequences is in overlap with TEs.

As a consequence, many mature lncRNA transcripts contain combinations of multiple repeat fragments reminiscent of protein domain structures. Besides, human endogenous retroviruses (hERV) are strongly enriched in lncRNAs compared to the genomic background. In contrast, families including the highly numerous Alu, L1 and L2 classes are significantly depleted from lncRNA.

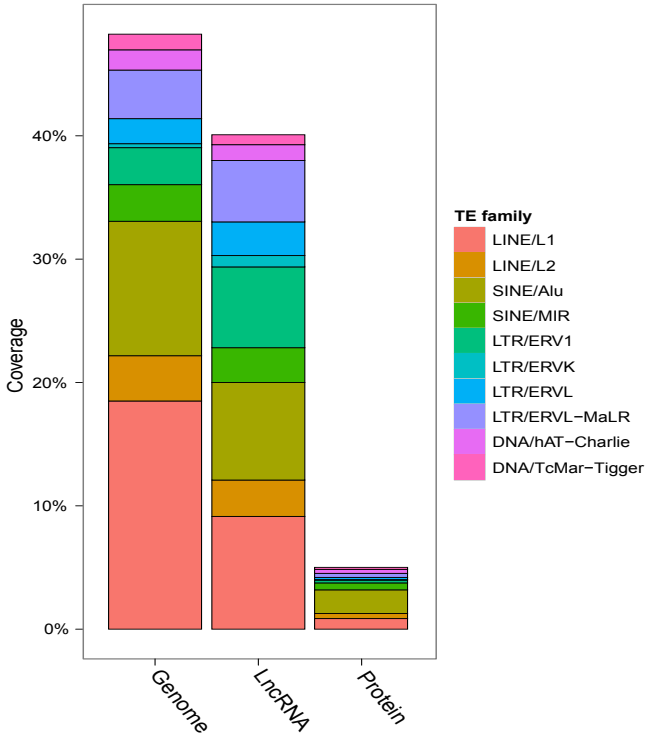

Figure 2.4 Transposable element composition of human genome. TEs compose less lncRNA sequence than genomic background but much more than protein coding genes. Figure is modified from Kelley and Rinn, 2012.

These patterns suggest that the presence of TE fragments within mature lncRNA sequence might have been selected for or against during evolution. The lncRNAs promoted by hERV elements are specifically up-regulated in pluripotent cell types, which is consistent with previous observations of the overexpression of these elements in human embryonic stem cells (Santoni et al., 2012).

Another study reported the presence of TEs in lncRNA exons of human (75%), mouse (68%) and zebrafish (66%) more than that expected by chance and the TEs were more likely to overlap a transcriptional start site in a lncRNA as compared to a coding gene (Kapusta et al., 2013). The longest studied lncRNAs, Xist, appear to have at least three repeat-regions with distinct functionalities. Silencing capability is dependent on the presence of A-repeat region, which has most likely originated as an endogenous retrovirus, ERVB5, and it is conserved across eutherians (Elisaphenko et al., 2008; Wutz et al., 2002). The A-repeat domain interacts with the repressive PRC2 complex to repress chromatin (Zhao et al., 2008). The localization of Xist seems to be dependent on the C-repeat region that is murine-specific, having homology to another endogenous retrovirus (ERVB4, Elisaphenko et al., 2008). This region interacts with the transcription factor, YY1, that directs XIST to specific genomic loci through DNA binding (Jeon and Lee, 2011). Most recently, it has been shown the F-repeat region of Xist that has originated from a DNA transposon and it is critical for Jarid2 (jumonji, AT rich interactive domain 2) recruitment (Elisaphenko et al., 2008). Jarid2 is implicated in the initial Xist-induced targeting of PRC2 complex to the inactive X chromosome (da Rocha et al., 2014). Thus, the functionalities of Xist appear to have evolved from three distinct transposable elements. The antisense lncRNA Uchl1-as1 contains an embedded SINEB2 element that stimulates the neuron-specific Uchl1 translation through a mechanism that remains still unclear (Carrieri et al., 2012). Removal of SINE2B element completely abrogated the translational effect of the transcript. Other antisense lncRNAs also bind their sense: BACE1-as binds to and increases the stability of BACE1 mRNA, whereas another neural antisense transcript, BDNFOS, negatively regulates BDNF mRNA (Faghihi et al., 2008; Lipovich et al., 2012). Both of these transcripts contain multiple TE insertions, although these have

not yet been strictly linked to any function. A recent study of the coronary artery disease-associated lncRNA, ANRIL (antisense non-coding RNA in the INK4 locus), showed that SINE (Alu) elements within its sequence were necessary for its biological activity; and loss of embedded Alu reversed ANRIL's promotion of growth, adhesion, and motility in cellular models (Holdt et al., 2013). Another TE-derived DNA-binding domain has been identified in the mouse lncRNA, Fendrr, which is necessary for mouse heart and body wall development. Inspection of the putative DNA-binding domain of Fendrr, shows that it is derived from a LINE1 element and binds to at least two gene promoters to which it recruits various chromatin-remodeling factors (Grote et al., 2013). A recent study on a rare neurodegenerative condition, the infantile encephalopathy, which is restricted to a small population from the island of Reunion, further support the significance of TE-derived lncRNA (Cartault et al., 2012). The authors identified a novel lncRNA associated with L1 element that shows a brain specific expression. In-vitro knockdown of this lncRNA induced apoptosis, suggesting that its expression is necessary for neuronal survival. Several other repeat-rich lncRNAs have been described. The transcript linc-RoR which modulated reprogramming of fibroblasts to a pluripotent state, is almost entirely composed of TE-derived sequence from seven different TE families and has an ERV at its transcription start site (TSS) (Kelley and Rinn, 2012). In mouse, a brain-specific non-coding transcript (AK046052), regulated by the master neural transcriptional repressor REST, is largely a mosaic of TE-derived sequence (Johnson et al., 2009). Other examples of lncRNAs largely composed of TEs are TUG1 (taurine up-regulated gene 1), interacts with methylated Polycomb 2 protein to modulate its recognition of histone modifications and BANCR (BRAF-regulated lncRNA 1), regulates melanoma cell migration (Flockhart et al., 2012; Young et al., 2005). Overall, we might conclude

that TE sequence within lncRNA is the rule rather than the exception, and it is strictly necessary for lncRNA's biological activity. The functional transposable element sequences within lncRNA discussed so far are summarised in the **Table 2.2**.

Table 2.2 Known cases of functional element sequences within lncRNA.

TE	LncRNA	Reference
ERVB5	Xist (A-repeat)	Elisaphenko et al., 2008; Wutz et al., 2002
ERVB4	Xist (C-repeat)	Elisaphenko et al., 2008
DNA transposon	Xist (F-repeat)	Elisaphenko et al., 2008
SINEB2	Uchl1-as1	Carrieri et al., 2012
Alu	BDNFOS	Lipovich et al., 2012
Not-specified	BACE-AS	Faghihi et al., 2008
Alu	ANRIL	Holdt et al., 2013
LINE1	Fendrr	Grote et al., 2013
LINE1	SLC7A2-IT1	Cartault et al., 2012
ERV	linc-RoR	Kelley and Rinn, 2012
Multiple TE families	AK046052	Johnson et al., 2009
Multiple TE families	TUG1	Young et al., 2005
Multiple TE families	BANCR	Flockhart et al., 2012

Chapter 3 – Genomic toolkit for Cephalopods

3.1 Genomic toolkit for Cephalopods

Until very recently, the genome-scale analyses have favoured the sequencing of deuterostomes among bilateria animals (Telford and Copley, 2011). In particular, many genome projects were primarily focused on vertebrates with an expanding study of other chordates and selected non-chordates such as sea urchin and hemicordates (Freeman et al., 2012; Sodergren et al., 2006). Among the protostomes, ecdysozoans, such as fruit fly (*Drosophila melanogaster*) and roundworm (*Caenorhabditis elegans*) genomes have been sequenced (Adams et al., 2000; C. elegans Sequencing Consortium, 1998). In contrast, there has been far less genomic analysis of Lophotrochozoans, with genomes published for only a handful of organisms, including three trematode parastic worms (*Schistosoma mansoni*, *S. japonicum* and *S. haematobium*) (Berriman et al., 2009; Young et al., 2012; Zhou et al., 2009), two annelids (*Capitella teleta*, *Helobdella robusta*) (Simakov et al., 2013) and four molluscs (*Pinctada fucata*, *Lottia gigantea*, *Crassostrea gigas* and *Aplysia californica*) (Simakov et al., 2013; Takeuchi et al., 2012; Zhang et al., 2012a; Broad Institute, 2009) but no one cephalopod genome has been sequenced so far (Berriman et al., 2009; Simakov et al., 2013; Takeuchi et al., 2012; Young et al., 2012; Zhang et al., 2012). The genomes of cephalopods are known to be larger and more repeat-rich than any previously sequenced metazoan genomes (Yoshida et al., 2011). Currently, studies regarding cephalopods have been restricted at the transcriptome level. EST (Expressed Sequence Tag) collections were performed for *Euprymna scolopes*, *Loligo pealei* and *Sepia officinalis* during embryonic and the adult stages (Chun et al., 2006; DeGiorgis et al., 2011; Bassaglia et al., 2012) (**Table 3.1**). Other transcriptome studies

have been published on the *O. vulgaris* using both EST library and RNA-sequencing (RNA-seq) approaches. The first approach has been used to perform comparative study of ESTs from the Octopus eye with genes in the human eye (Ogura et al., 2004). They sequenced 1,052 nonredundant ESTs that have matches with protein database. Comparing these ESTs with those of the human eye, 729 (69.3%) genes were commonly expressed between the human and octopus. Ogura and co-workers also compared Octopus-eye ESTs with some deuterostome genomes, founding that 1,019 out of the 1,052 genes already existed at the common ancestor of bilateria, and 875 genes were conserved between human and octopus thereby demonstrating molecular convergent evolution of this organ between the species. Subsequently, the RNA-seq approach was used for molecular evolutionary studies among several mollusc transcriptomes (Kocot et al., 2011; Smith et al., 2011). Kocot and Smith studies have respectively generated 20 and 14 transcriptomes from all major lineages of the phylum mollusca, including the *Octopus vulgaris*. They assembled 9,507 and 40,808 sequences from *Octopus vulgaris* neural and arm tissues, respectively. The purpose of these studies was to identify orthologous genes suitable for understanding the molluscan phylogeny. Although they used two independent datasets, a strong concordance was observed between them. Their results supported evolutionary framework for Mollusca consisting of two major clades: Aculifera, which includes a monophyletic Aplacophora sister to Polyplacophora, and Conchifera including Cephalopoda that represents the sister of all other conchiferan lineages (Pleistomollusca and Scapholopoda). Smith also includes Monoplacophora that unexpectedly reveals to be the sister group to Cephalopoda and not to other Conchifera.

More recently, the transcriptome profiling was used to study the central nervous system (Zhang et al., 2012b) and the hemocytes (Castellanos-Martínez et al., 2014) of the *O. vulgaris* (**Table 3.1**). Zhang and co-workers generated 59,859 contigs, including 31,909 sequences with length greater than 200 bp. Functional annotation analysis identified 10,412 (17.4%) sequences of the *Octopus* central nervous system. They compared those transcripts with CNS databases of different organisms demonstrating that the *Octopus* has a large number of homologous genes in both vertebrates and invertebrates. The authors also identified in the transcriptome all the important genes involved in adherens and tight junctions, suggesting that *Octopus vulgaris* may have a vertebrate-like Blood-Brain Barrier. Lastly, it has been assembled a transcriptome of hemocytes from health and silk octopuses infected by a gastrointestinal parasite (*Aggregata octopiana*). The transcriptome was assembled in 254,506 contigs, containing 87,408 sequences greater than 500 bp. A total of 48,225 (18.9%) contigs have matches with protein database. The authors revealed 538 transcripts differential expressed between health and silk octopuses. Those transcripts are involved in signalling pathways important for immune response such as nuclear factor-kB (NK-kB), Toll-like receptors (TLR) and apoptosis.

Table 3.1 Current available cephalopod transcriptome datasets.

Species	Tissues	Sequences	Technology	Source
Euprymna scolopes	Juvenile light organs	13,962	Sanger	Chun et al., 2006
Loligo pealei	Stellate ganglia	10,027	Sanger	DeGiorgis et al., 2011
Sepia officinalis	Whole embryos	19,780	Sanger	Bassaglia et al., 2012
Octopus vulgaris	Adult eye	1,052	Sanger	Ogura et al., 2004
Octopus vulgaris	Neural	9,507	454 Roche	Kocot et al., 2011
Octopus vulgaris	Arm	40,808	Illumina	Smith et al., 2011
Octopus vulgaris	Central Nervous System	59,859	Illumina	Zhang et al., 2012b
Octopus vulgaris	Hemocytes	254,506	Illumina	Castellanos-Martínez et al., 2014

Due to the lack of molecular tools for the cephalopods, it has been established the Cephalopod Sequencing Consortium (CephSeq Consortium), with the intention of using strategic genomic and transcriptomic sequencing of key cephalopod species to address unanswerable questions about the mainly research areas of cephalopod biology, from neuronal function at the cellular and systems levels to cephalopod population dynamics to the evolution of gene regulatory elements mediating body plan variation. Researches in the CephSeq Consortium will start to sequence the genomes of 10 organisms with the aim to provide molecular tools for cephalopod biology to move from the pre-genomic to the post genomic age (Albertin et al., 2012).

Chapter 4 – Materials and Methods

4.1 *In-silico* studies

4.1.1 Collection of public transcriptomes

In order to perform comparative analysis, I downloaded a broad representation of transcriptomes for multicellular organisms, containing both vertebrate and invertebrate species. Thirty-nine transcriptomes have been selected belonging to the organisms belonging to seven different phyla. The dataset corresponds to at least 10,000 unique assembled transcripts, to avoid bias given by transcriptome incompleteness (**Table 4.1**). Chordates, arthropods, echinoderms, annelida and nematodes transcriptomes were downloaded from UniGene (http://www.ncbi.nlm.nih.gov/unigene). Available assembled mollusc and brachiopod transcriptomes were downloaded from Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.24cb8) (Smith et al., 2011). *Crassostrea gigas* transcriptome was downloaded from GigaDB (http://gigadb.org/dataset/100030) (Zhang et al., 2012). *Sepia officinalis*, *Loligo pealei* and *Euprymna scolopes* expressed sequence tags (ESTs) were downloaded from NCBI dbEST (http://www.ncbi.nlm.nih.gov/dbEST/) using the following keywords: LIBEST_027716, LIBEST_027407, txid6613. The EST sequences were cleaned of vector using Cross_match (v 0.990329; http://www.phrap.org) and CAP3 was used with default parameters to cluster and assemble the transcripts (Huang and Madan, 1999).

This collection of thirty-nine transcriptomes was filtered out for sequences shorter than 200 bp and used as input sequence data to predict the noncoding potential using Portrait software (v1.1) (Arrial et al., 2009). Sequences are translated

by the software and are evaluated for noncoding potential by a support vector machine (SVM) on the *ab-initio* properties of longest ORF and the nucleotide composition. The software reports as output a probability score for a transcript to be non-coding. The long non-coding RNAs (lncRNAs) were selected with a non-coding score greater than 0.95 (95%).

Table 4.1. List of transcriptomes utilized. The data are arranged by phyla and species included (both in alphabetical order). In the header, **Name:** scientific name of the organisms; **Phylum:** taxonomic identification; **N. assembled transcripts:** number of transcripts assembled; **N. transcripts ≥ 200 bp:** number of transcripts filtered on the length; **Source:** origin of the data, i.e. public database and version of the relative dataset.

Name	Phylum	N. assembled transcripts	N. transcripts ≥ 200 bp	Source
A. pompejana	Annelida	14191	14177	UniGene_Build#2
D. melanogaster	Arthropoda	17168	16809	UniGene_Build#75
L. vannamei	Arthropoda	72348	72200	UniGene_Build#5
L. anatina	Brachiopoda	14861	14396	Dryad Digital Repository
B. floridae	Chordata	15165	15086	UniGene_Build#13
C. intestinalis	Chordata	29522	28578	UniGene_Build#28
D. rerio	Chordata	53558	52610	UniGene_Build#126
G.. aculeatus	Chordata	16728	16710	UniGene_Build#7
G.. gallus	Chordata	33850	33590	UniGene_Build#46
G.. morhua	Chordata	41275	41225	UniGene_Build#12
H. sapiens	Chordata	130056	125876	UniGene_Build#236
M. fascicularis	Chordata	12569	12519	UniGene_Build#19
M. mulatta	Chordata	26849	26774	UniGene_Build#18
M. musculus	Chordata	80240	78398	UniGene_Build#194
O. latipes	Chordata	21803	21499	UniGene_Build#30
O. niloticus	Chordata	17426	17179	UniGene_Build#2
P. anubis	Chordata	11659	11601	UniGene_Build#7
X. laevis	Chordata	31306	30817	UniGene_Build#94
S. purpuratus	Echinodermata	14718	14690	UniGene_Build#20
A. californica	Mollusca	24709	24083	UniGene_Build#9
A. entalis	Mollusca	12827	11145	Dryad Digital Repository
C. apiculata	Mollusca	24362	21451	Dryad Digital Repository
C. gigas	Mollusca	28027	27336	GigaDB
E. scolopes	Mollusca	17138	17042	Taxonomy Browser Database
G.. neomeniomorph	Mollusca	59536	50823	Dryad Digital Repository
G.. tolmiei	Mollusca	78360	69221	Dryad Digital Repository
L. gigantea	Mollusca	15623	15583	UniGene_Build#2
L. hyalina	Mollusca	11917	11549	Dryad Digital Repository
L. littorea	Mollusca	27651	24529	Dryad Digital Repository
L. pealei	Mollusca	10242	9578	EST Database
N. expansa	Mollusca	64402	56730	Dryad Digital Repository

Name	Phylum	N. assembled transcripts	N. transcripts ≥ 200 bp	Source
N. megatrapezata	Mollusca	31371	27503	Dryad Digital Repository
N. pompilius	Mollusca	31771	30257	Dryad Digital Repository
P. lucaya	Mollusca	22008	20450	Dryad Digital Repository
S. officinalis	Mollusca	18712	18055	EST Database
S. pectinata	Mollusca	33154	30788	Dryad Digital Repository
S. velum	Mollusca	27319	25769	Dryad Digital Repository
Y. limatula	Mollusca	14726	12995	Dryad Digital Repository
C. elegans	Nematoda	23151	22673	UniGene_Build#52

4.2 Generation of the transcriptome

4.2.1 RNA extraction, sequencing and quality filtering

Three adult individuals of *Octopus vulgaris* were collected through local fisherman from the Bay of Naples (Southern Tyrrhenian Sea, Italy). Animals were humanely-killed according to available standardized procedures (Grimaldi et al., 2007; Andrews et al., 2013) of G. Fiorito research group at the Stazione Zoologica Anton Dohrn. Sampling occurred late summer of 2012. For each animal, total RNA was isolated from the supra- (SEM) and sub- (SUB) esophageal masses, the optic lobes (OL) and the arms (ARM) using TRIzol reagent (Life Technologies) according to the manufacturer's protocol. The total RNA of the arms includes both muscle and nerve tissues. Contaminating DNA was degraded by treating samples with Turbo DNase Kit (Ambion) according to the instruction manual. RNA was checked for quantity and quality using NanoDrop (Thermo-Fisher) and RNA BioAnalyzer chip (Agilent Technologies, Santa Clara, CA, USA). Paired-end libraries were prepared using the Illumina TruSeq RNA sample library preparation kit (Illumina, San Diego, CA, USA). Each sample was barcoded and all samples were pooled in two lanes of the Illumina HiSeq 2000 platform (2x50bp read length).

The quality of the raw reads obtained from the sequencing platform was assessed using FastQC tool (v0.10.1). Raw reads were filtered and trimmed based on quality and adapter inclusion using Trimmomatic (Bolger et al., 2014) (parameters: -threads 24 -phred 64 ILLUMINACLIP: illumina_adapters.fa:2:40:15 LEADING: 3 TRAILING: 3 SLIDINGWINDOW: 3:20 MINLEN: 25). Trimmed and filtered reads were normalized using the normalize_by_kmer_coverage.pl script from the Trinity software (Release 2013-08-14) (Grabherr et al., 2011) (parameters: --seqType fq --JM 240G --max_cov 30 --JELLY_CPU 24). Only the read pairs with both members passing the quality-filtering test were further considered.

4.2.2 *De novo* assembly and annotation of assembled transcripts

The transcriptome was assembled using Trinity on the trimmed, filtered and normalized reads using the Jaccard clip (parameters: --seqType fq --JM 240G --inchworm_cpu 24 --bflyHeapSpaceInit 24G --bflyHeapSpaceMax 240G --bflyCalculateCPU --CPU 24 --jaccard_clip --min_kmer_cov 2). Trinity is a next generation sequencing (NGS) transcriptome assembler that does not rely on a sequenced genome while also addressing alternative isoform reconstruction. Trinity was developed to efficiently *de novo* reconstruct transcriptome, consisting of three modules: Inchworm, Chrysalis and Butterfly. Inchworm first assembles reads into single transcripts for a dominant isoform and then reports the unique portions of alternatively spliced transcripts. Next, Chrysalis clusters related contigs (e.g. alternative spliced isoforms), and constructs a De Bruijn graph for each cluster. Finally, Butterfly analyses the paths taken by the reads and reports all plausible transcript sequences (Grabherr et al., 2011).

To measure the relative abundance of each transcripts, the raw reads were mapped on the assembled transcriptome using the Bowtie software (v1.0) (Langmead et al., 2009) (parameters: -p 24 --chunkmbs 10240 --maxins 500 --trim5 2 --trim3 2 --seedlen 15 --tryhard –a). Sam output file from Bowtie were converted in bam, sorted, indexed and counted using the sort, index and idxstats programs from the samtools collection, respectively (Li et al., 2009). A final table containing the number of reads mapping on each transcript from each sample was built using a custom R script on the output of the samtools idxstats program. All the transcripts showing less than 0.5 reads mapping per million mapped reads (CPM) in more than 2 samples were discarded from the transcriptome as being expressed at too low levels, and potentially deriving by transcriptional noise or assembly artefacts.

The CPM values of transcripts filtered were used to delineate the distribution of the biological samples in the bi-dimensional principal component analysis (PCA) using the plotPCA function in DESeq package (v.1.18). A custom Perl script was used to calculate the basic statistics (GC content, N50 value, median, average, minimum and maximum length) of the assembled transcriptome.

The CEGMA (Core Eukaryotic Genes Mapping Approach) (v2.5) was used to measure the completeness of the transcriptome data using a set of 248 Core Eukaryotic Genes (CEGs) (Parra et al., 2007). The CEGMA software tool uses sensitive Hidden Markov Models (HMMs) to identify which of these genes are present in a given assembly.

Annotation of the assembled transcripts was performed using the Annocript pipeline (v0.2) (Musacchia et al., 2015). It executes BLASTX (Release 2.2.27+) algorithm (parameters: word_size = 4 evalue = 10^{-5} num_descriptions = 5 num_alignments = 5 threshold = 18) against UniRef90 and Swiss-Prot databases

(Release 2013_09). Protein domains are identified by running RPSBLAST (Release 2.2.27+) algorithm (parameters: evalue = 10^{-5} num_descriptions = 20 num_alignments = 20) against the Conserved Domains Database (CDD; v3.10) quering multiple sequence alignment models for domains from Pfam, SMART, COG, PRK, TIGRFAM (Marchler-Bauer et al., 2013). Non-coding RNAs (rRNA, tRNA, snoRNA, miRNA) are identified by performing a BLASTN (Release 2.2.27+) search (parameters: evalue = 10^{-5} num_descriptions = 1 num_alignments = 1) against an integrated database of Rfam (v11.0) and ribosomal RNAs extracted from GenBank. Each transcript is associated to Gene Ontology (GO) functional classification (Ashburner et al., 2000), Enzyme Commission (EC) numbers (Bairoch, 2000) and Pathways through cross-mapping of the best match from Uniref90 or Swiss-Prot using mapping tables from Uniprot. Virtual Ribosome tool (Dna2pep v1.1) was used to predict the longest open reading frame (ORF) by searching across all reading frames without the constraint for a specific start codon (parameters: o none r all). The final step of the annotation is related to the calculation of noncoding potential for all the input sequence data with Portrait software (v1.1) (Arrial et al., 2009; Wernersson, 2006). Finally, all sequences greater than 200 nucleotides without any homology based on the Blast results, containing an ORF smaller than 100 amino acids and a Portrait noncoding potential score greater than 0.95 were predicted as putative lncRNAs.

The comparison of the percentage of lncRNAs was done against *Homo sapiens* (h38), *Mus musculus* (m38), *Danio rerio* (BDGP5) and *Saccharomyces cerevisiae* (WBcel235) obtained by Ensembl with BioMart.

4.3 Gene Ontology enrichment analysis of the CNS and PNS expressed transcripts

To identify the tissue-expressed transcripts, the count per million (CPM) values has been used. For each tissue, the transcripts with values greater than 1.5 CPM in all the three biological replicates were considered expressed. Transcripts resulting expressed in the SEM, SUB and/or OL tissues were classified as central nervous system: CNS-expressed transcripts. The ARM-expressed transcripts were used as a proxy for the peripheral nervous system: PNS-expressed transcripts. The Gene Ontology (GO) (Ashburner et al., 2000) enrichment analysis was performed on the GO mapping done by Annocript pipeline using a custom R script exploiting the Fisher exact text and P value False Discovery Rate (FDR) correction to select enriched GO classes (minimum number of transcripts associated to a GO class = 100; FDR ≤ 0.01).

4.4 Classification of repetitive elements

Repetitive elements of each transcriptome were annotated using RepeatMasker (v4.0.5) (parameters: -species bilateria, -s, -gff) against Repbase database (v19.06 - Release 20140131) (Jurka et al., 2005; Saha et al., 2008). The RepeatMasker is a program that screens DNA sequences for interspersed repeats and low complexity sequences. The output of the program is a detailed annotation of the query sequence in which all the annotated repeats have been masked. Custom Perl script was used to count, in the RepeatMasker output, the repeats that were present, at least once, in each transcript for each transcriptome. The script is also able to gather together repeats based on their groups and classes (e.g. Retroelements and SINEs). A matrix was built including the percentage of transcripts containing repeats related to

retroelements, DNA transposons, satellites, simple-repeats, low complexity, other and unknown for each transcriptome. The principal component analysis (PCA) was performed on these data in R using the plot function from the graphics package (v3.1.2) followed by the biplot analysis using the bpca function from the bpca package (v1.2-2) (parameters: scale=F, var.rb=T, var.rd=T). The PCA has been used to understand how different organisms correlate among them based on their repeats content.

4.5 Identification of conserved transposable elements

A custom Perl script was written to identify full-length repeat sequences: it parses the RepeatMasker output calculating the position of the match into the repeat. It separates the repeat in three ranges of the same lengths, defining the 5'end (5P), medium (M) and 3'end (3P) regions along the repeat. Then, it classifies the match according to the overlap into the calculated ranges coverage. A custom R script filters the output from the previous Perl script to obtain transcripts containing Long INterspersed Elements (LINEs) with at least 80% of coverage on the repeat consensus sequences. The LINE element with the highest coverage was used to perform the further analysis. Virtual Ribosome tool (Dna2pep v1.1) was used to predict its longest ORF by searching across all reading frames with methionine as start codon (parameters: o strict r all). The InterPro tool (http://www.ebi.ac.uk/interpro/) was used to predict and classify its domains. The amino acids essential for the retrotransposition of the LINE were identified comparing its peptide sequence with those of the 15 L1 and L1-like elements as reported in the Clements paper (Clements and Singer, 1998) (**Table 4.2**).

The evolutionary tree for the full-length LINEs was generated using the 98 protein sequences collected from the Ohshima paper (Ohshima and Okada, 2005) (**Appendix**). The InterPro tool was used to identify the endonuclease and reverse transcriptase domains in the LINE sequences. Multiple sequence alignment was performed by MAFFT program (v7.0 - http://mafft.cbrc.jp/alignment/server/) using UPGMA algorithm with default parameters. Then, the phylogenetic trees were built with FigTree program (v1.4.2 - http://tree.bio.ed.ac.uk/software/figtree/).

Table 4.2. **List of L1 and L1-like elements used**. In the header, **LINE**: Name of LINE elements; **Species**: Scientific name of the organisms; **Accession Number**: ID of LINE available through NCBI nucleotide database.

LINE	Species	Accession Number
L1Hs	*Homo sapiens*	M80343
L1Md	*Mus musculus*	M13002
L1Rn	*Rattus Norvegicus*	X53581
L1Oc	*Oryctolagus cuniculus*	X15965
L1NC	*Nycticebus coucang*	P08548
Tx1	*Xenopus laevis*	M26915
Cin4	*Zea mays*	Y00086
Tal1-1	*Arabidopsis thaliana*	L47193
Zepp	*Bombyx mori*	D85594
RT1	*Anopheles gambiae*	M93690
I	*Drosophila melanogaster*	M14954
Sart-1	*Bombyx mori*	D85594
R1Bm	*Bombyx mori*	M19755
Tad1-1	*Neurospora crassa*	L25662
DRE	*Dictyostelium discoideum*	S20106

4.6 Identification of tissue-specific and candidate transcripts

To select the tissue-specific transcripts, the count per million (CPM) values has been used. The transcripts with values greater than 0.5 CPM in all the three biological replicates in a tissue and lesser in others have been considered tissue-specific and

were represented in the heatmap using MeV software (multi-experiment viewer – v4.8.1).

The coding and non-coding sequences from the tissue-specific transcripts have been identified using the Annocript classification. The candidate coding-transcripts for the further validations were chosen based on the expression levels greater than 1.5 CPM values in all the biological replicates. While the candidate lncRNAs were chosen based on their overlap with a SINE element using the RepeatMasker annotation. Both coding and lncRNA candidates were validated by using polymerase chain reaction (PCR) as reported in the paragraph 2.7.

A custom perl script was used which could identify the keywords (**Table 4.3**) associated to nociception in the Annocript output table. The keywords were searched in the columns of the annotation table related to the description of Swiss-Prot, UniRef, Enzyme, Biological process, Molecular function and Cellular component.

Table 4.3. **List of keywords utilized to identify transcripts involved in the nociception.**

Keywords	
5-hydroxytryptamine	Histamine
acid sensing ion channel family	neurokinin peptide family
anandamide	opioid neuropeptides
arachidonyl glycerol	prostaglandin
arachidonyl serine	substance P
endocannabinoids	synaptosomal associated protein
enkephalin	transient receptor potential channel family
extracellular ATP	tyrosine hydroxylase
family of ion channels	voltage gated sodium channel family
FMRFamide	

4.7 Polymerase chain reaction (PCR)

The cDNA synthesis was generated from 200 ng of total RNA by using Superscript VILO cDNA Synthesis Kit (Life Technologies) in a 20 µl reaction volume. After cDNA synthesis, PCR was assembled using 20 ng of cDNA, 0.25 µl of Taq DNA Polymerase (5U/µl) (Roche), 1 µl of each specific forward and reverse primers (25pmol/µl), 2.5 µl of PCR reaction buffer (10x) and 2.5 µl of dNTP mix (10x), and water up to a final volume of 25 µl. The *Ubiquitin* gene was used as internal control of PCR analysis. The sequences of the primers used are listed in **Table 4.4**. The reactions for the coding transcripts were 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 7 min for 72°C. The reactions for the non-coding transcripts changed for the annealing temperature that was 58°C.

Table 4.4 Sequences of primers utilized for PCR experiments.

Gene Symbol	Primer sequences
UBI	UBIF: TGTCAAGGCAAAGATTCAAGA UBIR: GGCCATAAACACACCAGCTC
ARX	ARXF: TCCCTGCCTTCTCAACACAT ARXR: TCCGAACTTCCACGCTTACT
HOXB5a	HOXB5AF: GTGGCGAGGAATTTAGGAAG HOXB5AR: GCAACAGTCATAGTCCGAACAG
PHOX2B	PHOX2BF: AATGGGGTGAGATCCTTTCC PHOX2BR: TTCATTGCAATCTCCTCTCG
MEOX2	MEOX2F: TCCAGAACCGTCGGATGAAA MEOX2R: TACGTAAAGGGCACACACCT
LncRNAs	**Primer sequences**
SEMl	SEMlF: CACTTGTGCAAGGTACCACG SEMlR: AGGTCTCCTTAAATTTATTTCTGTGCA
SUBl	SUBlF: ACAGAGCATCTTGAGTCTCACT SUBlR: CACTCCTGCGCCTTTCATTT
OLl	OLlF: GGATTGACCCTGCAACTTGG OLlR: CAGTGATGACGGACTTGCAA
ARMl	ARMlF: GTACCCCACAAAATTAAATC ARMlR: CACTCACAAGGCTTTAGTTGGC

Chapter 5 – Results and Discussion

5.1 General statistics from the *de-novo* assembly

The *Octopus vulgaris* nervous system shows the highest level of complexity among cephalopods (Borrelli and Fiorito, 2008). This is recognizable in: the relative brain size, the number of neuronal cells, and the neuroanatomical organization, comparable to that of vertebrates. *Octopus* is considered a good candidate for the analysis of the molecular and neuronal mechanisms underlying complex behaviour. Unfortunately, there is a limited amount of literature on molecular studies so far and a deeper understanding of such complexity is therefore needed (Kocot et al., 2011; Smith et al., 2011; Zhang et al., 2012b).

To this aim, next generation sequencing (NGS) approach has been used on the RNA samples of the main areas of both the central (CNS) and the peripheral (PNS) nervous systems of the *O. vulgaris*. The supra- (SEM), sub- (SUB) esophageal masses and optic lobes (CNS), and the arm (ARM) tissues (PNS) were collected from three adult animals. The RNA-sequencing has generated approximately 850 million paired-end reads, accounting for 85 GB of sequence data. After cleaning of raw sequences, I performed the assembly by combining all the reads across all samples using Trinity software. I obtained 98,174 transcripts longer than 200 bp. Then, I filtered out the transcripts expressed at very low expression levels that may be associated with assembly error. A set of 64,477 unique expressed transcripts was obtained and further clustered in 39,220 putative genes, based on Trinity classification (**Table 5.1**). The sequences showed a N50 value of 2,087 bp, an average length of 1,308 bp with a maximum sequence size of about 20 kb and an average GC-content of 37.9%. The

quality of the transcriptome was compared to an already available transcriptome of the CNS of *O. vulgaris* (Zhang et al., 2012b). The transcriptome I assembled has two times number of sequences in respect to transcriptome published (31,909 sequences with length ≥ 200 bp) with a median length of 795 bp (ONS) and 362 bp (CNS), respectively (**Figure 5.1**).

In order to understand the completeness of the transcriptome generated, I used CEGMA (Core Eukaryotic Genes Mapping Approach) pipeline (Parra et al., 2007). It relieves a set of 248 core proteins (CEGs) that are highly conserved in a wide range of taxa. The CEGMA analysis estimated completeness higher than 97% by taking into account the percentage of CEGs that were complete (more than 70% of protein length aligned) in the dataset. Moreover, the completeness became even higher (98.4%) when considering the percentage of the partial core proteins (fragmented or truncated alignment). The transcriptome published instead shows a less completeness with 68.2% of complete and partial CEGs. These results demonstrate that the transcriptome assembled has a considerable number of unique transcripts in respect to the public data and a great completeness of the eukaryotic conserved proteins.

Table 5.1 Basic statistics about the assembled transcripts.

Assembled and filtered 'transcripts'	64477
Number of 'genes'	39220
Total assembled bases	84399088
GC content (%)	37.9
Contig N50	2087
Median transcript length (bp)	795
Average transcript length (bp)	1308
Minimum length (bp)	201
Maximum length (bp)	20031
248 CEGs Complete (%)	97.2
248 CEGs Complete + Partial (%)	98.4

CEGs: Core Eukaryotic Genes

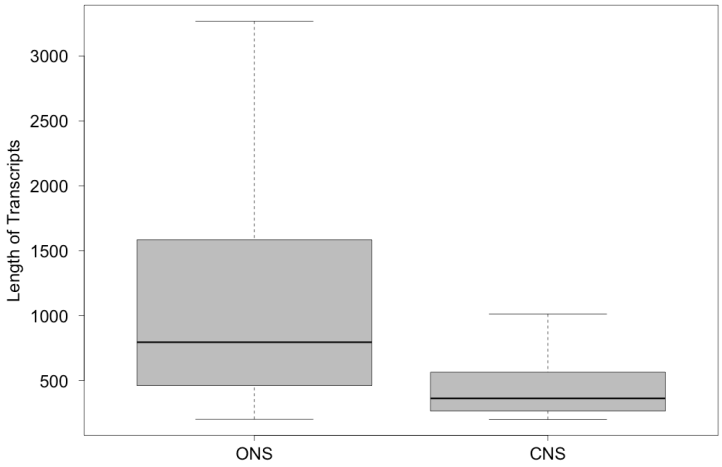

Figure 5.1 Comparison of the *Octopus* nervous transcriptome with public data. The box-plot shows the length distribution of transcripts in the *Octopus* nervous transcriptome (ONS) and the public transcriptome on the Central nervous system (CNS) reported by Zhang et al., 2012.

To inspect the quality of the data and eventually evaluate the degree of similarity between the central and peripheral nervous system, I performed a principal component analysis (PCA) of the samples using the level of expressions of all the transcripts. This analysis is a mathematical algorithm that can be used to examine the correlations among samples based on their patterns of expression levels (Ringnér, 2008). The PCA plot showed a high reproducibility of the transcript expression levels in biological replicates indicating a general good quality of the data (**Figure 5.2**). In addition, SEM, SUB and OL, cluster each other on the principal component 1 (PC1) whereas they clearly differ when the principal component 2 (PC2) is considered. This analysis also revealed that the arm tissue, i.e. PNS, is well separated from the CNS tissues (**Figure 5.2**).

These results show that expression profiles of transcripts can discriminate the CNS from the PNS.

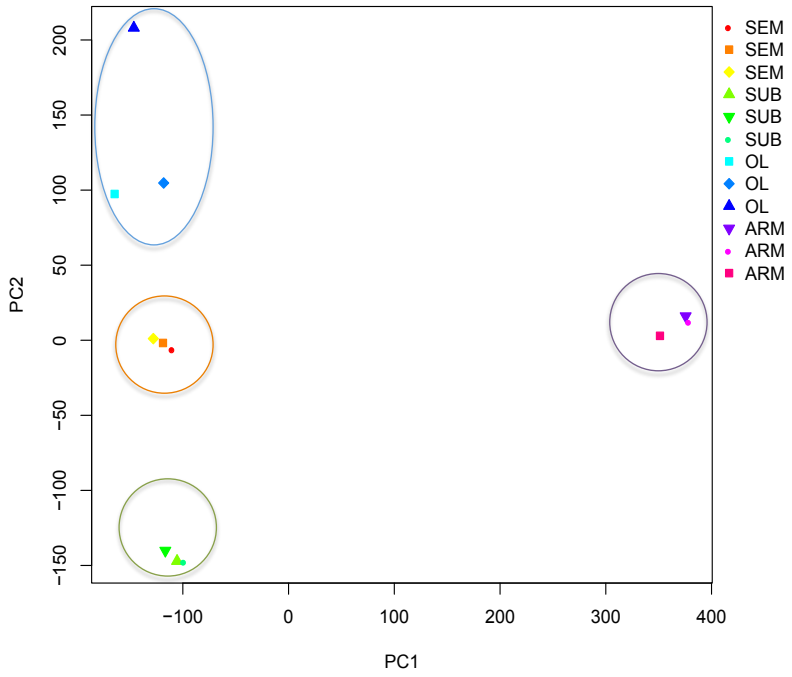

Figure 5.2 Clustering of samples based on their expression levels. Each dot represents one biological sample. Samples from the same tissue type are grouped together in a circle and their positions is defined by principal component 1 (PC1) and principal component 2 (PC2). The orange dots correspond to the supra- (SEM) esophageal mass. The green dots correspond to the sub- (SUB) esophageal mass. The cyan dots correspond to the optic lobe (OL) tissue. The pink dots correspond to the arm (ARM) tissue.

The eukaryotic genomes are dynamic entities in which transposable elements have the ability to move from one chromosomal location to another (Finnegan, 2012). Virtually all organisms harbour transposable elements that have amplified in copy number over evolutionary time via DNA or RNA intermediates (Levin and Moran, 2011). Moreover, the genomes of cephalopods are known to be larger and more repeat-rich than many previously sequenced metazoan genomes (Yoshida et al., 2011).

Therefore, I evaluated the contribution of repeats into the *Octopus* nervous system transcriptome using RepeatMasker software (Saha et al., 2008). This analysis revealed that approximately 7.8% of the transcriptome is represented by repeats

(**Table 5.2**). Among them, interspersed repeats (retroelements, DNA transposons) are the majority and correspond to the 4.5% of the total bases. Other classes of repeats represented into the transcriptome are: simple repeats (2.8%), low complexity (0.6%), and satellites (0.2%). In order to understand the repeats distribution within the *Octopus* transcriptome, I evaluated the percentage of transcripts containing at least one repetitive element. About 46,944 (72.8%) of transcripts contain repeats; in particular, 35.5% of the generated transcripts contain interspersed repeats, with retroelements being the most frequent ones (26.2%). Thus, transposable elements are the most abundant class of repetitive elements into the *Octopus* nervous system transcriptome.

Table 5.2 Repeats composition for the assembled transcriptome. The table shows the number of nucleotides and transcripts with their related percentage containing the different classes of repetitive elements. The percentage values are referred to the total number of nucleotides and transcripts analyzed, respectively.

	Nucleotides	Percentage	Transcripts	Percentage
Bases Masked	6584938	7.8	46944	72.8
Retroelements	2102784	2.5	16926	26.2
DNA transposons	1501995	1.8	12601	19.5
Unclassified	169576	0.2	250	0.4
Total interspersed repeats	3774355	4.5	22915	35.5
Satellites	150431	0.2	1326	2.1
Simple repeats	2368596	2.8	34833	54.0
Low complexity	470441	0.6	7704	11.9

5.2 Functional annotation of the transcriptome

In order to understand the biological functions of the assembled transcriptome for the *Octopus* nervous system, I annotated the sequences using the Annocript pipeline (Musacchia et al., 2015). This pipeline combines the annotation of protein

coding transcripts with the prediction of putative long noncoding RNAs (lncRNAs) in the transcriptome. The functional annotation of the *Octopus* transcriptome has been obtained for 21,030 (32.6%) protein-coding transcripts. This result is a notable improvement considering that the only study on the CNS transcriptome of the *Octopus* annotated 10,412 (17.4%) transcripts (Zhang et al., 2012b).

To understand the most abundant functional classes in the transcriptome, I selected the top 10 Gene Ontology (GO) terms with the highest number of transcripts for biological processes, molecular functions and cellular components (**Figure 5.3**). Based on this filtering, a high number of transcripts result involved in biological processes such as "*RNA-dependent DNA replication*", "*homophilic cell adhesion*" and "*regulation of transcription DNA-dependent*" (**Figure 5.3 A**). The "*RNA-dependent DNA replication*" shows the most abundant class of transcripts (2.4%) is implicated into the DNA replication process that uses RNA as a template to synthesize the new strands, that is the mechanism used by retrotransposons to jump from a chromosome to another in the host genome (Finnegan, 2012). The *"homophilic cell adhesion"* includes 1.3% of transcripts encoding adhesion molecules for the plasma membrane, whereas, the *"regulation of transcription DNA-dependent"* contains about 1.2% of transcripts that modulate the rate of cellular DNA-templated transcription.

According to the molecular function classification, the most common classes are related to single-molecule binding such as "*zinc ion binding*", "*nucleic acid binding*" and "*ATP bindin*g" (**Figure 5.3 B**). The most abundant molecular function (8%) encodes molecules that interact selectively with zinc (Zn) ions. Another 10% of transcripts are equally distributed in the interaction with any nucleic acid and ATP (adenosine 5'-triphosphate). Cellular component annotations suggest that the expressed protein products mainly localize into the membrane bilayer and within the

nucleus (**Figure 5.3 C**). The most representative protein domains are Zn-finger, protein kinase and cadherin domains (**Figure 5.3 D**). In particular, the Zn-finger domains are generally found among the most frequent both in vertebrates (e.g. human, mouse and zebrafish) as well as invertebrates (e.g Drosophila and yeast) (**Table 5.3**).

The eukaryotic transcriptome is also composed by a great counterpart of long noncoding RNAs (lncRNAs) which play specific and diverse roles especially in the nervous system. In particular, in the central nervous system, more than half of known lncRNAs are expressed in different neuron types (Derrien et al., 2012; Knauss and Sun, 2013; Mercer et al., 2008). Also, aberrant expression of lncRNAs has been associated with neurological disorders (Faghihi et al., 2008; St George-Hyslop and Haass, 2008; Ziats and Rennert, 2013). These findings suggest that lncRNAs may be crucial effectors in cognitive and behavioural repertoires in mammals (Fatica and Bozzoni, 2014).

Hence, I decided to search for lncRNAs in the whole transcriptome of the *Octopus* nervous system, using the Annocript pipeline. The prediction of putative lncRNAs is based on an heuristic process that takes in account their lengths, the lack of similarity with protein, domain and other short ncRNA, the length of the open reading frame and the noncoding potential score. The constraints used to identify potential lncRNAs are very stringent. Indeed, in published studies different combinations of analyses are used to identify lncRNAs (Arrial et al., 2009; Cabili et al., 2011; Ulitsky et al., 2011; Pauli et al., 2012). In summary the features searched in these studies are:

1. Lack of similarity with proteins.
2. Lack of similarity with domain profiles.

3. Lack of similarity with other non-coding RNAs (ribosomal, snoRNA, miRNA).
4. Transcript and ORF lengths.
5. Non-coding potential score (mainly SVM classification based on compositional patterns).

Annocript executes these analyses in a single pipeline and performs a final step to identify potential lncRNAs, keeping into account all the collected results. By default a transcript is considered long noncoding if it satisfies all the following conditions:

1. Length >= 200 nucleotides.
2. Lack of similarity with any protein, domain, rRNA, other short ncRNA from Rfam.
3. ORF < 100 amino acids.
4. Non-coding potential score >= 0.95.

The Annocript pipeline was able to predict around 12.1% (7,806) of transcripts as putative lncRNAs. It is an interesting result because the amount of lncRNAs in the *Octopus* is more comparable to the vertebrates (12.5% human, 8.3% mouse and 6.7% zebrafish) rather than the invertebrates' transcriptomes (2.4% *Drosophila* and 0.2% yeast).

Altogether the functional annotation of the *Octopus* transcriptome shows widespread transcription of long non-coding transcripts and several transcripts of the protein-coding counterpart appear to be implicated in reverse transcription and signalling process.

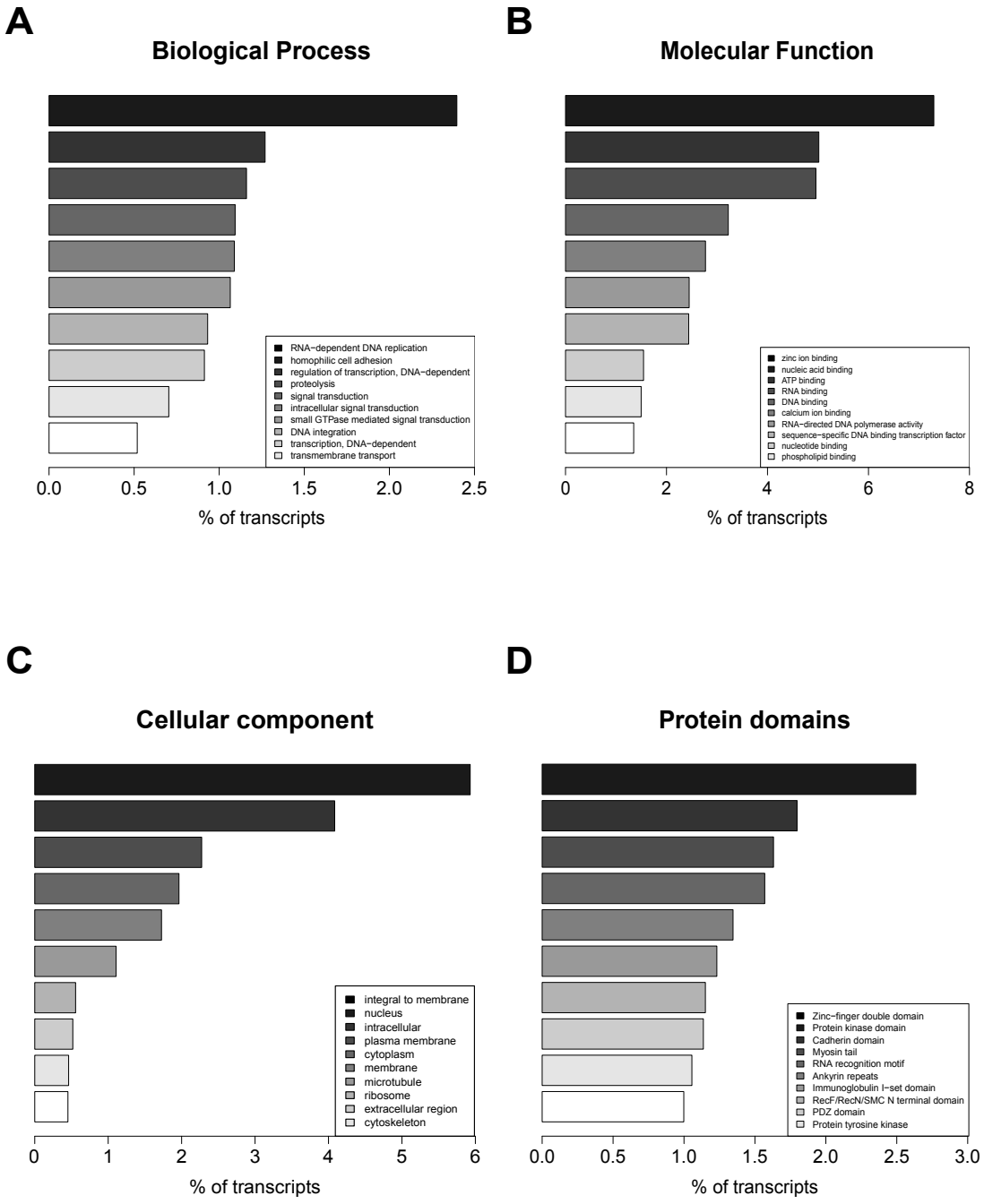

Figure 5.3. Gene Ontology analysis of the *Octopus* nervous system transcriptome. Distribution of the top ten functional classes in the biological processes (**A**), molecular functions (**B**), cellular components (**C**) and protein domains (**D**). On the y axis are the GO-terms reported in the legend, on the x axis are the percentage of transcripts associated to each GO-term.

Table 5.3 Amount of Zinc finger domains in vertebrates and invertebrates. The table shows the number of genes containing the Zinc finger domain (IPR007087) in the species selected from Ensembl. The rank shows the position of the Zinc finger domain in the top 500 InterPro hits.

Species	Number of genes	Rank	Ensembl release
Homo sapiens	811	3	h37
Mus musculus	708	5	m37
Danio rerio	710	8	zv8
Drosophila melanogaster	326	2	BDGP5.25
Saccharomyces cerevisiae	49	25	EF2

5.3 Gene Ontology enrichment analysis of the CNS and PNS transcripts

In order to inspect, at the molecular level, the functions of the main areas of the *Octopus* nervous system, I performed a Gene Ontology (GO) enrichment analysis.

I identified the transcripts with expression levels greater than 1.5 count per million (CPM) in the biological replicates from the supra-, sub-esophageal masses and the optic lobe samples for the CNS and from the arm (**Table 5.4**). Then, I selected the transcripts expressed in at least one tissue of the CNS. A total of 57,445 transcripts have been selected in this case.

Table 5.4 Tissue-expressed transcripts in the *Octopus* nervous system. The table shows the number of transcripts highly expressed and annotated in the each tissue of the CNS (SEM, SUB, OL) and PNS (ARM). The percentages are refereed to the total number of transcripts and transcripts expressed, respectively.

Tissues	N. of transcripts expressed	N. of transcripts annotated
SEM	52929 (82.1%)	19791 (37.4%)
SUB	49420 (76.6%)	19223 (38.9%)
OL	49568 (76.9%)	19154 (38.6%)
CNS (SEM+SUB+OL)	57445 (89.1%)	19939 (34.7%)
ARM	26119 (40.5%)	15010 (57.5%)

About 89.1% of the total transcripts are expressed at high levels in the CNS tissues while only the 40.5% are highly expressed in the PNS. I performed a GO enrichment analysis among those sets of transcripts to understand which are the difference between CNS and PNS of the *Octopus* nervous transcriptome.

In general, the PNS shows several functional enrichments for many diverse functions and locations in respect to CNS transcripts (**Figure 5.4 A, B and C**). This is due to the fact that the PNS contains a significantly higher fraction of annotated transcripts in respect to the CNS (prop.test p-value < 2.2e-16). Indeed, CNS expressed transcripts represent about 90% of the whole generated transcriptome and as expected maintain the same distribution of functions and localization. Notably, despite the similarity between the CNS and the whole neural transcriptome, the CNS still maintain a significantly higher amount of transcripts associated to "*RNA-dependent DNA replication*" (p-value ≤ 1.00e-3) (**Figure 5.4 A**) and "*RNA–directed DNA polymerase activity*" (p-value ≤ 1.37e-3) (**Figure 5.4 B**). In addition, the most significantly enriched classes associated to PNS expressed transcripts belong to mainly signalling and metabolic functions such as signal transduction (**Figure 5.4 A**)

and ATP binding (**Figure 5.4 B**). These results indicate a gain of transcripts involved in the DNA replication mediated by reverse transcriptase and signalling pathway for the CNS and PNS, respectively.

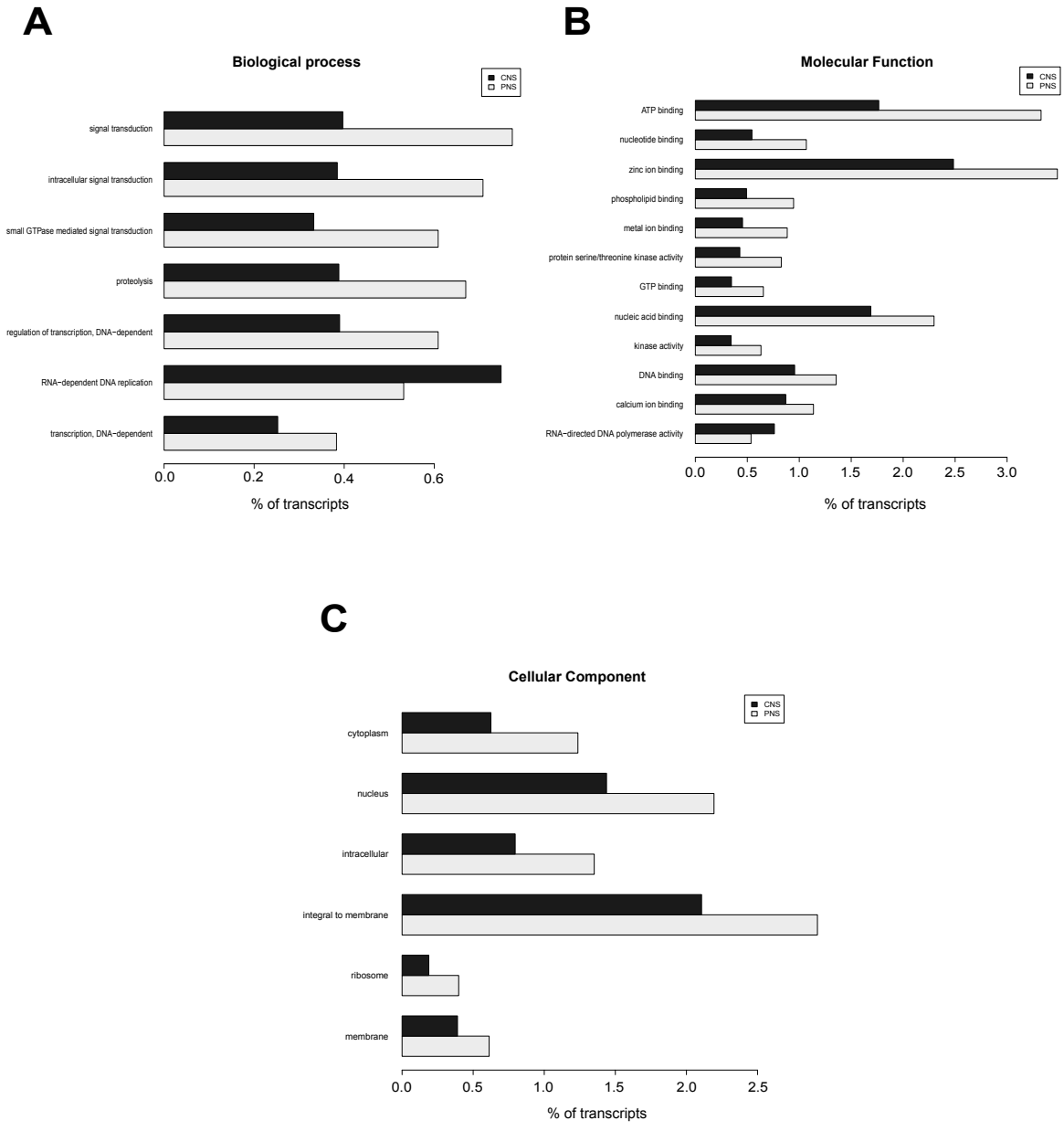

Figure 5.4. Gene Ontology enrichment analysis among the CNS and PNS expressed transcripts. The bar-plots show the percentage of transcripts which are defined by GO terms enriched in the biological processes (**A**), molecular functions (**B**) and cellular components (**C**) for the CNS (black bar) and for the PNS (grey bar)**.** On the y axis are the GO-terms, on the x axis are the percentage of transcripts associated to each GO-term.

5.4 Distribution of TEs and lncRNAs in the *Octopus* neural transcriptome

During the last 10 years, several evidences showed that transposable elements (TEs) can mobilize in the genome of brain neurons somatically (Muotri et al., 2005; Coufal et al., 2009; Baillie et al., 2011). Even though this phenomenon is well studied in mammals, it has been demonstrated that it is not restricted only to vertebrates (Perrat et al., 2013). Therefore, I decided to evaluate the abundance of TEs in the different compartments of the *Octopus* nervous system. I compared the TE-content of transcripts expressed in each tissue of the central and peripheral nervous systems. These analyses revealed that the Short INterspersed Elements (SINEs) are the unique retroelements to be significantly enriched in transcripts expressed in the CNS (Bonferroni-adjusted p-values ≤ 3.32e-4), while the Long INterspersed Elements (LINEs) and the Long Terminal Repeats (LTRs) are enriched in PNS expressed transcripts (Bonferroni-adjusted p-values ≤ 1.04e-10 and 3.07e-15, respectively) (**Figure 5.5 A**). DNA transposons are also enriched in transcripts expressed in the PNS (Bonferroni-adjusted p-values ≤ 1.72e-2), as occur for the Mavericks and Helitrons elements (Bonferroni-adjusted p-values ≤ 4.47e-2 and 1.91e-2, respectively).

Different studies have identified pervasive transcription of long noncoding RNAs (lncRNAs) in the CNS (Derrien et al., 2012; Knauss and Sun, 2013; Mercer et al., 2008), and the increase in number of lncRNAs has been linked to the evolutionary complexity (Qureshi and Mehler, 2012). For these reasons, I compared the percentage of lncRNAs in each tissue of the CNS and PNS. Similarly to SINE elements, lncRNAs show enrichment in the CNS (**Figure 5.5 B**). Indeed, more than 10% of transcripts expressed into the CNS are lncRNAs compared to about 7.3% in PNS

(Bonferroni-adjusted p-value ≤ 1.97e-44) confirming also in the *Octopus* what has been observed in other organisms (Ponjavic et al., 2009).

Recent studies suggest a positive correlation between lncRNAs and repetitive elements in vertebrate transcriptomes (Kapusta et al., 2013; Kelley and Rinn, 2012). Particularly, it has been demonstrated in human, mouse and zebrafish transcriptomes that repetitive elements are more abundant in non-coding rather than in coding transcripts. Thereafter, I decided to investigate whether this kind of correlation could also occur in the *Octopus* transcriptome.

Interestingly, among the different classes of retroelement, SINEs are significantly enriched in lncRNAs compared to coding transcripts (Bonferroni-adjusted p-value ≤ 2.50e-37) (**Figure 5.5 C**). In contrast, LINEs and LTRs seem to be enriched in the portion of coding transcripts (Bonferroni-adjusted p-values ≤ 4.01e-25 and 7.09e-8, respectively). This last result is in contrast with those reported in mammals (Kelley and Rinn, 2012). It could be explained in two different ways. First, transcripts containing coding portion of LINE and LTR have been classified as coding in the annotation analysis due to the presence of coding-protein domains necessary for the mechanisms of retrotransposition. Second, these RNA-seq data were generated with a non-strand specific library protocol and therefore, following the annotation step, putative antisense noncoding transcripts (AS) of sense protein coding genes (S) have been classified as coding when the S/AS overlap interested coding exons.

Finally, the three categories of DNA transposons – cut-and-paste, Mavericks and Helitrons – are instead equally distributed in the coding as well as in the non-coding transcripts (Bonferroni-adjusted p-values ≤ 0.05).

Overall these data suggest that the organization of the *Octopus* neural transcriptome might reflect, at least in part, the mammalian one in terms of noncoding and repeats content in the brain.

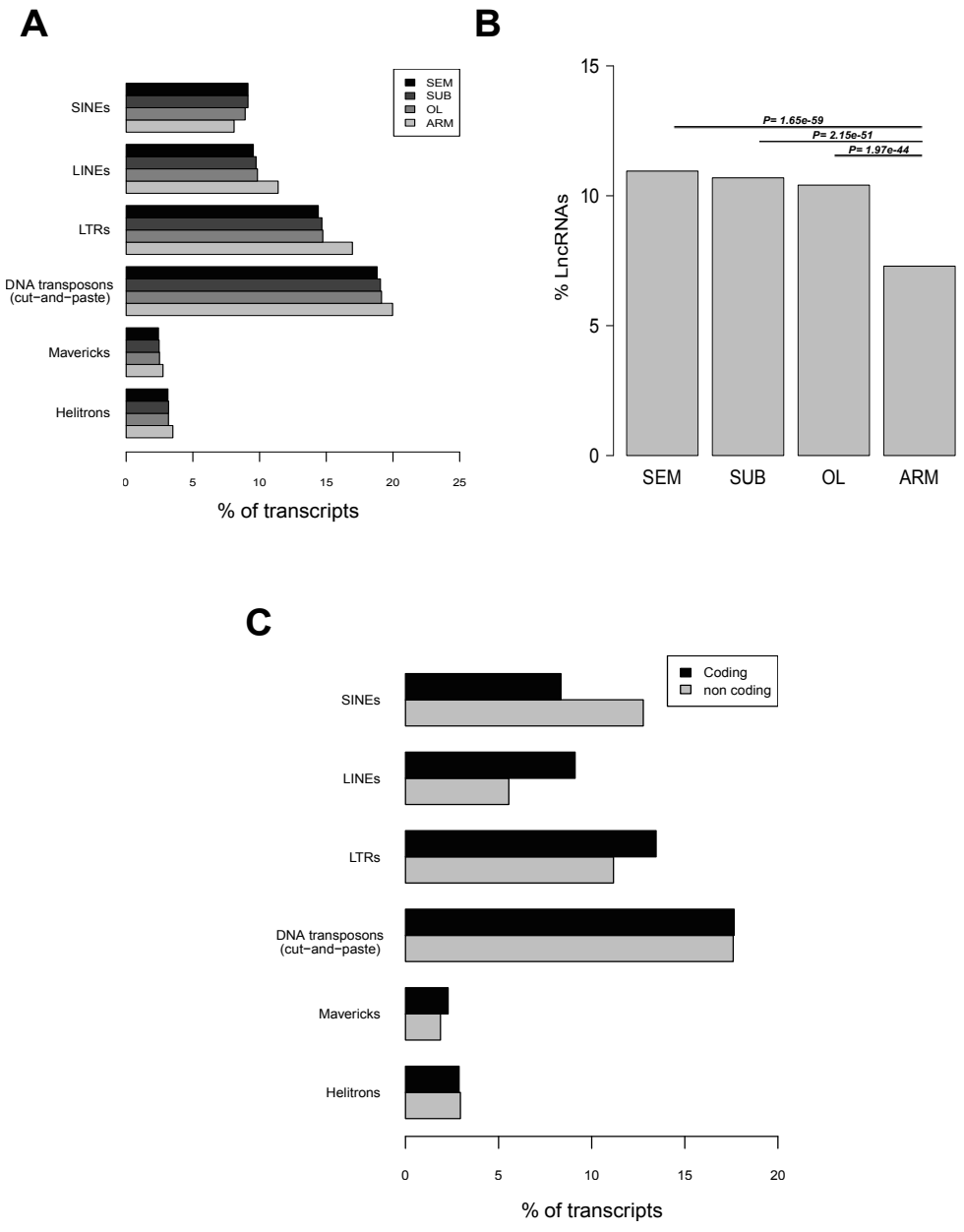

Figure 5.5 Distribution of transposable elements and lncRNAs in the *Octopus* nervous system. (**A**) Proportion of transcripts containing transposable elements (**A**) and lncRNAs (**B**) in different areas of the CNS (SEM, SUB, OL) and PNS (ARM) of *Octopus*; (**C**) Proportion of transposable elements in coding (black bar) and lncRNAs (grey bar). A and C report types of transposable elements on the y-axis and the percentage of transcripts containing those repeats on the x-axis, respectively. The opposite orientation is used in figure B for the lncRNAs.

5.5 Comparison of the repeat and lncRNA contents among the species

In order to understand the composition of the *Octopus* transcriptome in terms of repeat and lncRNA contents from an evolutionary point of view, I compared it with the transcriptomes of thirty-nine organisms belonging to seven different phyla. I used Portrait software to predict the putative lncRNAs and compared their amounts in each organism tested.

My results demonstrate that lncRNAs are enriched in the transcriptome of molluscan species (**Figure 5.6**). The only exception is the *Crassostrea gigas*; this is probably due to the fact that, as suggested in the original study, the published transcriptome has been depleted of non-annotated and therefore noncoding sequences (Zhang et al., 2012a). Interstingly, *O. vulgaris* and *Siphonaria pectinata* have high percentage of lncRNAs (14.6% and 15.8%, respectively), resulting more similar to mammals *Papio anubis* (18.2%) and *Homo sapiens* (17.1%).

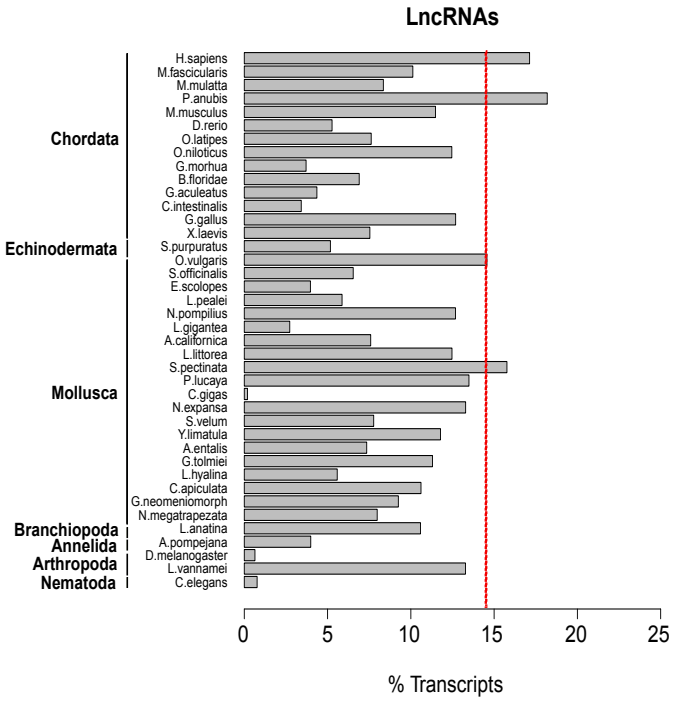

Figure 5.6 Long noncoding RNAs abundance among species. Proportion of lncRNAs in the transcriptomes of the species tested. On the y-axis are listed the species selected; the x-axis shows the percentage of putative lncRNAs. The red dashed line represents the percentage of lncRNAs in *Octopus vulgaris*.

In addition, I also evaluated the proportion of transcripts containing repetitive elements in the *Octopus* and in the thirty-nine other organisms (**Figure 5.7 A**).

Strikingly, the retroelements content in the *O. vulgaris* transcriptome (26.2%) results comparable to the mammals, such as *Macaca fascicularis* (27.1%) and *M. mulatta* (26.4%). Among the molluscs, only cephalopod species (*Loligo pealei*, *Sepia officinalis, Euprymna scolopes*) contain a retroelement proportion higher than the invertebrate average, and more similar to the *Octopus* (16.6%, 10.3% and 8.2% respectively). *O. vulgaris* transcriptome show a high percentage of DNA transposons (18.5%) and this results second only to *Danio rerio* (29.5%) (**Figure 5.7 B**).

Looking at the other repeat groups, the satellites are more abundant in the chordates, *M. mulatta* (6.1%) and *D. rerio* (3.3%), and in the molluscs, *Solemya velum* (2.1%) and *O. vulgaris* (2.1%).

Cephalopods, *E. scolopes*, *L. pealei* and *O. vulgaris* (84.3%, 55.8% and 54% respectively), together with *Drosophila melanogaster* (55.4%), show higher percentage of simple repeats than all the other species (**Figure 5.7 C**). The low-complexity sequences are abundant in *Strongylocentrotus purpuratus* (24.7%), *L. pealei* (14.0%), *D. melanogaster* (13.9%) and *O. vulgaris* (11.9%).

These analyses reveal that *Octopus* neural transcriptome has a high percentage of transcripts containing interspersed repeats, and the high retroelement content suggests similarity to the mammals.

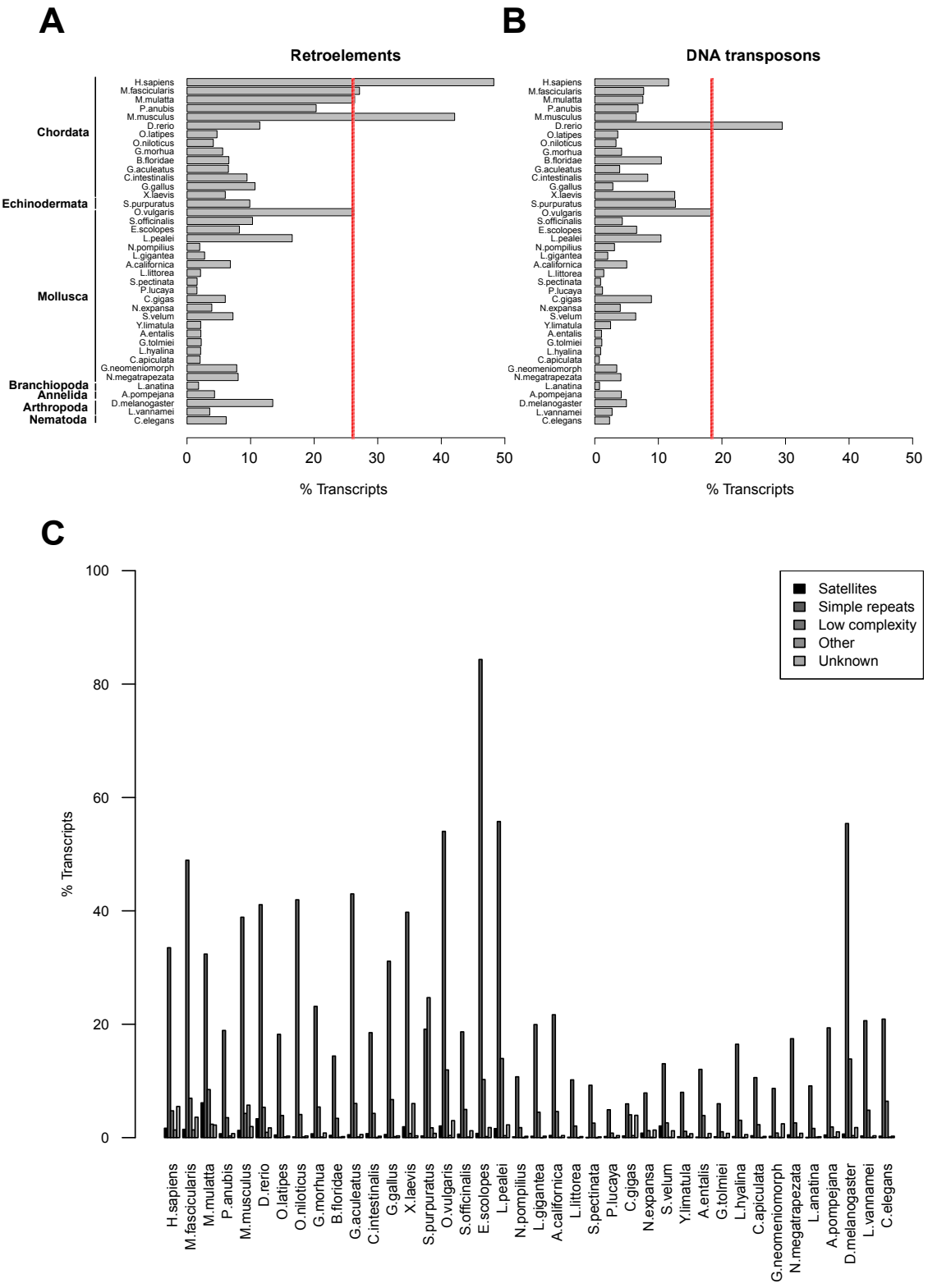

Figure 5.7 Repetitive elements abundance in the organisms. Proportion of transcripts containing retroelements (**A**), DNA transposons (**B**) and other repeat groups (satellites, simple repeats, low complexity, other and unknown) (**C**) among species tested. The red dashed lines represent the percentage of retroelements and DNA transposons in the *Octopus vulgaris* transcriptome. A and B report the species selected (y–axis) and the percentage of transcripts containing retroelements and DNA transposons (x–axis). The opposite orientation is used in figure C for the other repeats groups.

Thereafter, I decided to understand which retroelements (SINEs, LINEs and LTRs) and DNA transposons (cut-and-paste, Mavericks and Helitrons) are more abundant in the *Octopus* transcriptome in respect to the thirty-nine organisms selected. To reach this aim, I identified the percentage of transcripts containing at least one of those elements.

SINE elements are highly abundant in *O. vulgaris* (8.9%), second only to the higher vertebrates (27.7% *H. sapiens*, 27.3% *M. musculus*, 17.1% *M. fascicularis*, 15.4% *M. mulatta*) (**Figure 5.8 A**). The cephalopod *L. pealei* (7.1%) is the sole other mollusc species showing a percentage of SINEs comparable to *Octopus*. *O. vulgaris* transcriptome also contains a higher percentage of transcripts containing LINE (8.7%) and LTR (13.2%) elements than all the other mollusc species (**Figure 5.8 B-C**).

It is already known that retroelements are widespread among mammalian genomes and non-LTR retrotransposons are especially prolific in mammalian genomes (Treangen and Salzberg, 2012). These results confirm that, as expected, SINEs and LINEs constitute a large portion of mammalian transcriptomes. Surprisingly, the *Octopus* is the unique organism to have an abundance of SINEs comparable to the mammals.

The canonical cut-and-paste is the class of the DNA transposons more abundant in the organisms selected. The *Octopus* shows a high content of those DNA transposons (17.6%) and it is second only to the *Danio rerio* (29.3%) (**Figure 5.9 A**). Regarding the other DNA transposon classes, the Mavericks are particularly enriched in the *Strongylocentrotus purpuratus* (13.1%) whereas the Helitrons show higher levels in *Crassostrea gigas* (3.8%) and *Octopus vulgaris* (2.9%) than all the other organisms (**Figure 5.9 B-C**).

This comparative analysis shows an expansion of lncRNAs and transposable elements in the cephalopoda taxon. Among the Cephalopoda *Octopus vulgaris* is the organism in which this expansion results more dramatic. Its transcriptome shows an enrichment for retroelements, especially SINEs, in respect to other invertebrates and this leads to the grouping of *O. vulgaris* with mammals.

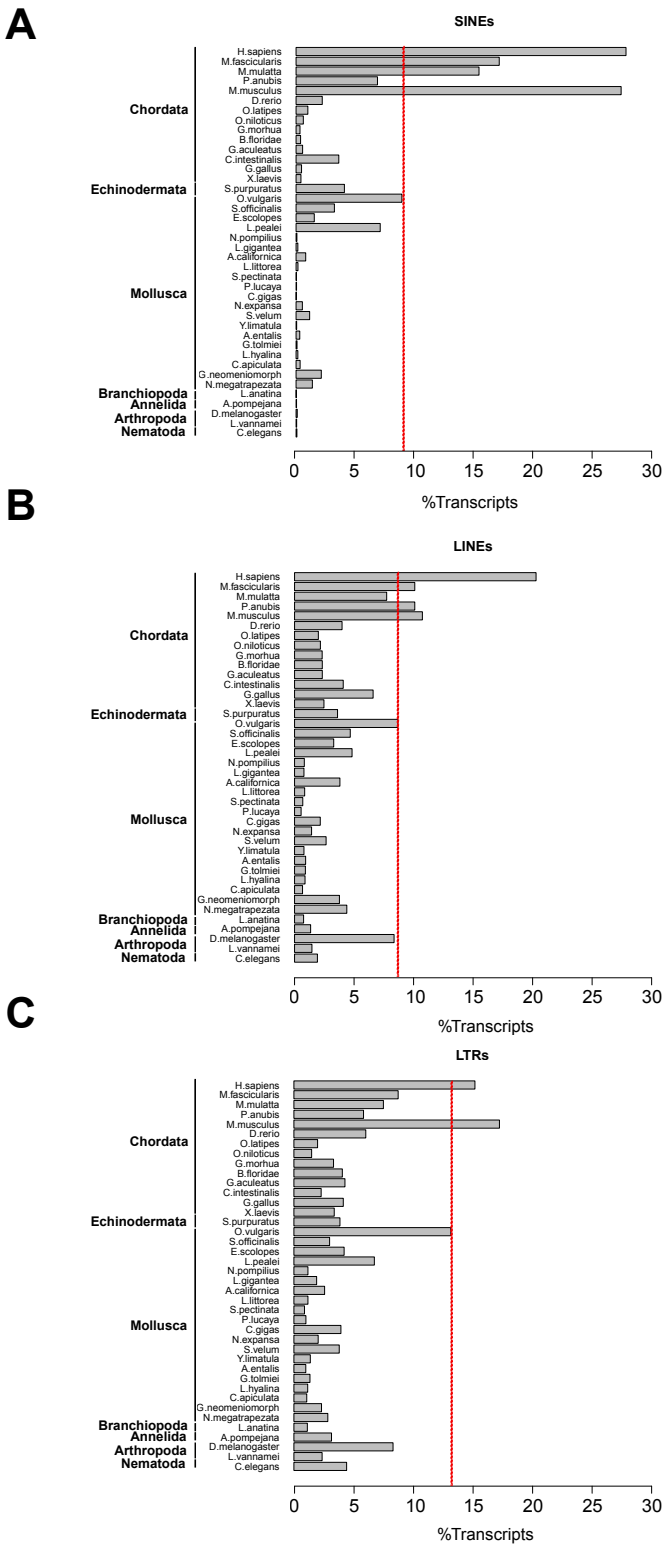

Figure 5.8 Abundance of Retroelement classes in the organisms. Proportion of SINEs (**A**), LINEs (**B**) and LTRs (**C**) among species tested; On the y-axis are listed the species selected; the x-axis show the percentage of transcripts containing SINEs, LINEs and LTRs.The red dashed lines represent the percentage those retroelements in the *Octopus vulgaris* transcriptome.

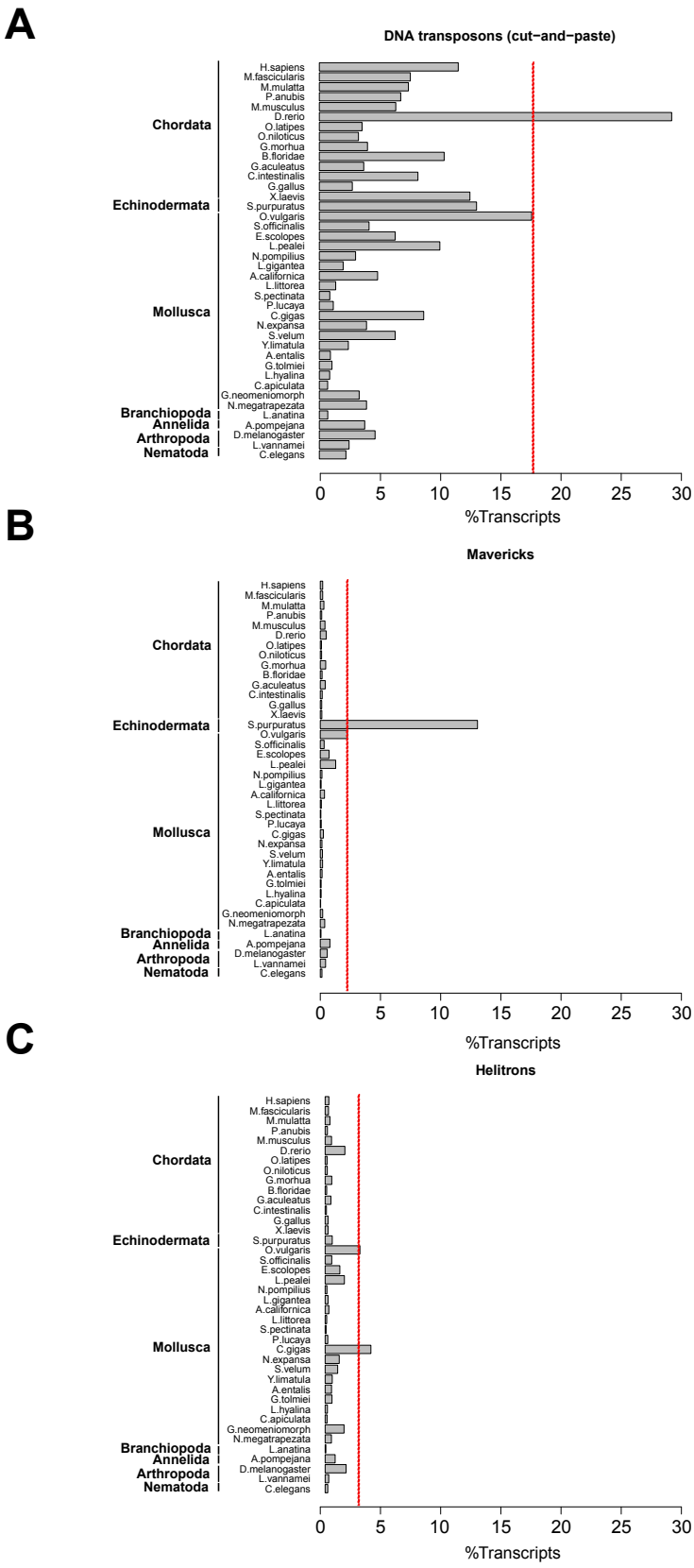

Figure 5.9 Abundance of DNA transposons classes in the organisms. Proportion of DNA transposons (cut-and-paste) (**A**), Mavericks (**B**) and Helitrons (**C**) among species tested; the y-axis lists species selected; the x-axis shows the percentage of transcripts containing DNA transposons (cut-and-paste), Mavericks and Helitrons. The red dashed lines represent the percentage those DNA transposons in the *Octopus vulgaris* transcriptome.

To understand how the selected organisms correlate each other based on their composition in repeats, I generated a bi-plot graph based on a principal component analysis (PCA). The principal components 1 (PC1) and 2 (PC2) reveal more than 91% of variance for the repeat-content in the organisms tested. The bi-plot highlights that *O. vulgaris* together with *L. pealei* and *E. scolopes* are far from all the molluscs (in pink) and appear more closed to the chordates (in blue); in addition, *O. vulgaris* appears to be the closest organism to the mammals. This analysis also permits to understand which repetitive elements drive distribution of organisms selected. Indeed, the similarity of the *Octopus* to the mammals seems to be determined by transposable elements, mainly retrotransposons and, at a lesser extent, DNA transposons (**Figure 5.10**).

These results highlight an unexpected high frequency of retroelements embedded in octopus transcripts, much more similar to that shown in mammals and much higher than in any other invertebrate and vertebrate non-mammalian species.

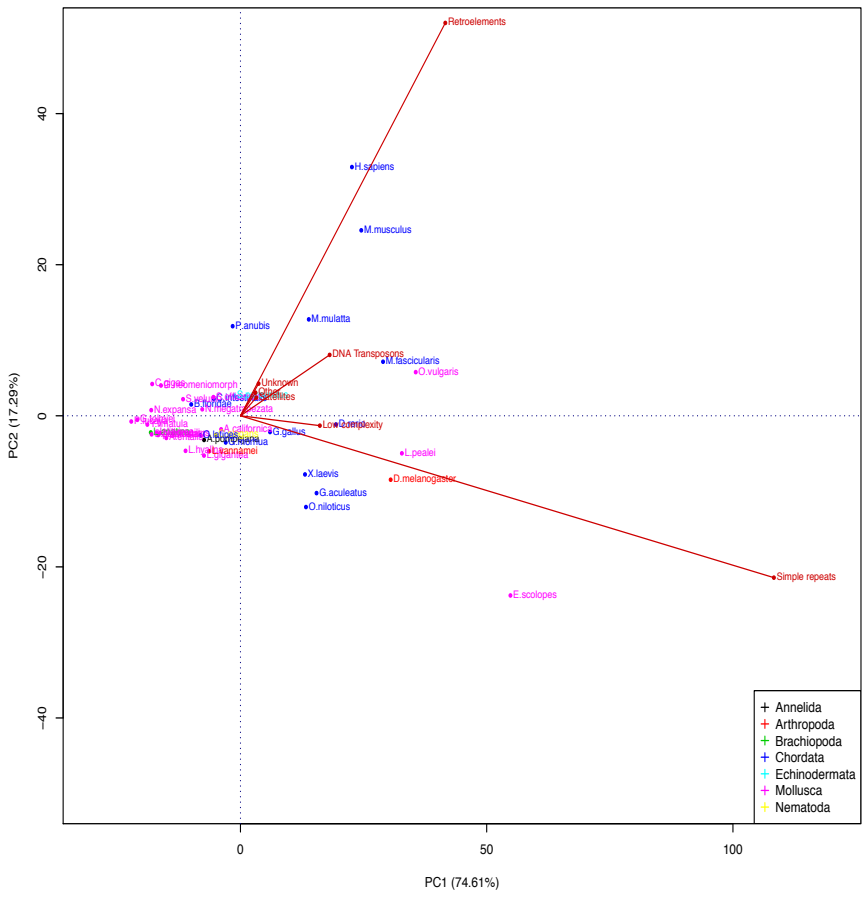

Figure 5.10 Correlations among the organisms based on their repetitive elements content. The bi-plot shows the distribution of the different species according to the proportion of repeats in the transcriptome. The species are coloured based on their phyla. The length of lines indicates the variances of the variables used.

5.6 Structure and conservation of transposable elements

In the previous paragraph, it has been shown that the *Octopus* transcriptome has high percentage of SINEs embedded in the transcripts, which is comparable with the mammals. The mechanism of mobility that is generally used by SINE and LINE elements initiates with the synthesis of their RNA intermediate (Finnegan, 2012). LINEs are autonomous elements whose active form encode a protein with endonuclease (Feng et al., 1996) and reverse transcriptase (Mathias et al., 1991) domains that permit the reverse transcription of the RNA intermediate and its

integration in the other parts of the genome. SINEs are instead non-autonomous retroelements, which do not code for any of the proteins needed for retrotransposition activity but they utilize the protein machinery coded by LINEs to expand into the genomes (Dewannieux and Heidmann, 2005; Dewannieux et al., 2003). Hence, I searched full-length LINE elements in the *Octopus* transcriptome. To identify them, I used the annotation of the repeats obtained by the RepeatMasker output. This analysis allowed to found a single LINE element belonging to the RTE clade, similar to the RTE-2_CPB which was recently identified in the western painted turtle (*Chrysemys picta bellii*) genome (Shaffer et al., 2013). Members of the RTE clade have been studied in nematods and insects (Malik and Eickbush, 1998), teleost fishes (Volff et al., 1999) and mammals (Szemraj et al., 1995). They are the shortest of non-LTR retrotransposons (~3.2 kb in length) with a small ORF1, not always present, and a highly conserved ORF2 encoding apurinic/apyrimidinic endonuclease and reverse transcriptase domains (Župunski et al., 2001).

The assembled transcript is long 4,649 bp and shows an open reading frame (ORF) of 3,327 nucleotides (**Figure 5.11 A**). The translation is 1,109 amino acids (aa) long and contains three domains: 1) a C-terminal apurinic/apyrimidinic endonuclease (EN) domain whose function is to cleave the target DNA for retrotransposition, generating a reverse transcription primer, 2) a reverse transcriptase (RT) domain capable of synthetizing DNA from RNA using the primer generated by the EN and, 3) a C2H2 zing-finger (Znf) domain present downstream the RT (Malik and Eickbush, 1998).

To understand whether this element could still be active, I compared the 11 amino acids essential for the endonuclease, reverse transcriptase and retrotransposition activities (Clements and Singer, 1998) included in the ORF proteins from 15 L1 and

L1-like elements (**Figure 5.11 B**). Those amino acids are highly conserved suggesting that RTE-2_OV is fully competent for the retrotransposition in the *Octopus* neural transcriptome. Then, I evaluated its expression level in the tissues belonging to the central and peripheral nervous systems (**Figure 5.11 C**). The RTE-2_OV is highly transcribed in the CNS, especially in the supra- and sub-esophageal masses. Finally, in order to understand the closeness or the distance of the LINE element to the main LINE clades, I performed a phylogenetic analysis using the amino acid sequences of 98 full-length LINEs (**Appendix**) and only the regions that encode endonuclease and reverse transcriptase domains respectively for 96 and 70 of those LINEs. The phylogenetic tree for the full-length sequences shows that the RTE-2_OV is included in the RTE clade and it clusters in the same branch with Rex3 and RTE-1_AG belonging to the *Xiphophorus maculatus* and *Anopheles gambiae* respectively (**Figure 5.12 A**). The same occurs when I compared the endonuclease and reverse transcriptase domains (**Figure 5.12 B-C**). Moreover, the RTE clade shares the branch with the L1 clade, again suggesting an interesting closeness to the mammals.

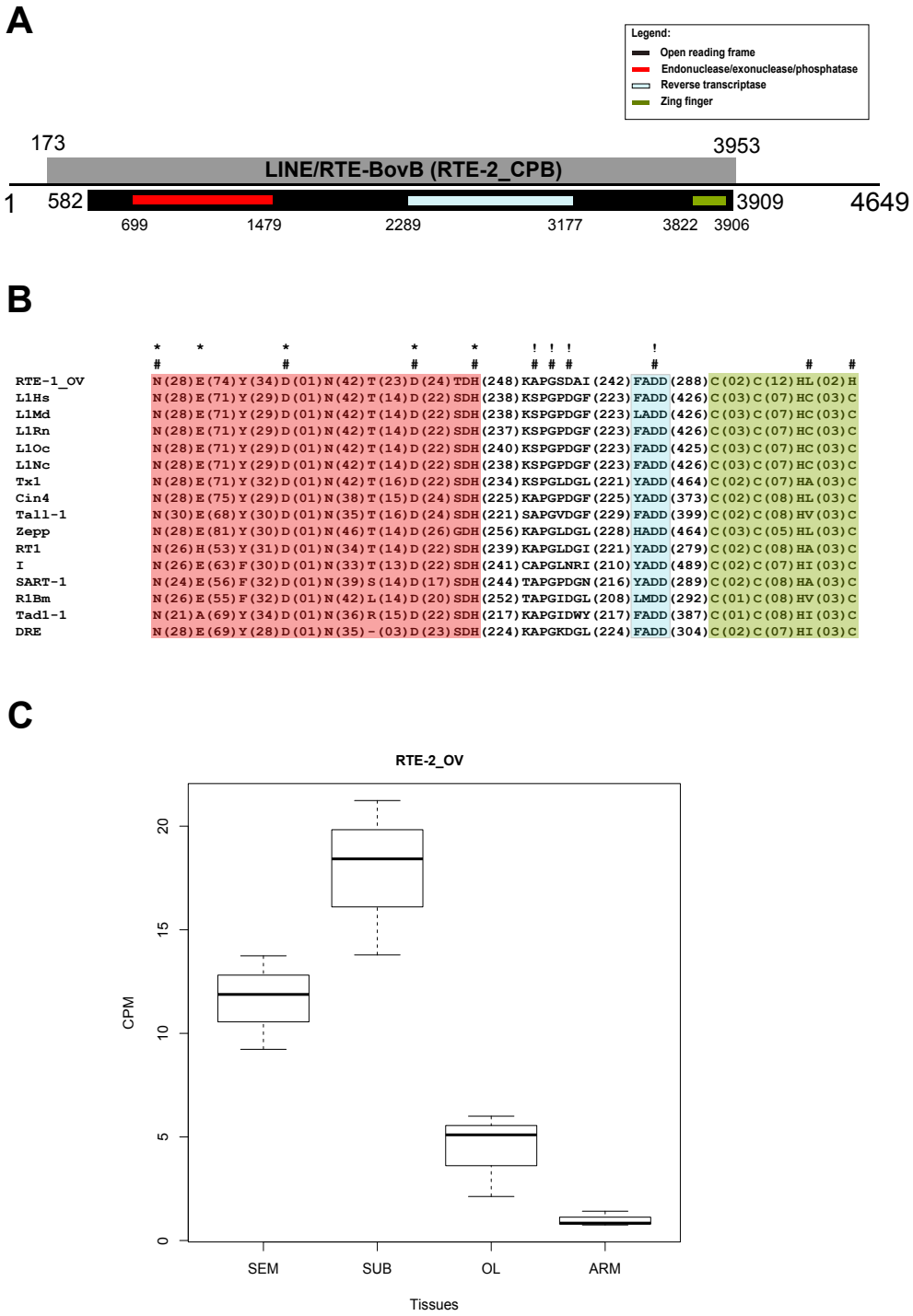

Figure 5.11 Full-length LINE element in the *Octopus* nervous transcriptome. (**A**) Assembled transcript containing the full-length LINE element (grey box). The open reading frame (ORF) (black box) includes endonuclease (red box), reverse transcriptase (cyan box) and zinger finger (green box) domains. (**B**) Comparison of the ORF proteins from 16 LINE elements. Numbers in parenthesis indicate the number of amino acids (aas) that separate the aas critical for *in vitro* endonuclease (EN) activity (*); aas critical for *in vitro* and *in vivo* reverse transcriptase (RT) activity (!); aas critical for retrotransposition (#). (**C**) Expression levels of the RTE-2_OV element in the supra- (SEM), sub- (SUB) esophageal masses, optic lobe (OL) and arm (ARM) tissues of the *Octopus* nervous system. On the y axis are the count per million (CPM), on the x axis are tissues of the *Octopus* nervous system.

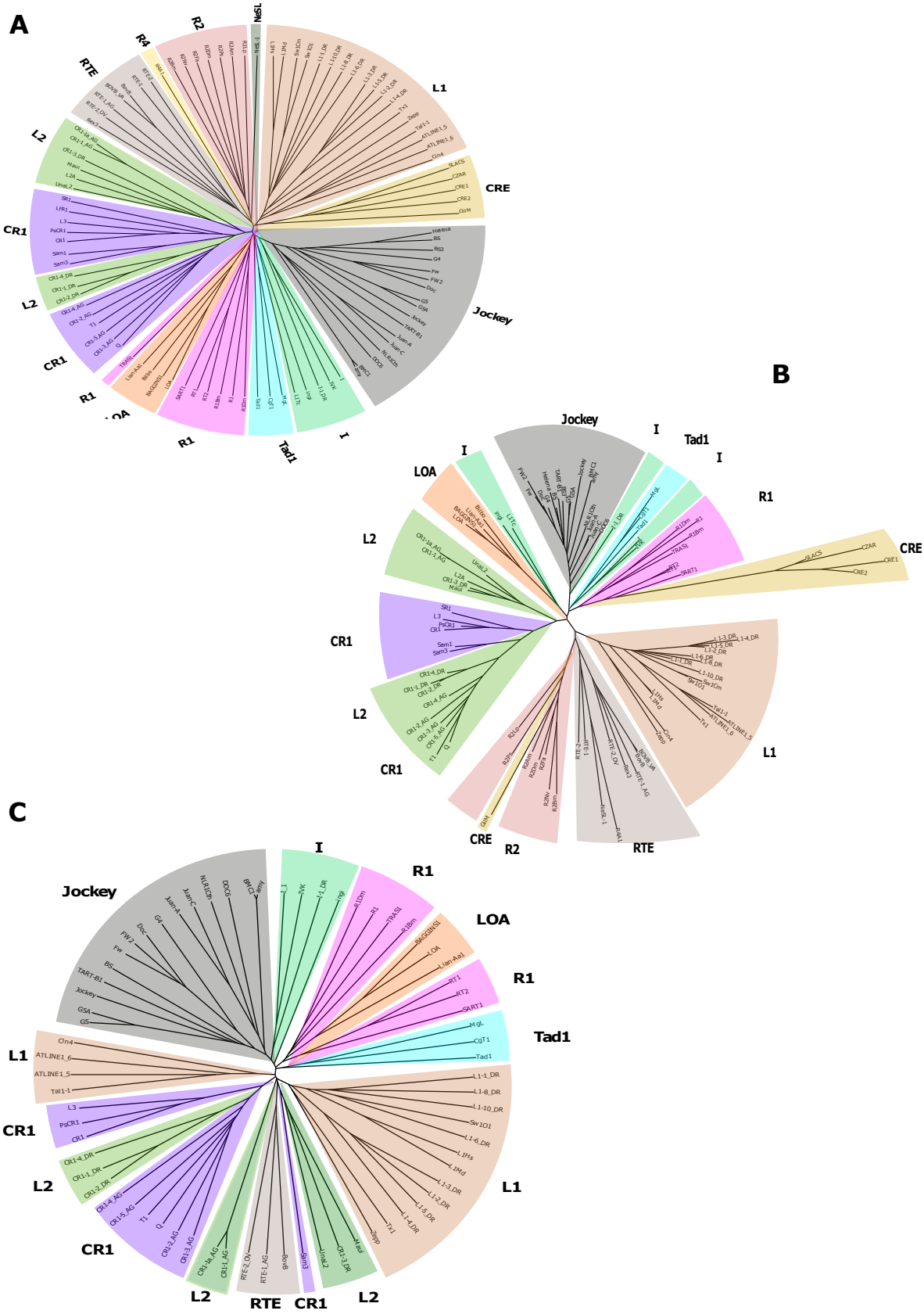

Figure 5.12 Phylogenetic trees of LINE clades. Phylogenetic trees for the amino acids sequences of the full-length (**A**), endonuclease (EN) (**B**) and reverse transcriptase (RT) (**C**) domains of known LINE families. The different colours indicate the LINE clades to whom belong the LINE elements.

5.7 Inspection of the CNS- and PNS-specific transcripts

In order to investigate the expression patterns of the octopus nervous system, I explored the transcripts expression levels of the main areas for the central and peripheral nervous systems. The transcripts with higher values in all the three biological replicates in a tissue than in others have been considered tissue-specific. This approach allowed me to identify 1,841 tissue-specific transcripts that are spread in the SEM (382), SUB (406), OL (657) for the CNS and in the ARM (396) for the PNS (**Figure 5.13**). Those transcripts are classified in putative coding (1,566) and long non-coding (275) transcripts based on the Annocript classification.

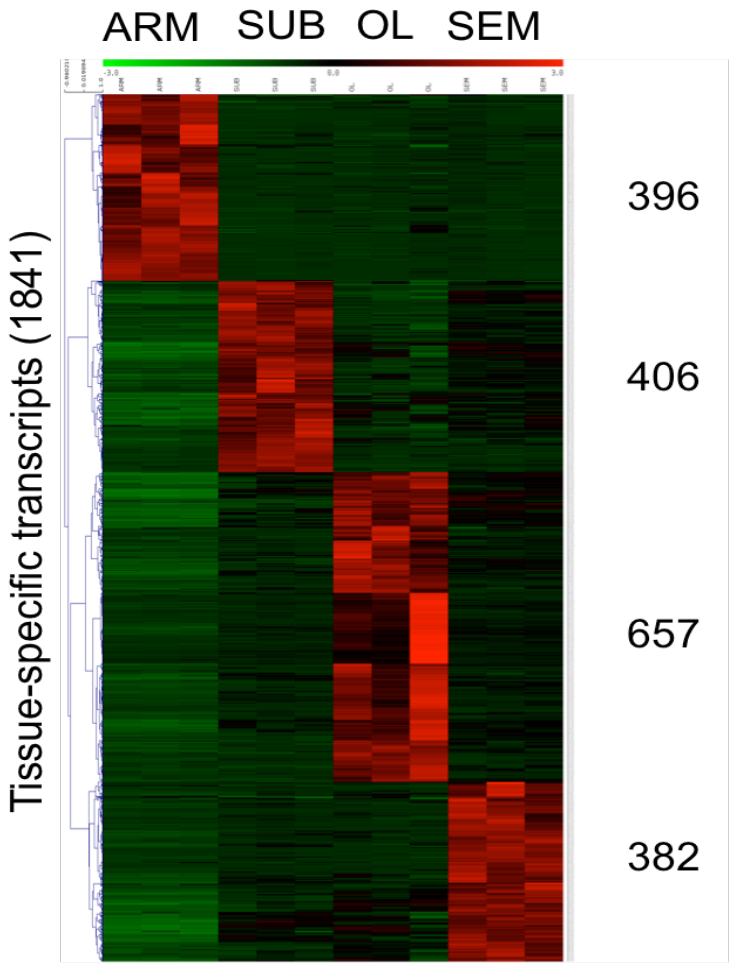

Figure 5.13 Heatmap of the *Octopus* nervous system tissues-specific transcripts. Hierarchical clustering of tissue-specific transcripts (rows) based on their relative expression levels in each sample (column). Red and green indicate high and low expression levels respectively.

Out of those coding, only 54 transcripts (3.5%) are annotated and are respectively distributed in SEM (5), SUB (12), OL (9) and ARM (28) tissues (**Table 5.5**).

Table 5.5 List of transcripts specifically expressed in the *Octopus* nervous system. The table lists tissue-specific transcripts with values greater than 1.5 count per million (CPM) in all the three biological replicates in a tissue and lesser in others. SEM: Supra-esophageal mass; SUB: Sub-esophageal mass; OL: Optic lobe; ARM: Arm.

Tissues	Gene symbol	SwissProt ID	UniRef ID	Description	Length
SEM	*PTF1A*	Q4ZHW1	UniRef90_K1QDK9	Pancreas transcription factor 1 subunit alpha	982
SEM	*ARX*	O42115	UniRef90_R7T3T5	Aristaless-related homeobox protein	1203
SEM	*N/A*	-	UniRef90_P91717	Transposase (Fragment)	545
SEM	*N/A*	-	UniRef90_UPI0003594A72	PREDICTED fukutin-related protein-like	1296
SEM	*DDC*	P14173	UniRef90_R7UY01	Aromatic-L-amino-acid decarboxylase	2070
SUB	*ERICH3*	Q5RHP9	UniRef90_R0LIX1	Uncharacterized protein C1orf173	1807
SUB	*EFCAB1*	Q3KQ77	UniRef90_A4UUI7	EF-hand calcium-binding domain-containing protein 1	811
SUB	*HOXB5A*	Q1KKX9	UniRef90_Q8WQS1	Homeobox protein Hox-B5a	627
SUB	*LRG1*	P02750	UniRef90_M1R4W2	Leucine-rich alpha-2-glycoprotein	1185
SUB	*TRPM2*	Q91YD4	UniRef90_UPI000359EB15	Transient receptor potential cation channel subfamily M member 2	1455
SUB	*HOX3*	P50901	UniRef90_Q6VS81	Homeobox protein HOX3	573
SUB	*MIB2*	Q5ZIJ9	UniRef90_K1PPV3	E3 ubiquitin-protein ligase MIB2	1411
SUB	*MIB2*	Q5ZIJ9	UniRef90_K1PPV3	E3 ubiquitin-protein ligase MIB2	1423
SUB	*N/A*	-	UniRef90_UPI00026542FF	UPI00026542FF related cluster	2343
SUB	*N/A*	-	UniRef90_UPI00026542FF	UPI00026542FF related cluster	2841
SUB	*GFI1B*	O42409	UniRef90_C3YTR0	Zinc finger protein Gfi-1b	614
SUB	*CGI_10024084*	-	UniRef90_K1RF32	Uncharacterized protein	458
OL	*WNT8B*	P51029	UniRef90_D1LXI5	Protein Wnt-8b	940
OL	*WNT8B*	P51029	UniRef90_H2T971	Protein Wnt-8b	1251
OL	*PHOX2B*	O35690	UniRef90_L7SWL8	Paired mesoderm homeobox protein 2B	680
OL	*LGR4*	Q9Z2H4	UniRef90_UPI00035A1CC6	Leucine-rich repeat-containing G-protein coupled receptor 4	2182
OL	*N/A*	-	UniRef90_UPI000359EF97	PREDICTED uncharacterized protein LOC101863547	983
OL	*N/A*	-	UniRef90_UPI000359EF97	PREDICTED uncharacterized protein LOC101863547	625
OL	*GGT3P*	A6NGU5	UniRef90_H0XKF4	Putative gamma-glutamyltranspeptidase 3	910
OL	*N/A*	-	UniRef90_K1QH48	Uncharacterized protein	562
OL	*C38C10.2*	Q03567	UniRef90_K1Q326	Uncharacterized transporter C38C10.2	915
ARM	*PAPL*	A5D6U8	UniRef90_F2UJ11	Iron/zinc purple acid phosphatase-like protein	1376
ARM	*TYR-3*	Q19673	UniRef90_UPI0001CBAFCE	Putative tyrosinase-like protein tyr-3	501
ARM	*CG32809*	Q7KW14	UniRef90_UPI000359D056	Coiled-coil domain-containing protein CG32809	2935
ARM	*TWIST2*	P97831	UniRef90_F7IYW0	Twist-related protein 2	1397
ARM	*TWK-18*	Q18120	UniRef90_R7U1G0	TWiK family of potassium channels protein 18	1231
ARM	*MEOX2*	P39021	UniRef90_K1Q1H4	Homeobox protein MOX-2	1134
ARM	*ORCT*	Q9VCA2	UniRef90_K1R3T8	Organic cation transporter protein	1420
ARM	*MEOX2*	P39021	UniRef90_K1Q1H4	Homeobox protein MOX-2	1205
ARM	*SLC22A4*	A9CB25	UniRef90_K1R3T8	Solute carrier family 22 member 4	1923
ARM	*SLC22A4*	Q9H015	UniRef90_K1R3T8	Solute carrier family 22 member 4	1796
ARM	*TTN*	A2ASS6	UniRef90_H2MVA6	Titin	326
ARM	*N/A*	-	UniRef90_I1GA96	Uncharacterized protein	778
ARM	*TCF15*	Q60539	UniRef90_K1S6P7	Transcription factor 15	1195

Tissues	Gene symbol	SwissProt ID	UniRef ID	Description	Length
ARM	*TTN*	Q8WZ42	UniRef90_K1QVQ5	Titin	640
ARM	*TTN*	A2ASS6	UniRef90_UPI00035960E3	Titin	915
ARM	*UNC-89*	O01761	UniRef90_E1FJE7	Muscle M-line assembly protein unc-89	1017
ARM	*UNC-22*	Q23551	UniRef90_K1QVQ5	Twitchin	254
ARM	*MYLK*	P11799	UniRef90_UPI00032913C4	Myosin light chain kinase smooth muscle	860
ARM	*UNC-89*	O01761	UniRef90_F1KPF2	Muscle M-line assembly protein unc-89	1297
ARM	*MYLK*	Q6PDN3	UniRef90_K1QVQ5	Myosin light chain kinase smooth muscle	2811
ARM	*N/A*	-	UniRef90_UPI0001CB9E99	UPI0001CB9E99 related cluster	441
ARM	*CAPTEDRAFT_122246*	-	UniRef90_R7VD42	Uncharacterized protein (Fragment)	408
ARM	*CAPTEDRAFT_106883*	-	UniRef90_R7UC16	Uncharacterized protein	438
ARM	*N/A*	-	UniRef90_UPI0001CB9E99	UPI0001CB9E99 related cluster	420
ARM	*CIT*	O14578	UniRef90_K1QJU8	Citron Rho-interacting kinase	1907
ARM	*ZNF276*	Q8N554	UniRef90_K1PJ99	Zinc finger protein 276	1122
ARM	*ORCT*	Q9VCA2	UniRef90_K1PJJ4	Organic cation transporter protein	1840
ARM	*N/A*	0	UniRef90_E2C9T0	Transposable element Tcb2 transposase (Fragment)	690

Interestingly, some of them encode homeobox proteins; these are an important transcription factors family, which plays a fundamental role in a diverse set of functions that include body plan specification, pattern formation and cell fate determination during metazoan development (Banerjee-Basu and Baxevanis, 2001). Four transcripts have been identified as homeobox-containing transcription factors that result specifically expressed in each tissue (**Table 5.6**).

The transcript for the Aristaless-related homeobox protein is expressed in the supra-esophageal mass of the *Octopus* central nervous system; it is encoded by *ARX* gene that was found to be expressed in the vertebrate central nervous system and in particular in the dorsal telencephalon and diencephalon of mouse embryos (Miura et al., 1997). The *HOXB5a* (homeobox B5a) transcript is expressed in the sub-esophageal mass; it belongs to a development regulatory system which acts as pattern regional tissue identity along the anterior-posterior axis during animal embryonic development (Lyon et al., 2013). The *PHOX2B* (Paired-like homeobox 2b) gene is expressed in the optic lobe tissue and it is required throughout the developing

sympathetic, parasympathetic and enteric ganglia, regulating the noradrenergic phenotype in vertebrates (Pattyn et al., 1999; Tiveron et al., 1996). The *MEOX2* (mesenchyme homeobox 2) transcript is expressed in the arm tissue that encode for a transcription factor with a role in the somitogenesis, a process by give rise the axial skeleton, skeletal muscles and dermis in vertebrates (Mankoo et al., 2003).

Growing evidence highlighted an intriguing relationship between transposable elements (TEs) and long non-coding RNAs (lncRNAs) in mammals genomes as well as in zebrafish (Kapusta et al., 2013; Kelley and Rinn, 2012). Moreover, a recent study identified in mice the *Uchl1-as1* lncRNA, an antisense transcript to the neuron-specific *Uchl1* gene, containing an embedded SINEB2 element that stimulates the *Uchl1* translation; removal of SINE2B element completely abrogated the translational effect of the transcript (Carrieri et al., 2012). Thereby, from the 275 tissue-specific long non coding RNAs, I searched the lncRNAs including retroelements based on the RepeatMasker annotation. The analysis revealed 22 lncRNAs transcripts (8%) embedded by SINE elements distributed in SEM (7), SUB (2), OL (6) and ARM (7) tissues, respectively. Finally, I manually selected one specific lncRNAs for each each tissue taking in consideration the length, the expression levels and the non-coding portrait score (**Table 5.6**). Both coding and long non-coding candidates are used to perform validations as reported in the 3.8 paragraph.

Table 5.6 List of coding and lncRNA candidate transcripts specifically expressed in the *Octopus* nervous system. The three CPM values for each transcript in each tissue represent the biological replicates. SEM: Supra-esophageal mass; SUB: Sub-esophageal mass; OL: Optic lobe; ARM: arm. CPM value: count per million value

Gene Symbol	Length	SEM (CPM values)			SUB (CPM values)			OL (CPM values)			ARM (CPM values)		
ARX	1203	10.04	10.39	12.80	0.17	0.49	0.15	0.23	0.20	0.07	0.00	0.00	0.04
HOXB5a	627	0.22	0.16	0.28	7.35	8.88	6.85	0.28	0.00	0.02	0.08	0.04	0.15
PHOX2b	680	0.18	0.16	0.36	0.37	0.31	0.34	1.81	1.57	2.15	0.00	0.00	0.04
MOX2	1134	0.13	0.06	0.04	0.02	0.00	0.00	0.00	0.04	0.15	3.28	2.01	2.61
LncRNAs	**Length**	**SEM (CPM values)**			**SUB (CPM values)**			**OL (CPM values)**			**ARM (CPM values)**		
SEML	475	1.29	0.71	1.03	0.32	0.18	0.34	0.46	0.31	0.35	0.06	0.01	0.00
SUBL	609	0.11	0.04	0.04	0.55	1.20	1.15	0.00	0.00	0.00	0.02	0.00	0.00
OLL	514	0.18	0.33	0.41	0.22	0.22	0.23	1.12	1.14	1.58	0.14	0.07	0.17
ARML	370	0.07	0.02	0.02	0.00	0.00	0.02	0.03	0.06	0.15	1.42	1.30	2.38

5.8 Experimental validations of candidate coding and lncRNAs

In order to validate the coding and long noncoding transcripts candidates, it has been performed a polymerase chain reaction (PCR) analysis (**Figure 5.14**). It allows to evaluate the expression levels of those transcripts in each tissue of the *Octopus* nervous system. The results show that *ARX* and *HOXB5a* expression is strongly enriched in SEM and SUB tissues, respectively, as expected. However, they also show a faint expression in the other tissues (**Figure 5.14 A**). *PHOX2b* transcript is expressed in all the CNS tissues with a slight increase in the optic lobe (OL). In contrast, *MEOX2* shows a specific, although low, expression in the arm (ARM) tissue.

Then, it has been validated the lncRNAs specifically expressed in the SEM (SEMl), SUB (SUBl), OL (OLl) for the CNS and ARM (ARMl) for the PNS (**Figure 5.14 B**). The SEML lncRNA has higher expression level in the SEM tissue than all the other tissue, although it exhibits a discrete expression also in the OL. The SUBl is highly expressed in the SUB tissue with a faint expression in the SEM.

Similar result is obtained for the OLl lncRNA that shows a slight expression in the SUB tissue. The ARMl, as occurring for the ARM-specific coding transcripts, has very

specific expression pattern. The PCR validations confirmed the tissue-specific expression patterns for both coding and long non-coding transcripts as observed by RNA-seq analysis.

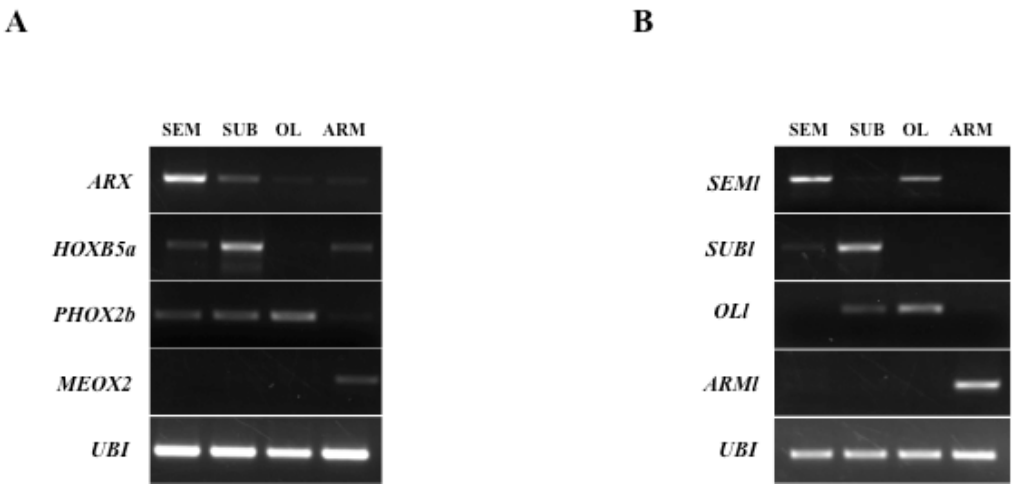

Figure 5.14 PCR analysis of coding and noncoding transcripts in the *Octopus* nervous system. Coding (**A**) and noncoding (**B**) transcripts expression levels have been evaluated by PCR analysis. The results are referred to a single animal but are representative of all the samples analyzed (N=3). The *(UBI*) transcripts has been used as internal control. The numbers indicate the amplicons length.

5.9 Exploring the *Octopus*-resources

The application of the next-generation sequencing (NGS) technologies has yielded an unprecedented wealth of data for the *Octopus* transcriptome. This has represented an important resource for research group I belong to providing a deeper understanding of molecular basis of the *O. vulgaris* nervous system. The research group (GFLab) of my supervisor, Dr. Graziano Fiorito, at the Stazione Zoologica is interested into the study of the behavioural plasticity and the learning capabilities of *Octopus vulgaris.* Therefore, the assembled transcriptome represents an invaluable resource to look at for the presence and at the sequence of genes associated with

these processes. One of them is the perception of the pain or nociception. The pain and nociception are two sides of the same coin; where nociception refers to the noxious stimuli arriving in the central nervous system resulting from activation of specialized sensory receptors called nociceptors that provide information about tissue damage (Julius and Basbaum, 2001). The pain is then defined as the unpleasant emotional experience that usually accompanies nociception. Almost all animals exhibit defensive responses to the noxious stimuli (Kavaliers, 1988; Walters, 1994). Nociceptors can be activated by mechanical, chemical (e.g. acid, capsaicin) and thermal stimuli and are classified as polymodal if an individual nociceptive neurone responds to multiple classes of stimulus Primary nociceptors encoding noxious stimuli have been reported in several invertebrate and vertebrate species; most of the research has focused on mammalian receptors, however. Outside of mammals, very little is known about nociceptors sensitive to stimuli. They have been described in selected invertebrates: *Caenorhabditis elegans*, *Drosophila melanogaster*, *Hirudo medicinalis* and *Aplysia californica* (Smith and Lewin, 2009). Cephalopods, even more octopuses, have the most neurally and behaviorally complex brain among the invertebrates, enforcing the opinion that cephalopods may experience pain through nociceptive alterations with interesting similarities to those reported in mammals. Recent studies have shown how cephalopods receive and process mechanical and chemical stimuli (Alupay et al., 2014; Crook et al., 2013; Hague et al., 2013). To better understand the molecular mechanisms of the nociception in *O. vulgaris*, GFLab selected from the literature molecules implicated as neurotransmitters in pain pathways in cephalopods and in other species. The research of those molecules among the annotated sequences resulted in the identification of hundreds of transcripts that have been filtered based on our interest in the 15 listed in the table

below (**Table 5.7**). Currently, we are validating the sequences obtained from the transcriptome by Sanger sequencing method and subsequently, we will evaluate the expression levels of those pain-related genes before and after noxious stimuli.

Table 5.7. List of candidate transcripts related to the nociception in the *Octopus* nervous system.

Gene Symbol	Description	Stimulus	Function	Reference
TRPA1	Transient receptor potential cation channel subfamily A member 1 homolog	Termal	Non-selective cation channel	Kwan et al. 2006, Chatzigeorgiou et al. 2010
P2X	P2X purinoceptor 4	Mechanical	Receptor for ATP that acts as a ligand-gated ion channel	Ulmann et al. 2008
PIEZO1	Piezo-type mechanosensitive ion channel component 1	Mechanical	Promotes endothelial cell organization and alignment in the direction of blood flow	Kim et al. 2012
PIEZO2	Piezo-type mechanosensitive ion channel component 2	Mechanical	Component of mechanosensitive channel required for rapidly adapting mechanically activated currents	Kim et al. 2012
ASIC1	Acid-sensing ion channel 1	Chemical (acidic pH)	Mediates glutamate-independent Ca2+ entry into neurons upon acidosis	Paukert et al. 2004
CGRP	Calcitonin gene-related peptide type 1 receptor	Chemical (CAP)	Vasodilation via cAMP pathway.	Li et al 2008
TKRP	Tachykinin-related peptide	Chemical (CAP)	Triggers the phospholipase C–inositol triphosphate–calcium signal transduction cascade via coupling to Gq receptors	Kunde at al. 2013
5HTr	5-hydroxytryptamine receptor	Any	Synaptic facilitation	Bogen et al. 2012
CAMKII	Calcium/calmodulin-dependent protein kinase kinase 2	Any	Lead to phosphorilation of TRPV1 (sensitization)	Vasileiou et al. 2009
CdK5	Cyclin-dependent kinase 5	Any	Involved in altering the MAPK pathway interacting with calcium calmodulin kinase II	Vasileiou et al. 2009
CPEB	Cytoplasmic polyadenylation element-binding protein 2	Any	Generation of pain memory in primary afferent nociceptors	Bogen et al. 2012
MAPK1	Mitogen-activated protein kinase kinase kinase 1	Any	Activation of signals that (by NGF, after TRPV1 act.) leads to a direct or indirect increase in the transcription of immediate early genes responsible for a transition from short-term adaptive processes to long-term hyperexcitability	Vasileiou et al. 2009, Ji et al. 2002
NMDA	Glutamate receptor ionotropic NMDA 2D	Any	Central sensitization	Lima et al. 2003
PgE2	Prostaglandin E2 receptor EP4 subtype	Any	Lead to phosphorilation of TRPV1 (sensitization)	Wang & Woolf 2005
TRP	Transient-receptor-potential-like protein	Any	Nociceptor signal transduction	Xu et al. 2000

Chapter 6 – Conclusions

Cephalopods have shown remarkable phenotypic plasticity and physiological divergence over the course of evolution. This dramatic change in behaviour became possible following the development of sophisticated motor, sensory, and cognitive capabilities, such as excellent vision, highly efficient flexible arms, and the ability to learn rapidly compared to other invertebrates (Borrelli and Fiorito, 2008). In comparison to lower molluscs, the Cephalopods show increased number and higher organization of nerve cells resulting in a brain with similar complexity to that of lower vertebrates. Indeed, the Cephalopod nervous system lies within the same range as vertebrate nervous systems smaller than birds and mammals but larger than fish and reptiles (Packard, 1972). The Cephalopod nervous system organization is most evolved in the common octopus, *Octopus vulgaris.* Therefore, *Octopus* is an ideal candidate for the analysis of molecular and neuronal mechanisms underlying complex behaviour in cephalopods. The lack of knowledge of the *Octopus* genome sequence is obviously a disadvantage (Albertin et al., 2012), and currently, the molecular resources of the *Octopus vulgaris* are limited at the transcriptome level (Ogura et al., 2004; Kocot et al., 2011; Smith et al., 2011; Zhang et al., 2012; Castellanos-Martínez et al., 2014). Besides, tissue diversity of transcriptomic data remains limited to the *Octopus* central nervous system (Zhang et al., 2012).

Here, I report the first reference transcriptome representing the principal domains in the Central nervous system (CNS) and Peripheral nervous system (PNS) of the *Octopus vulgaris.* The study aims to understand the molecular processes governing the organization, development and evolution of the *Octopus* nervous system (ONS).

I assembled a reference transcriptome of 64,477 transcripts with a median length of 795 bp. It is worth noting that the current transcriptome is at 2X the number and median length the previous published transcriptome (31,909 sequences with length ≥ 200 bp – median length: 362 bp). Moreover, the transcriptome shows a significantly higher percentage (97%) of complete core eukaryotic genes than reported previously (50%). The results indicate the current assembly to comprise of longer transcripts covering majority of *Octopus vulgaris* genes. I functionally annotated 21,030 (32,6%) transcripts, improving the 10,412 (17.4%) annotated sequences obtained from the CNS-transcriptome (Zhang et al., 2012).

Interestingly a high number of transcripts (7,806, 12,1%) were classified as long noncoding RNAs (lncRNAs) where more than 10% of transcripts expressed into the CNS are lncRNAs compared to about 7.3% in PNS. Thus, the *Octopus* lncRNAome shares an affinity to be expressed in the CNS similar to what is reported in higher vertebrates (Knauss and Sun, 2013; Mercer et al., 2008).GO term classification of the assembled transcriptome shows transcripts involved in retrotransposition to be enriched in the CNS with respect to the PNS. Thus, I decided to analyzed the repeat content in the *Octopus* transcriptome resulting in ~73% (46,900) of sequences to contain repeat elements. Notably, interspersed repeats are the major group where (~35.5%) with retroelements being the most frequent interspersed repeat (26.2%). Among the different classes of retroelements, I observed that Short INterspersed Elements (SINEs) are significantly enriched in transcripts expressed in the CNS, while a higher fraction of transcripts expressed in the PNS contains fragments from Long INterspersed Elements (LINEs), Long Terminal Repeats (LTRs) and DNA transposons – cut-and-paste, Mavericks and Helitrons –. It is interesting to note that SINEs are also the most significant retroelement enriched in lncRNAs in respect to the protein

coding genes. The results suggest that the organization of non-coding elements the *Octopus* neural transcriptome shows similarity to mammals indicating a convergent evolution of genomic features potentially playing a role in the evolution of nervous system (Kelley and Rinn, 2012).

Comparative analysis of the *Octopus* transcriptome highlighted a high frequency of transcript-embedded retroelements, similar to that of mammals and much higher than other vertebrates and invertebrates. Inspecting the proportions of transposable elements in the different species, SINEs, LINEs and LTRs are the most frequently embedded elements in the *Octopus* transcriptome after mammals while the three categories of DNA transposons are higher in the other invertebrates. Interestingly, other cephalopods also show a trend towards expansion of transposons consistently with previous observation (Yoshida et al., 2011). However, they are more distant from the *Octopus* and mammals because of their lower retroelement content and higher content of simple and low complexity repeats. The results indicate a selective pressure to retain retroelement expansion among cephalopods. Such expansion may lead to further exaptation of retroelements into functional non-coding transcripts which might play an important role in the evolution of a complex CNS in *Octopus vulgaris*.

I found a unique LINE element transcribed in the CNS samples with intact functional domains and amino acids required for the retrotransposition activity. I performed the phylogenetic analyses demonstrating that the *Octopus* LINE is conserved in the RTE clade, which highlights a relevant closeness to L1 clade in mammals. Currently, the research group driven by Dr. Fiorito is validating the LINE expression in the main areas of the CNS using in-situ hybridization assay.

An active LINE could also explain the expansion of non-autonomous SINEs in the *Octopus* nervous system. Hence, further studies could show the LINE is really active measuring its mobilization through southern blot (Muotri et al., 2005) and copy number variation (Coufal et al., 2009) assays, demonstrating that somatic retrotransposition occurs in the *Octopus* brain, as in mammals.

Subsequently, I studied the transcriptome dynamics of the *Octopus* nervous system selecting the tissue-specific transcripts and validating Hox transcription factors expression in each tissue for both CNS and PNS. Finally, I explored the transcriptome resource generated to identify a subset of transcripts involved in the nociception field to provide support of a current project from the laboratory of my supervisor Dr. Fiorito.

The findings of this study suggest a convergent evolutionary process, driven by retroelements and lncRNAs, has led to the evolution of mammalian-like molecular traits in the *Octopus* nervous system, contributing to develop its cognitive abilities, unique among invertebrates.

References

Abbott, N.J., and Pichon, Y. (1987). The glial blood-brain barrier of crustacea and cephalopods: a review. J. Physiol. (Paris) *82*, 304–313.

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185–2195.

Adelson, D.L., Raison, J.M., and Edgar, R.C. (2009). Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. Proc. Natl. Acad. Sci. U. S. A. *106*, 12855–12860.

Akagi, K., Li, J., Stephens, R.M., Volfovsky, N., and Symer, D.E. (2008). Extensive variation between inbred mouse strains due to endogenous L1 retrotransposition. Genome Res. *18*, 869–880.

Albertin, C.B., Bonnaud, L., Brown, C.T., Crookes-Goodson, W.J., Fonseca, R.R. da, Cristo, C.D., Dilkes, B.P., Edsinger-Gonzales, E., Robert J. Freeman, J., Hanlon, R.T., et al. (2012). Cephalopod Genomics: A Plan of Strategies and Organization. Stand. Genomic Sci. *7*.

Alupay, J.S., Hadjisolomou, S.P., and Crook, R.J. (2014). Arm injury produces long-term behavioral and neural hypersensitivity in octopus. Neurosci. Lett. *558*, 137–142.

Ambrozová, K., Mandáková, T., Bures, P., Neumann, P., Leitch, I.J., Koblízková, A., Macas, J., and Lysak, M.A. (2011). Diverse retrotransposon families and an AT-rich satellite DNA revealed in giant genomes of Fritillaria lilies. Ann. Bot. *107*, 255–268.

Arendt, D., Technau, U., and Wittbrodt, J. (2001). Evolution of the bilaterian larval foregut. Nature *409*, 81–85.

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

Athanikar, J.N., Badge, R.M., and Moran, J.V. (2004). A YY1-binding site is required for accurate human LINE-1 transcription initiation. Nucleic Acids Res. *32*, 3846–3855.

Baidouri, M.E., Carpentier, M.-C., Cooke, R., Gao, D., Lasserre, E., Llauro, C., Mirouze, M., Picault, N., Jackson, S.A., and Panaud, O. (2014). Widespread and frequent horizontal transfers of transposable elements in plants. Genome Res. *24*, 831–838.

Baillie, J.K., Barnett, M.W., Upton, K.R., Gerhardt, D.J., Richmond, T.A., De Sapio, F., Brennan, P.M., Rizzu, P., Smith, S., Fell, M., et al. (2011). Somatic retrotransposition alters the genetic landscape of the human brain. Nature *479*, 534–537.

Baker, M. (2011). Long noncoding RNAs: the search for function. Nat. Methods *8*, 379–383.

Banerjee-Basu, S., and Baxevanis, A.D. (2001). Molecular evolution of the homeodomain family of transcription factors. Nucleic Acids Res. *29*, 3258–3269.

Bassaglia, Y., Bekel, T., Da Silva, C., Poulain, J., Andouche, A., Navet, S., and Bonnaud, L. (2012). ESTs library from embryonic stages reveals tubulin and reflectin diversity in Sepia officinalis (Mollusca — Cephalopoda). Gene *498*, 203–211.

Batista, P.J., and Chang, H.Y. (2013). Long noncoding RNAs: cellular address codes in development and disease. Cell *152*, 1298–1307.

Bennetzen, J.L. (2005). Transposable elements, gene creation and genome rearrangement in flowering plants. Curr. Opin. Genet. Dev. *15*, 621–627.

Bennetzen, J.L., Ma, J., and Devos, K.M. (2005). Mechanisms of recent genome size variation in flowering plants. Ann. Bot. *95*, 127–132.

Bernard, D., Prasanth, K.V., Tripathi, V., Colasse, S., Nakamura, T., Xuan, Z., Zhang, M.Q., Sedel, F., Jourdren, L., Coulpier, F., et al. (2010). A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression. EMBO J. *29*, 3082–3093.

Bernstein, H.D., Zopf, D., Freymann, D.M., and Walter, P. (1993). Functional substitution of the signal recognition particle 54-kDa subunit by its Escherichia coli homolog. Proc. Natl. Acad. Sci. U. S. A. *90*, 5229–5233.

Berriman, M., Haas, B.J., LoVerde, P.T., Wilson, R.A., Dillon, G.P., Cerqueira, G.C., Mashiyama, S.T., Al-Lazikani, B., Andrade, L.F., Ashton, P.D., et al. (2009). The genome of the blood fluke Schistosoma mansoni. Nature *460*, 352–358.

Biémont, C., and Vieira, C. (2006). Genetics: Junk DNA as an evolutionary force. Nature *443*, 521–524.

Bingham, P.M., Kidwell, M.G., and Rubin, G.M. (1982). The molecular basis of P-M hybrid dysgenesis: the role of the P element, a P-strain-specific transposon family. Cell *29*, 995–1004.

Boeke, J.D., and Corces, V.G. (1989). Transcription and Reverse Transcription of Retrotransposons. Annu. Rev. Microbiol. *43*, 403–434.

Boeke, J.D., Garfinkel, D.J., Styles, C.A., and Fink, G.R. (1985). Ty elements transpose through an RNA intermediate. Cell *40*, 491–500.

Bone, Q., Pulsford, A., and Chubb, A.D. (1981). Squid mantle muscle. J. Mar. Biol. Assoc. U. K. *61*, 327–342.

Bonnaud, L., Bassaglia, Y., Bassaglia, Y., and Bassaglia, Y. (2014). Cephalopod development: what we can learn from differences. OA Biol. *2*, 6.

Borrelli, L., and Fiorito, G. (2008). 1.31 - Behavioral Analysis of Learning and Memory in Cephalopods. In Learning and Memory: A Comprehensive Reference, J.H. Byrne, ed. (Oxford: Academic Press), pp. 605–627.

Bourque, G., Leong, B., Vega, V.B., Chen, X., Lee, Y.L., Srinivasan, K.G., Chew, J.-L., Ruan, Y., Wei, C.-L., Ng, H.H., et al. (2008). Evolution of the mammalian transcription factor

binding repertoire via transposable elements. Genome Res. *18*, 1752–1762.

Bowen, N.J., and McDonald, J.F. (2001). Drosophila Euchromatic LTR Retrotransposons are Much Younger Than the Host Species in Which They Reside. Genome Res. *11*, 1527–1540.

Britten, R.J., and Kohne, D.E. (1968). Repeated Sequences in DNA. Science *161*, 529–540.

Broad Institute (2009). Aplysia Genome Project. Www.broadinstitute.org.

Brown, C.J., Hendrich, B.D., Rupert, J.L., Lafrenière, R.G., Xing, Y., Lawrence, J., and Willard, H.F. (1992). The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. Cell *71*, 527–542.

Budelmann, B.U. (1995). The cephalopod nervous system: What evolution has made of the molluscan design. In The Nervous Systems of Invertebrates: An Evolutionary and Comparative Approach, P.D.D.O. Breidbach, and P.D.W. Kutsch, eds. (Birkhäuser Basel), pp. 115–138.

Budelmann, B.U. (1996). Active marine predators: The sensory world of cephalopods. Mar. Freshw. Behav. Physiol. *27*, 59–75.

Bullock, T.H., and Horridge, G.A. (1965). Structure and function in the nervous systems of invertebrates (W. H. Freeman).

Bundo, M., Toyoshima, M., Okada, Y., Akamatsu, W., Ueda, J., Nemoto-Miyauchi, T., Sunaga, F., Toritsuka, M., Ikawa, D., Kakita, A., et al. (2014). Increased L1 Retrotransposition in the Neuronal Genome in Schizophrenia. Neuron *81*, 306–313.

Burke, W.D., Malik, H.S., Rich, S.M., and Eickbush, T.H. (2002). Ancient lineages of non-LTR retrotransposons in the primitive eukaryote, Giardia lamblia. Mol. Biol. Evol. *19*, 619–630.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. *25*, 1915–1927.

Cantrell, M.A., Scott, L., Brown, C.J., Martinez, A.R., and Wichman, H.A. (2008). Loss of LINE-1 activity in the megabats. Genetics *178*, 393–404.

Carl, M., Loosli, F., and Wittbrodt, J. (2002). Six3 inactivation reveals its essential role for the formation and patterning of the vertebrate eye. Dev. Camb. Engl. *129*, 4057–4063.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al. (2005). The transcriptional landscape of the mammalian genome. Science *309*, 1559–1563.

Carrieri, C., Cimatti, L., Biagioli, M., Beugnet, A., Zucchelli, S., Fedele, S., Pesce, E., Ferrer,

I., Collavin, L., Santoro, C., et al. (2012). Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. Nature *491*, 454–457.

Cartault, F., Munier, P., Benko, E., Desguerre, I., Hanein, S., Boddaert, N., Bandiera, S., Vellayoudom, J., Krejbich-Trotot, P., Bintner, M., et al. (2012). Mutation in a primate-conserved retrotransposon reveals a noncoding RNA as a mediator of infantile encephalopathy. Proc. Natl. Acad. Sci. U. S. A. *109*, 4980–4985.

Castellanos-Martínez, S., Arteta, D., Catarino, S., and Gestal, C. (2014). De Novo Transcriptome Sequencing of the Octopus vulgaris Hemocytes Using Illumina RNA-Seq Technology: Response to the Infection by the Gastrointestinal Parasite Aggregata octopiana. PLoS ONE *9*, e107873.

Cavalier-Smith, T. (1998). A revised six-kingdom system of life. Biol. Rev. Camb. Philos. Soc. *73*, 203–266.

C. elegans Sequencing Consortium (1998). Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. Science *282*, 2012–2018.

Chun, C.K., Scheetz, T.E., Bonaldo, M.F., Brown, B., Clemens, A., Crookes-Goodson, W.J., Crouch, K., DeMartini, T., Eyestone, M., Goodson, M.S., et al. (2006). An annotated cDNA library of juvenile Euprymna scolopes with and without colonization by the symbiont Vibrio fischeri. BMC Genomics *7*, 154.

Clements, A.P., and Singer, M.F. (1998). The human LINE-1 reverse transcriptase: Effect of deletions outside the common reverse transcriptase domain. Nucleic Acids Res. *26*, 3528–3535.

Coufal, N.G., Garcia-Perez, J.L., Peng, G.E., Yeo, G.W., Mu, Y., Lovci, M.T., Morell, M., O'Shea, K.S., Moran, J.V., and Gage, F.H. (2009). L1 retrotransposition in human neural progenitor cells. Nature *460*, 1127–1131.

Crook, R.J., Hanlon, R.T., and Walters, E.T. (2013). Squid have nociceptors that display widespread long-term sensitization and spontaneous activity after bodily injury. J. Neurosci. Off. J. Soc. Neurosci. *33*, 10021–10026.

DeGiorgis, J.A., Cavaliere, K.R., and Burbach, J.P.H. (2011). Identification of molecular motors in the Woods Hole squid, Loligo pealei: an expressed sequence tag approach. Cytoskelet. Hoboken NJ *68*, 566–577.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. *22*, 1775–1789.

Dewannieux, M., and Heidmann, T. (2005). L1-mediated retrotransposition of murine B1 and B2 SINEs recapitulated in cultured cells. J. Mol. Biol. *349*, 241–247.

Dewannieux, M., Esnault, C., and Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. Nat. Genet. *35*, 41–48.

Dilly, P.N. (1963). Delayed Responses in Octopus. J. Exp. Biol. *40*, 393–401.

Edelman, D.B., and Seth, A.K. (2009). Animal consciousness: a synthetic approach. Trends Neurosci. *32*, 476–484.

Elisaphenko, E.A., Kolesnikov, N.N., Shevchenko, A.I., Rogozin, I.B., Nesterova, T.B., Brockdorff, N., and Zakian, S.M. (2008). A dual origin of the Xist gene from a protein-coding gene and a set of transposable elements. PloS One *3*, e2521.

Faghihi, M.A., Modarresi, F., Khalil, A.M., Wood, D.E., Sahagan, B.G., Morgan, T.E., Finch, C.E., St Laurent, G., Kenny, P.J., and Wahlestedt, C. (2008). Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. Nat. Med. *14*, 723–730.

Fatica, A., and Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. Nat. Rev. Genet. *15*, 7–21.

Feng, Q., Moran, J.V., Kazazian, H.H., and Boeke, J.D. (1996). Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell *87*, 905–916.

Feschotte, C., and Wessler, S.R. (2001). Treasures in the attic: rolling circle transposons discovered in eukaryotic genomes. Proc. Natl. Acad. Sci. U. S. A. *98*, 8923–8924.

Finnegan, D.J. (1989). Eukaryotic transposable elements and genome evolution. Trends Genet. TIG *5*, 103–107.

Finnegan, D.J. (1992). Transposable elements. Curr. Opin. Genet. Dev. *2*, 861–867.

Finnegan, D.J. (2012). Retrotransposons. Curr. Biol. CB *22*, R432–R437.

Fiorito, G., and Chichery, R. (1995). Lesions of the vertical lobe impair visual discrimination learning by observation in Octopus vulgaris. Neurosci. Lett. *192*, 117–120.

Fiorito, G., and Scotto, P. (1992). Observational Learning in Octopus vulgaris. Science *256*, 545–547.

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

Fischer, S.E.J., Wienholds, E., and Plasterk, R.H.A. (2003). Continuous exchange of sequence information between dispersed Tc1 transposons in the Caenorhabditis elegans genome. Genetics *164*, 127–134.

Flash, T., and Hochner, B. (2005). Motor primitives in vertebrates and invertebrates. Curr. Opin. Neurobiol. *15*, 660–666.

Flockhart, R.J., Webster, D.E., Qu, K., Mascarenhas, N., Kovalski, J., Kretz, M., and

Khavari, P.A. (2012). BRAFV600E remodels the melanocyte transcriptome and induces BANCR to regulate melanoma cell migration. Genome Res. *22*, 1006–1014.

Florey, P.D.E., and Kriebel, D.M.E. (1969). Electrical and mechanical responses of chromatophore muscle fibers of the squid, Loligo opalescens, to nerve stimulation and drugs. Z. Für Vgl. Physiol. *65*, 98–130.

Ganko, E.W., Bhattacharjee, V., Schliekelman, P., and McDonald, J.F. (2003). Evidence for the contribution of LTR retrotransposons to C. elegans gene evolution. Mol. Biol. Evol. *20*, 1925–1931.

Gentles, A.J., Wakefield, M.J., Kohany, O., Gu, W., Batzer, M.A., Pollock, D.D., and Jurka, J. (2007). Evolutionary dynamics of transposable elements in the short-tailed opossum Monodelphis domestica. Genome Res. *17*, 992–1004.

Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E., et al. (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature *428*, 493–521.

Goff, S.A., Ricke, D., Lan, T.-H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., et al. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science *296*, 92–100.

Grahn, R.A., Rinehart, T.A., Cantrell, M.A., and Wichman, H.A. (2005). Extinction of LINE-1 activity coincident with a major mammalian radiation in rodents. Cytogenet. Genome Res. *110*, 407–415.

Grau, J.H., Poustka, A.J., Meixner, M., and Plötner, J. (2014). LTR retroelements are intrinsic components of transcriptional networks in frogs. BMC Genomics *15*, 626.

Graziadei, P. (1965). Muscle Receptors in Cephalopods. Proc. R. Soc. Lond. B Biol. Sci. *161*, 392–402.

Greilhuber, J., Borsch, T., Müller, K., Worberg, A., Porembski, S., and Barthlott, W. (2006). Smallest angiosperm genomes found in lentibulariaceae, with chromosomes of bacterial size. Plant Biol. Stuttg. Ger. *8*, 770–777.

Grote, P., Wittler, L., Hendrix, D., Koch, F., Währisch, S., Beisaw, A., Macura, K., Bläss, G., Kellis, M., Werber, M., et al. (2013). The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. Dev. Cell *24*, 206–214.

Gupta, S., Gallavotti, A., Stryker, G.A., Schmidt, R.J., and Lal, S.K. (2005). A novel class of Helitron-related transposable elements in maize contain portions of multiple pseudogenes. Plant Mol. Biol. *57*, 115–127.

Guttman, M., and Rinn, J.L. (2012). Modular regulatory principles of large non-coding RNAs. Nature *482*, 339–346.

Haapa-Paananen, S., Wahlberg, N., and Savilahti, H. (2014). Phylogenetic analysis of Maverick/Polinton giant transposons across organisms. Mol. Phylogenet. Evol. *78*, 271–274.

Hague, T., Florini, M., and Andrews, P.L.R. (2013). Preliminary in vitro functional evidence for reflex responses to noxious stimuli in the arms of Octopus vulgaris. J. Exp. Mar. Biol. Ecol. *447*, 100–105.

Halanych, K.M., Bacheller, J.D., Aguinaldo, A.M., Liva, S.M., Hillis, D.M., and Lake, J.A. (1995). Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. Science *267*, 1641–1643.

Hanlon, R.T., and Messenger, J.B. (1998). Cephalopod Behaviour (Cambridge University Press).

Hardie, R.C., and Raghu, P. (2001). Visual transduction in Drosophila. Nature *413*, 186–193.

Hejnol, A., and Martindale, M.Q. (2008). Acoel development indicates the independent evolution of the bilaterian mouth and anus. Nature *456*, 382–386.

Hobbs, M., and Young, J. (1973). Cephalopod Cerebellum. Brain Res. *55*, 424–430.

Hochner, B., Shomrat, T., and Fiorito, G. (2006). The octopus: a model for a comparative analysis of the evolution of learning and memory mechanisms. Biol. Bull. *210*, 308–317.

Holdt, L.M., Hoffmann, S., Sass, K., Langenberger, D., Scholz, M., Krohn, K., Finstermeier, K., Stahringer, A., Wilfert, W., Beutner, F., et al. (2013). Alu elements in ANRIL non-coding RNA at chromosome 9p21 modulate atherogenic cell functions through trans-regulation of gene networks. PLoS Genet. *9*, e1003588.

Houseley, J., Rubbi, L., Grunstein, M., Tollervey, D., and Vogelauer, M. (2008). A ncRNA modulates histone modification and mRNA induction in the yeast GAL gene cluster. Mol. Cell *32*, 685–695.

Huang, X., Lu, G., Zhao, Q., Liu, X., and Han, B. (2008). Genome-wide analysis of transposon insertion polymorphisms reveals intraspecific variation in cultivated rice. Plant Physiol. *148*, 25–40.

Huarte, M., and Rinn, J.L. (2010). Large non-coding RNAs: missing links in cancer? Hum. Mol. Genet. *19*, R152–R161.

Huffard, C.L. (2013). Cephalopod neurobiology: an introduction for biologists working in other model systems. Invertebr. Neurosci. IN *13*, 11–18.

Huffard, C.L., Boneka, F., and Full, R.J. (2005). Underwater bipedal locomotion by octopuses in disguise. Science *307*, 1927.

Ietswaart, R., Wu, Z., and Dean, C. (2012). Flowering time control: another window to the connection between antisense RNA and chromatin. Trends Genet. TIG *28*, 445–453.

Jeffs, A.R., Benjes, S.M., Smith, T.L., Sowerby, S.J., and Morris, C.M. (1998). The BCR gene recombines preferentially with Alu elements in complex BCR-ABL translocations

of chronic myeloid leukaemia. Hum. Mol. Genet. *7*, 767–776.

Jeon, Y., and Lee, J.T. (2011). YY1 tethers Xist RNA to the inactive X nucleation center. Cell *146*, 119–133.

Johnson, J.M., Edwards, S., Shoemaker, D., and Schadt, E.E. (2005). Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. Trends Genet. TIG *21*, 93–102.

Johnson, R., Teh, C.H.-L., Jia, H., Vanisri, R.R., Pandey, T., Lu, Z.-H., Buckley, N.J., Stanton, L.W., and Lipovich, L. (2009). Regulation of neural macroRNAs by the transcriptional repressor REST. RNA N. Y. N *15*, 85–96.

Julius, D., and Basbaum, A.I. (2001). Molecular mechanisms of nociception. Nature *413*, 203–210.

Jurka, J., Kapitonov, V.V., Kohany, O., and Jurka, M.V. (2007). Repetitive sequences in complex genomes: structure and evolution. Annu. Rev. Genomics Hum. Genet. *8*, 241–259.

Kapitonov, V.V., and Jurka, J. (1999). Molecular paleontology of transposable elements from Arabidopsis thaliana. Genetica *107*, 27–37.

Kapitonov, V.V., and Jurka, J. (2001). Rolling-circle transposons in eukaryotes. Proc. Natl. Acad. Sci. U. S. A. *98*, 8714–8719.

Kapitonov, V.V., and Jurka, J. (2006). Self-synthesizing DNA transposons in eukaryotes. Proc. Natl. Acad. Sci. U. S. A. *103*, 4540–4545.

Kapitonov, V.V., and Jurka, J. (2007). Helitrons on a roll: eukaryotic rolling-circle transposons. Trends Genet. TIG *23*, 521–529.

Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X., Ramsay, L., Bourque, G., Yandell, M., and Feschotte, C. (2013). Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet. *9*, e1003470.

Kavaliers, M. (1988). Evolutionary and comparative aspects of nociception. Brain Res. Bull. *21*, 923–931.

Kazazian, H.H. (2004). Mobile elements: drivers of genome evolution. Science *303*, 1626–1632.

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

Knauss, J.L., and Sun, T. (2013). Regulatory mechanisms of long noncoding RNAs in vertebrate central nervous system development and function. Neuroscience *235*, 200–214.

Kocot, K.M., Cannon, J.T., Todt, C., Citarella, M.R., Kohn, A.B., Meyer, A., Santos, S.R.,

Schander, C., Moroz, L.L., Lieb, B., et al. (2011). Phylogenomics reveals deep molluscan relationships. Nature *477*, 452–456.

Konkel, M.K., and Batzer, M.A. (2010). A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome. Semin. Cancer Biol. *20*, 211–221.

Kordis, D. (2009). Transposable elements in reptilian and avian (sauropsida) genomes. Cytogenet. Genome Res. *127*, 94–111.

Kröger, B., Vinther, J., and Fuchs, D. (2011). Cephalopod origin and evolution: A congruent picture emerging from fossils, development and molecules: Extant cephalopods are younger than previously realised and were under major selection to become agile, shell-less predators. BioEssays News Rev. Mol. Cell. Dev. Biol. *33*, 602–613.

Kumar, A., and Bennetzen, J.L. (1999). Plant retrotransposons. Annu. Rev. Genet. *33*, 479–532.

Kuwabara, T., Hsieh, J., Muotri, A., Yeo, G., Warashina, M., Lie, D.C., Moore, L., Nakashima, K., Asashima, M., and Gage, F.H. (2009). Wnt-mediated activation of NeuroD1 and retro-elements during adult neurogenesis. Nat. Neurosci. *12*, 1097–1105.

Lal, S.K., Giroux, M.J., Brendel, V., Vallejos, C.E., and Hannah, L.C. (2003). The maize genome contains a helitron insertion. Plant Cell *15*, 381–391.

Lampe, D.J., Witherspoon, D.J., Soto-Adames, F.N., and Robertson, H.M. (2003). Recent horizontal transfer of mellifera subfamily mariner transposons into insect lineages representing four different orders shows that selection acts only during horizontal transfer. Mol. Biol. Evol. *20*, 554–562.

Lee, J.T., and Bartolomei, M.S. (2013). X-inactivation, imprinting, and long noncoding RNAs in health and disease. Cell *152*, 1308–1323.

Levin, H.L., and Moran, J.V. (2011). Dynamic interactions between transposable elements and their hosts. Nat. Rev. Genet. *12*, 615–627.

Li, W., Prazak, L., Chatterjee, N., Grüninger, S., Krug, L., Theodorou, D., and Dubnau, J. (2013). Activation of transposable elements during aging and neuronal decline in Drosophila. Nat. Neurosci. *16*, 529–531.

Lipkow, K., Buisine, N., Lampe, D.J., and Chalmers, R. (2004). Early intermediates of mariner transposition: catalysis without synapsis of the transposon ends suggests a novel architecture of the synaptic complex. Mol. Cell. Biol. *24*, 8301–8311.

Lipovich, L., Dachet, F., Cai, J., Bagla, S., Balan, K., Jia, H., and Loeb, J.A. (2012). Activity-dependent human brain coding/noncoding gene regulatory networks. Genetics *192*, 1133–1148.

Llinás, R.R. (2003). The contribution of Santiago Ramon y Cajal to functional

neuroscience. Nat. Rev. Neurosci. *4*, 77–80.

Lohe, A.R., Moriyama, E.N., Lidholm, D.A., and Hartl, D.L. (1995). Horizontal transmission, vertical inactivation, and stochastic loss of mariner-like transposable elements. Mol. Biol. Evol. *12*, 62–72.

Lyon, R.S., Davis, A., and Scemama, J.-L. (2013). Spatio-temporal expression patterns of anterior Hox genes during Nile tilapia (Oreochromis niloticus) embryonic development. Gene Expr. Patterns GEP *13*, 104–108.

Malik, H.S., and Eickbush, T.H. (1998). The RTE class of non-LTR retrotransposons is widely distributed in animals and is the origin of many SINEs. Mol. Biol. Evol. *15*, 1123–1134.

Mankoo, B.S., Skuntz, S., Harrigan, I., Grigorieva, E., Candia, A., Wright, C.V.E., Arnheiter, H., and Pachnis, V. (2003). The concerted action of Meox homeobox genes is required upstream of genetic pathways essential for the formation, patterning and differentiation of somites. Dev. Camb. Engl. *130*, 4655–4664.

Martin, S.L. (1991). Ribonucleoprotein particles with LINE-1 RNA in mouse embryonal carcinoma cells. Mol. Cell. Biol. *11*, 4804–4807.

Maruyama, K., and Hartl, D.L. (1991). Evidence for interspecific transfer of the transposable element mariner between Drosophila and Zaprionus. J. Mol. Evol. *33*, 514–524.

Mathias, S.L., Scott, A.F., Kazazian, H.H., Boeke, J.D., and Gabriel, A. (1991). Reverse transcriptase encoded by a human transposable element. Science *254*, 1808–1810.

Matzner, H., Gutfreund, Y., and Hochner, B. (2000). Neuromuscular system of the flexible arm of the octopus: physiological characterization. J. Neurophysiol. *83*, 1315–1328.

McClintock, B. (1950). The Origin and Behavior of Mutable Loci in Maize. Proc. Natl. Acad. Sci. U. S. A. *36*, 344–355.

McClintock, B. (1987). The discovery of characterization of transposable elements: the collected papers of Barbara McClintock. Genes.

Mercer, T.R., Dinger, M.E., Sunkin, S.M., Mehler, M.F., and Mattick, J.S. (2008). Specific expression of long noncoding RNAs in the mouse brain. Proc. Natl. Acad. Sci. U. S. A. *105*, 716–721.

Mercer, T.R., Dinger, M.E., and Mattick, J.S. (2009). Long non-coding RNAs: insights into functions. Nat. Rev. Genet. *10*, 155–159.

Messenger, J.B. (1967a). The peduncle lobe: a visuo-motor centre in octopus. Proc. R. Soc. Lond. Ser. B Contain. Pap. Biol. Character R. Soc. G. B. *167*, 225–251.

Messenger, J.B. (1967b). The effects on locomotion of lesions to the visuo-motor system in octopus. Proc. R. Soc. Lond. Ser. B Contain. Pap. Biol. Character R. Soc. G. B.

167, 252–281.

Messenger, J.B. (1996). Neurotransmitters of cephalopods. Invert. Neurosci. *2*, 95–114.

Miller, W.J., Hagemann, S., Reiter, E., and Pinsker, W. (1992). P-element homologous sequences are tandemly repeated in the genome of Drosophila guanche. Proc. Natl. Acad. Sci. U. S. A. *89*, 4018–4022.

Milligan, B., Curtin, N., and Bone, Q. (1997). Milligan BJ, Curtin NA, Bone Q. A thin-slice preparation of cuttlefish mantle muscle for study of contraction. J. Physiol.-Lond. 504P 3-4.

Miura, H., Yanazawa, M., Kato, K., and Kitamura, K. (1997). Expression of a novel aristaless related homeobox gene "Arx" in the vertebrate telencephalon, diencephalon and floor plate. Mech. Dev. *65*, 99–109.

Muñoz-López, M., and García-Pérez, J.L. (2010). DNA Transposons: Nature and Applications in Genomics. Curr. Genomics *11*, 115–128.

Muntz, W.R.A. (1999). Visual systems, behaviour, and environment in cephalopods. In Adaptive Mechanisms in the Ecology of Vision, S.N. Archer, M.B.A. Djamgoz, E.R. Loew, J.C. Partridge, and S. Vallerga, eds. (Springer Netherlands), pp. 467–483.

Muotri, A.R., Chu, V.T., Marchetto, M.C.N., Deng, W., Moran, J.V., and Gage, F.H. (2005a). Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature *435*, 903–910.

Muotri, A.R., Chu, V.T., Marchetto, M.C.N., Deng, W., Moran, J.V., and Gage, F.H. (2005b). Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature *435*, 903–910.

Muotri, A.R., Marchetto, M.C.N., Coufal, N.G., Oefner, R., Yeo, G., Nakashima, K., and Gage, F.H. (2010). L1 retrotransposition in neurons is modulated by MeCP2. Nature *468*, 443–446.

Musacchia, F., Basu, S., Petrosino, G., Salvemini, M., and Sanges, R. (2015). Annocript: a flexible pipeline for the annotation of transcriptomes also able to identify putative long noncoding RNAs. Bioinformatics btv106.

Ng, S.-Y., Johnson, R., and Stanton, L.W. (2012). Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. EMBO J. *31*, 522–533.

Niehrs, C. (2010). On growth and form: a Cartesian coordinate system of Wnt and BMP signaling specifies bilaterian body axes. Development *137*, 845–857.

Nixon and Young, M. (2003). The brains and lives of cephalopods (Oxford University Press,).

O'Dor, R.K., and Webber, D.M. (1986). The constraints on cephalopods: why squid

aren't fish. Can. J. Zool. *64*, 1591–1605.

Ogura, A., Ikeo, K., and Gojobori, T. (2004). Comparative Analysis of Gene Expression for Convergent Evolution of Camera Eye Between Octopus and Human. Genome Res. *14*, 1555–1561.

Ohshima, K., Hamada, M., Terai, Y., and Okada, N. (1996). The 3' ends of tRNA-derived short interspersed repetitive elements are derived from the 3' ends of long interspersed repetitive elements. Mol. Cell. Biol. *16*, 3756–3764.

Okada, N., and Hamada, M. (1997). The 3′ ends of tRNA-derived SINEs originated from the 3' ends of LINEs: A new example from the bovine genome. J. Mol. Evol. *44*, S52–S56.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., et al. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature *420*, 563–573.

O'Neil, J., Tchinda, J., Gutierrez, A., Moreau, L., Maser, R.S., Wong, K.-K., Li, W., McKenna, K., Liu, X.S., Feng, B., et al. (2007). Alu elements mediate MYB gene tandem duplication in human T-ALL. J. Exp. Med. *204*, 3059–3066.

Owen (1832). Memoir on the pearly nautilus (Nautilus pompilius, Linn). With illustrations of its external form and internal structure. R. Coll. Surg. Lond.

Packard (1988). The Mollusca, Vol. 11. Form and Function (San Diego: Academic Press).

Packard, A. (1972). Cephalopods and Fish: The Limits of Convergence. Biol. Rev. *47*, 241–307.

Parisod, C., Holderegger, R., and Brochmann, C. (2010). Evolutionary consequences of autopolyploidy. New Phytol. *186*, 5–17.

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

Passamaneck, Y.J., Schander, C., and Halanych, K.M. (2004). Investigation of molluscan phylogeny using large-subunit and small-subunit nuclear rRNA sequences. Mol. Phylogenet. Evol. *32*, 25–38.

Pattyn, A., Morin, X., Cremer, H., Goridis, C., and Brunet, J.F. (1999). The homeobox gene Phox2b is essential for the development of autonomic neural crest derivatives. Nature *399*, 366–370.

Pauli, A., Valen, E., Lin, M.F., Garber, M., Vastenhouw, N.L., Levin, J.Z., Fan, L., Sandelin, A., Rinn, J.L., Regev, A., et al. (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res. *22*, 577–591.

Pereira, V. (2004). Insertion bias and purifying selection of retrotransposons in the Arabidopsis thaliana genome. Genome Biol. *5*, R79.

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

Pfefferkorn, A. (1915). Das Nervensystem der Octopoden (W. Engelmann).

Plän, T. (1987). Funktionelle Neuroanatomie sensorisch-motorischer Loben im Gehirn von Octopus vulgaris. PhD Thesis Univesität Regensbg. Ger.

Ponder, W.F., and Lindberg, D.R. (2008). Phylogeny and Evolution of the Mollusca (University of California Press).

Ponjavic, J., Oliver, P.L., Lunter, G., and Ponting, C.P. (2009a). Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. PLoS Genet. *5*, e1000617.

Ponjavic, J., Oliver, P.L., Lunter, G., and Ponting, C.P. (2009b). Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. PLoS Genet. *5*, e1000617.

Ponte, G., Dröscher, A., and Fiorito, G. (2013). Fostering cephalopod biology research: past and current trends and topics. Invertebr. Neurosci. IN *13*, 1–9.

Porter, F.D., Drago, J., Xu, Y., Cheema, S.S., Wassif, C., Huang, S.P., Lee, E., Grinberg, A., Massalas, J.S., Bodine, D., et al. (1997). Lhx2, a LIM homeobox gene, is required for eye, forebrain, and definitive erythrocyte development. Dev. Camb. Engl. *124*, 2935–2944.

Prak, E.T., and Kazazian, H.H. (2000). Mobile elements and the human genome. Nat. Rev. Genet. *1*, 134–144.

Pritham, E.J., and Feschotte, C. (2007). Massive amplification of rolling-circle transposons in the lineage of the bat Myotis lucifugus. Proc. Natl. Acad. Sci. U. S. A. *104*, 1895–1900.

Qureshi, I.A., and Mehler, M.F. (2012). Emerging roles of non-coding RNAs in brain evolution, development, plasticity and disease. Nat. Rev. Neurosci. *13*, 528–541.

Ray, D.A., Feschotte, C., Pagan, H.J.T., Smith, J.D., Pritham, E.J., Arensburger, P., Atkinson, P.W., and Craig, N.L. (2008). Multiple waves of recent DNA transposon activity in the bat, Myotis lucifugus. Genome Res. *18*, 717–728.

Reeves, M.B., Davies, A.A., McSharry, B.P., Wilkinson, G.W., and Sinclair, J.H. (2007). Complex I binding by a virally encoded RNA regulates mitochondria-induced cell death. Science *316*, 1345–1348.

Richardson, J.M., Colloms, S.D., Finnegan, D.J., and Walkinshaw, M.D. (2009). Molecular architecture of the Mos1 paired-end complex: the structural basis of DNA transposition in a eukaryote. Cell *138*, 1096–1108.

Rinn, J.L., and Chang, H.Y. (2012). Genome regulation by long noncoding RNAs. Annu. Rev. Biochem. *81*, 145–166.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, H., Helms, J.A., Farnham, P.J., Segal, E., et al. (2007). Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Non-Coding RNAs. Cell *129*, 1311–1323.

Da Rocha, S.T., Boeva, V., Escamilla-Del-Arenal, M., Ancelin, K., Granier, C., Matias, N.R., Sanulli, S., Chow, J., Schulz, E., Picard, C., et al. (2014). Jarid2 Is Implicated in the Initial Xist-Induced Targeting of PRC2 to the Inactive X Chromosome. Mol. Cell *53*, 301–316.

Ruppert, E.E., Fox, R.S., and Barnes, R.D. (2003). Invertebrate Zoology: A Functional Evolutionary Approach (Belmont, CA: Cengage Learning).

Saha, S., Bridges, S., Magbanua, Z.V., and Peterson, D.G. (2008). Empirical comparison of ab initio repeat finding programs. Nucleic Acids Res. *36*, 2284–2294.

Saidel, W.M. (1982). Connections of the octopus optic lobe: an HRP study. J. Comp. Neurol. *206*, 346–358.

SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., et al. (1996). Nested retrotransposons in the intergenic regions of the maize genome. Science *274*, 765–768.

Santoni, F.A., Guerra, J., and Luban, J. (2012). HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. Retrovirology *9*, 111.

Schaack, S., Gilbert, C., and Feschotte, C. (2010). Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends Ecol. Evol. *25*, 537–546.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science *326*, 1112–1115.

Shaffer, H.B., Minx, P., Warren, D.E., Shedlock, A.M., Thomson, R.C., Valenzuela, N., Abramyan, J., Amemiya, C.T., Badenhorst, D., Biggar, K.K., et al. (2013). The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. Genome Biol. *14*, R28.

Shigeno, S., Sasaki, T., Moritaki, T., Kasugai, T., Vecchione, M., and Agata, K. (2008). Evolution of the cephalopod head complex by assembly of multiple molluscan body parts: Evidence from Nautilus embryonic development. J. Morphol. *269*, 1–17.

Shigeno, S., Takenori, S., and Von Boletzky, S. (2010). The origins of cephalopod body plans: A geometrical and developmental basis for the evolution of vertebrate-like organ systems. Tokai Univ. Press Tokyo p. 23–34.

Simakov, O., Marletaz, F., Cho, S.-J., Edsinger-Gonzales, E., Havlak, P., Hellsten, U., Kuo, D.-H., Larsson, T., Lv, J., Arendt, D., et al. (2013). Insights into bilaterian evolution from three spiralian genomes. Nature *493*, 526–531.

De Sio, F. (2011). Leviathan and the Soft Animal: Medical Humanism and the Invertebrate Models for Higher Nervous Functions, 1950s–90s. Med. Hist. *55*, 369–374.

Smith, E.S.J., and Lewin, G.R. (2009). Nociceptors: a phylogenetic view. J. Comp. Physiol. A Neuroethol. Sens. Neural. Behav. Physiol. *195*, 1089–1106.

Smith, M.R., and Caron, J.-B. (2010). Primitive soft-bodied cephalopods from the Cambrian. Nature *465*, 469–472.

Smith, S.A., Wilson, N.G., Goetz, F.E., Feehery, C., Andrade, S.C.S., Rouse, G.W., Giribet, G., and Dunn, C.W. (2011a). Resolving the evolutionary relationships of molluscs with phylogenomic tools. Nature *480*, 364–367.

Smith, S.A., Wilson, N.G., Goetz, F.E., Feehery, C., Andrade, S.C.S., Rouse, G.W., Giribet, G., and Dunn, C.W. (2011b). Resolving the evolutionary relationships of molluscs with phylogenomic tools. Nature *480*, 364–367.

St George-Hyslop, P., and Haass, C. (2008). Regulatory RNA goes awry in Alzheimer's disease. Nat. Med. *14*, 711–712.

Sumbre, G., Fiorito, G., Flash, T., and Hochner, B. (2006). Octopuses Use a Human-like Strategy to Control Precise Point-to-Point Arm Movements. Curr. Biol. *16*, 767–772.

Sun, C., Shepard, D.B., Chong, R.A., López Arriaza, J., Hall, K., Castoe, T.A., Feschotte, C., Pollock, D.D., and Mueller, R.L. (2012). LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders. Genome Biol. Evol. *4*, 168–183.

Swiezewski, S., Liu, F., Magusin, A., and Dean, C. (2009). Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. Nature *462*, 799–802.

Szemraj, J., Płucienniczak, G., Jaworski, J., and Płucienniczak, A. (1995). Bovine Alu-like sequences mediate transposition of a new site-specific retroelement. Gene *152*, 261–264.

Taft, R.J., Pang, K.C., Mercer, T.R., Dinger, M., and Mattick, J.S. (2010). Non-coding RNAs: regulators of disease. J. Pathol. *220*, 126–139.

Takeuchi, T., Kawashima, T., Koyanagi, R., Gyoja, F., Tanaka, M., Ikuta, T., Shoguchi, E., Fujiwara, M., Shinzato, C., Hisata, K., et al. (2012). Draft genome of the pearl oyster Pinctada fucata: a platform for understanding bivalve biology. DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes *19*, 117–130.

Tchénio, T., Casella, J.F., and Heidmann, T. (2000). Members of the SRY family regulate the human LINE retrotransposons. Nucleic Acids Res. *28*, 411–415.

Telford, M.J., and Copley, R.R. (2011). Improving animal phylogenies with genomic data. Trends Genet. TIG *27*, 186–195.

Telford, M.J., Bourlat, S.J., Economou, A., Papillon, D., and Rota-Stabelli, O. (2008). The evolution of the Ecdysozoa. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *363*, 1529–1537.

Thompson, J.T., and Voight, J.R. (2003). Erectile tissue in an invertebrate animal: the Octopus copulatory organ. J. Zool. *261*, 101–108.

Tiveron, M.C., Hirsch, M.R., and Brunet, J.F. (1996). The expression pattern of the transcription factor Phox2 delineates synaptic pathways of the autonomic nervous system. J. Neurosci. Off. J. Soc. Neurosci. *16*, 7649–7660.

Treangen, T.J., and Salzberg, S.L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. *13*, 36–46.

Tricarico, E., Borrelli, L., Gherardi, F., and Fiorito, G. (2011). I Know My Neighbour: Individual Recognition in Octopus vulgaris. PLoS ONE *6*.

Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., and Bartel, D.P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell *147*, 1537–1550.

Vance, K.W., Sansom, S.N., Lee, S., Chalei, V., Kong, L., Cooper, S.E., Oliver, P.L., and Ponting, C.P. (2014). The long non-coding RNA Paupar regulates the expression of both local and distal genes. EMBO J. *33*, 296–311.

Vassetzky, N.S., and Kramerov, D.A. (2013). SINEBase: a database and tool for SINE analysis. Nucleic Acids Res. *41*, D83–D89.

Vinther, J. (2015). The origins of molluscs. Palaeontology *58*, 19–34.

Volff, J.N., Körting, C., Sweeney, K., and Schartl, M. (1999). The non-LTR retrotransposon Rex3 from the fish Xiphophorus is widespread among teleosts. Mol. Biol. Evol. *16*, 1427–1438.

Walsh, A.M., Kortschak, R.D., Gardner, M.G., Bertozzi, T., and Adelson, D.L. (2013). Widespread horizontal transfer of retrotransposons. Proc. Natl. Acad. Sci. U. S. A. *110*, 1012–1016.

Walters, E.T. (1994). Injury-Related Behavior and Neuronal Plasticity: an Evolutionary Perspective on Sensitization, Hyperalgesia, and Analgesia. In International Review of Neurobiology, R.J.B. and R.A. Harris, ed. (Academic Press), pp. 325–427.

Wang, K.C., Yang, Y.W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B.R., Protacio, A., Flynn, R.A., Gupta, R.A., et al. (2011). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature *472*, 120–124.

Wells, M.J. (1994). The evolution of a racing snail.

Wells, M.J., and O'Dor, R.K. (1991). Jet propulsion and the evolution of cephalopods. Bull Mar Sci 49419-432.

Wells, M.J., and Wells, J. (1959). Hormonal Control of Sexual Maturity in Octopus. J. Exp. Biol. *36*, 1–33.

Willingham, A.T., Orth, A.P., Batalov, S., Peters, E.C., Wen, B.G., Aza-Blanc, P., Hogenesch, J.B., and Schultz, P.G. (2005). A Strategy for Probing the Function of Noncoding RNAs Finds a Repressor of NFAT. Science *309*, 1570–1573.

Wilusz, J.E., Sunwoo, H., and Spector, D.L. (2009). Long noncoding RNAs: functional surprises from the RNA world. Genes Dev. *23*, 1494–1504.

Woodhams, P.L. (1977). The ultrastructure of a cerebellar analogue in octopus. J. Comp. Neurol. *174*, 329–345.

Wutz, A. (2011). Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation. Nat. Rev. Genet. *12*, 542–553.

Wutz, A., Rasmussen, T.P., and Jaenisch, R. (2002). Chromosomal silencing and localization are mediated by different domains of Xist RNA. Nat. Genet. *30*, 167–174.

Yang, N., Zhang, L., Zhang, Y., and Kazazian, H.H. (2003). An important role for RUNX3 in human L1 transcription and retrotransposition. Nucleic Acids Res. *31*, 4929–4940.

Yoshida, M., Ishikura, Y., Moritaki, T., Shoguchi, E., Shimizu, K.K., Sese, J., and Ogura, A. (2011). Genome structure analysis of molluscs revealed whole genome duplication and lineage specific repeat variation. Gene *483*, 63–71.

Yoshida, M., Yura, K., and Ogura, A. (2014). Cephalopod eye evolution was modulated by the acquisition of Pax-6 splicing variants. Sci. Rep. *4*.

Young, J.Z. (1963). The Number and Sizes of Nerve Cells in Octopus. Proc. Zool. Soc. Lond. *140*, 229–254.

Young, J.Z. (1971). The anatomy of the nervous system of Octopus vulgaris (Clarendon Press).

Young, J.Z. (1976). The “cerebellum” and the control of eye movements in cephalopods. Nature *264*, 572–574.

Young, J.Z. (1977). The Nervous System of Loligo: III. Higher Motor Centres: The Basal Supraoesophageal Lobes. Philos. Trans. R. Soc. Lond. B Biol. Sci. *276*, 351–398.

Young, J.Z. (1985). Cephalopods and Neuroscience. Biol. Bull. *168*, 153–158.

Young, N.D., Jex, A.R., Li, B., Liu, S., Yang, L., Xiong, Z., Li, Y., Cantacessi, C., Hall, R.S., Xu, X., et al. (2012). Whole-genome sequence of Schistosoma haematobium. Nat. Genet. *44*, 221–225.

Young, T.L., Matsuda, T., and Cepko, C.L. (2005). The noncoding RNA taurine upregulated gene 1 is required for differentiation of the murine retina. Curr. Biol. CB *15*, 501–512.

Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., Yang, P., Zhang, L., Wang, X., Qi, H., et al. (2012a). The oyster genome reveals stress adaptation and complexity of shell formation. Nature *490*, 49–54.

Zhang, X., Mao, Y., Huang, Z., Qu, M., Chen, J., Ding, S., Hong, J., and Sun, T. (2012b). Transcriptome analysis of the Octopus vulgaris central nervous system. PloS One *7*, e40320.

Zhang, Y., Maksakova, I.A., Gagnier, L., van de Lagemaat, L.N., and Mager, D.L. (2008). Genome-wide assessments reveal extremely high levels of polymorphism of two active families of mouse endogenous retroviral elements. PLoS Genet. *4*, e1000007.

Zhao, H., and Bourque, G. (2009). Recovering genome rearrangements in the mammalian phylogeny. Genome Res. *19*, 934–942.

Zhao, J., Sun, B.K., Erwin, J.A., Song, J.-J., and Lee, J.T. (2008). Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science *322*, 750–756.

Zhou, Y., Zheng, H., Chen, Y., Zhang, L., Wang, K., Guo, J., Huang, Z., Zhang, B., Huang, W., Jin, K., et al. (2009). The Schistosoma japonicum genome reveals features of host–parasite interplay. Nature *460*, 345–351.

Ziats, M.N., and Rennert, O.M. (2013). Aberrant expression of long noncoding RNAs in autistic brain. J. Mol. Neurosci. MN *49*, 589–593.

Ziolkowski, P.A., Koczyk, G., Galganski, L., and Sadowski, J. (2009). Genome sequence comparison of Col and Ler lines reveals the dynamic nature of Arabidopsis chromosomes. Nucleic Acids Res. *37*, 3189–3201.

Zuccolo, A., Sebastian, A., Talag, J., Yu, Y., Kim, H., Collura, K., Kudrna, D., and Wing, R.A. (2007). Transposable element distribution, abundance and role in genome size variation in the genus Oryza. BMC Evol. Biol. *7*, 152.

Župunski, V., Gubenšek, F., and Kordis, D. (2001). Evolutionary Dynamics and Evolutionary History in the RTE Clade of Non-LTR Retrotransposons. Mol. Biol. Evol. *18*, 1849–1863.

Appendix

List of LINEs used to generate phylogenetic trees.

Clade	LINE	Species
CRE	SLACS	*Trypanosoma brucei*
	CZAR	*Trypanosoma cruzi*
	CRE1	*Crithidia fasciculata*
	CRE2	*Crithidia fasciculata*
	GilM	*Giardia intestinalis*
R4	R4A1	*Ascaris lumbricoides*
R1	R1Dm	*Drosophila melanogaster*
	R1	*Bradysia coprophila*
	RT1	*Anopheles gambiae*
	RT2	*Anopheles gambiae*
	R1Bm	*Bombyx mori*
	TRAS1	*Bombyx mori*
	SART1	*Bombyx mori*
LOA	LOA	*Drosophila silvestris*
	BAGGINS1	*Drosophila silvestris*
	Bilbo	*Drosophila subobscura*
	Lian-Aa1	*Aedes aegypti*
Tad1	Tad1	*Neurspora crassa*
	MgL	*Magnaporthe grisea*
	CgT1	*Glomerella cingulata*
Jockey	BMC1	*Bombyx mori*
	Doc	*Drosophila melanogaster*
	amy	*Bombyx mori*
	Juan-A	*Aedes aegypti*
	Juan-C	*Culex pipiens*
	NLR1Cth	*Chironomus thummi*
	Doc6	*Drosophila melanogaster*
	G5	*Drosophila melanogaster*
	G5A	*Drosophila melanogaster*
	Jockey	*Drosophila melanogaster*
	Fw	*Drosophila melanogaster*
	Fw2	*Drosophila melanogaster*
	G4	*Drosophila melanogaster*
	Helena	*Drosophila melanogaster*
	BS	*Drosophila melanogaster*
	BS3	*Drosophila melanogaster*
	TART-B1	*Drosophila melanogaster*
I	I-1_DR	*Danio rerio*
	IVK	*Drosophila melanogaster*
	I	*Drosophila melanogaster*
	ingi	*Trypanosoma brucei*
	L1Tc	*Trypanosoma cruzi*
CR1	Sam3	*Caenorhabditis elegans*
	Sam6	*Caenorhabditis elegans*
	LfR1	*Lepidosiren paradoxa*
	Sam1	*Caenorhabditis elegans*
	Q	*Anopheles gambiae*
	DMCR1A	*Drosophila melanogaster*
	CR1-3_AG	*Anopheles gambiae*
	CR1-5_AG	*Anopheles gambiae*
	T1	*Anopheles gambiae*
	CR1-2_AG	*Anopheles gambiae*
	CR1-4_AG	*Anopheles gambiae*
	CR1	*Gallus gallus*

Clade	LINE	Species
	PsCR1	*Platemys spixii*
	L3	*Homo sapiens*
	SR1	*Schistosoma mansoni*
L2	CR1-2_DR	*Danio rerio*
	UnaL2	*Anguilla japonica*
	L2A	*Eutheria*
	Maui	*Takifugu rubripes*
	CR1-1_DR	*Danio rerio*
	CR1-4_DR	*Danio rerio*
	CR1-3_DR	*Danio rerio*
	CR1-1_AG	*Anopheles gambiae*
	CR1-1a_AG	*Anopheles gambiae*
R2	R2Bm	*Bombyx mori*
	R2Nv	*Nasonia vitripennis*
	R2Fa	*Forficula auricularia*
	R2Dm	*Drosophila melanogaster*
	R2Ps	*Porcellio scaber*
	R2Am	*Anurida maritima*
	R2Lp	*Limulus polyphemus*
NeSL-1	NeSL-1	*Caenorhabditis elegans*
RTE	BovB	*Bos taurus*
	BovB_VA	*Vipera ammodytes*
	RTE-1	*Caenorhabditis elegans*
	RTE-2	*Caenorhabditis elegans*
	RTE-1_AG	*Anopheles gambiae*
	Rex3	*Xiphophorus maculatus*
L1	L1Hs	*Homo sapiens*
	L1Md	*Mus musculus*
	L1-1_DR	*Danio rerio*
	L1-10_DR	*Danio rerio*
	L1-6_DR	*Danio rerio*
	L1-8_DR	*Danio rerio*
	Sw1Cm	*Cyprinodon macularius*
	Sw1O1	*Oryzias latipes*
	L1-3_DR	*Danio rerio*
	L1-5_DR	*Danio rerio*
	L1-2_DR	*Danio rerio*
	L1-4_DR	*Danio rerio*
	ATLINE1_5	*Arabidopsis thaliana*
	Tal1-1	*Arabidopsis thaliana*
	ATLINE1_6	*Arabidopsis thaliana*
	Cin4	*Zea mays*
	Tx1	*Xenopus laevis*
	Zepp	*Chlorella vulgaris*

Clade: Groups of LINE **LINE:** Name of LINE elements **Species:** Scientific name of the organisms.