De-novo assembly of the Octopus Transcriptome integrating custom and public sequencing data

Petrosino G(1), Zarrella I(1), Ponte G(1), Basu S(1), Calogero RA(2), Fiorito G(1), Sanges R(1)

(1) Laboratory of Animal Physiology and Evolution, Stazione Zoologica Anton Dohrn, Naples, Italy
(2) Bioinformatics and Genomics Unit, MBC Centro di Biotecnologie Molecolari, University of Turin, Italy

Contact: gius.petrosino@gmail.com

Motivation

Cuttlefishes, squids and octopuses belong to Cephalopoda, a class of the Phylum Mollusca. During the course of evolution, Cephalopods changed dramatically their body plan from the molluscan ancestor, and differentiated their way of life. Among cephalopods, the common octopus, Octopus vulgaris, shows sophisticated motor, sensory and cognitive capabilities, including excellent vision, highly efficient flexibility of their arms, and learning capabilities; flexibility in behaviour necessary for competition with fishes [1]. Furthermore, morphological data account for a complex nervous system containing as many neurons as in the dog brain [2].

For these reasons, the octopus is considered one of the most intelligent invertebrates, and a key animal for neurosciences. However, the use of the octopus as model organism is largely affected by the lack of molecular tools such as a sequenced genome and transcriptomes. Therefore, the initial aim of this study is to move Octopus biology from a pre-transcriptomic to a post-transcriptomic era generating the reference transcriptome.

Methods

The O. vulgaris transcriptome was generated integrating two published RNA-seq data [3,4], gathered through Roche/454 and Illumina GAIIx sequencing technologies, and an EST library produced at Stazione Zoologica in Naples (SZN). MIRA [5] was used to perform an hybrid de-novo assemblies integrating together reads and sequences originated from the indicated different technologies. Non-redundant transcripts were then obtained using CAP3 with default parameters [6] and filtering out transcripts shorter than 500 bp. Annotations were collected using a pipeline able to perform blastx, rpsblast, blastn against Uniref, CDD, and RFAM databases, respectively. In addition, the pipeline is able to obtain the coding potential for each sequence using a heuristic score based on the Portrait software [7]. Mollusc transcriptomes assembled by Smith et al. were downloaded from the Dryad Digital Repository [8]. Mouse and human transcriptomes were downloaded from the NCBI UniGene database [9]. Repetitive elements composition of the transcriptomes was evaluated by using RepeatMasker software and Repbase database [10,11] searching for bilateral repeats. Custom perl and R scripts were prepared and used to evaluate the presence of specific repeat classes and families in both coding and non-coding transcripts of O. vulgaris, M. musculus and H. sapiens and the other used transcriptomes. Basically, we counted, in the RepeatMasker output, the repeats that were present, at least once, in each transcript for each transcriptome.

Results

The strategy for the hybrid assembly allowed us to predict about 45,000 different transcripts. We annotated 29,167 (64.8%) transcripts and predicted 5,115 (11.4%) lncRNAs from the Octopus transcriptome. The Gene Ontology (GO) analysis highlighted that an interesting subset of transcripts belongs to the RNA-dependent DNA replication class, indicating a crucial biological process related to retrotransposons. Moreover, Octopus transcriptome contains a high proportion of domains well known to be highly frequent in transcriptomes sequenced until now, such as Zn-finger; however, we found an unusual proportion of reverse transcriptase and transposase domains that overlap about 500 transcripts. These results suggest that repetitive elements are frequently transcribed and might be highly active into the Octopus transcriptome.

Based on these results, we performed an interspecies repetitive elements transcriptome analysis.

We found that the Octopus transcriptome contains a more high percentage of bases in repeats than those of all the other mollusc species considered. Finally, we found that SINEs and LTRs in the Octopus transcriptome are significantly more abundant in lncRNAs than in protein coding transcripts. Recently, a correlation between lncRNAs and repetitive elements has been reported in the human and mouse transcriptome [12] and we believe that a similar phenomenon also occurs in Octopus vulgaris.

References

1. Hochner B et al.; The octopus: a model for a comparative analysis of the evolution of learning and memory mechanisms. Biol Bull. 2006 Jun;210(3):308-317.
2. Hochner B; An embodied view of octopus neurobiology. Curr Biol. 2012 Oct 23;22(20).
3. Kocot KM et al.; Phylogenomics reveals deep molluscan relationships. Nature. 2011 Sep 4;477(7365).
4. Smith SA et al.; Resolving the evolutionary relationships of molluscs with phylogenomics tools. Nature. 2011 Oct 26;480(7377).
5. Huang X et al.; CAP3: A DNA sequence assembly program. Genome Res. 1999 Sep;9(9):868-877.
6. Chevreux B. et al.; Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res. 2004 Jun;14(6):1147-1159.
7. Arrial RT et al.; Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus Paracoccidioides brasiliensis. BMC Bioinformatics. 2009 Aug 4;10:239.
8. Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.24cb8.
9. UniGene: http://www.ncbi.nlm.nih.gov/unigene; Mm build 193 and Hs build 235.
10. Saha S et al.; Empirical comparison of ab initio repeat finding programs. Nucleic Acids Res. 2008 Apr;36(7):2284-2294.
11. Jurka J et al.; (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110(1-4):462-467.
12. Kelley D et al.; Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol. 2012 Nov 26;13(11).