Discovery of conserved long non-coding RNAs in vertebrates

S. Basu✉, **R. Sanges**
Bioinformatics - Animal Physiology and Evolution, Stazione Zoologica Anton Dohrn, Napoli, Italy

Motivations

Long non-coding RNAs (lncRNA) have been reported as a major class of novel transcripts related to organism development and early neural expression pattern [1-4]. They are reported to be expressed in large numbers in the mammalian transcriptomes [5,6] and recently reported to be expressed in the teleost fishes [7,8]. Computational identification and characterization of lncRNA from public sequence resources have been performed by different groups [9-11]. The focus of attention has been on the mammalian genomes starting by the assumption that they are not well conserved in term of sequence. However, systematic studies measuring their levels of conservation among vertebrates are lacking. Hence we want to computationally evaluate the existence of vertebrate conserved lncRNAs through systematic conservation analyses of both sequence as well as genomic architecture.

Methods

Mouse lncRNAs reported in an earlier study [2] and predicted by the EnsEMBL pipeline were considered as a reference dataset. Homology search of the lncRNAs against the zebrafish conserved phastcons elements was performed with the BLAST program. The phastcons elements are regions of conservation in the zebrafish genome with human, mouse, western clawed frog and two teleost fishes, tetraodon and stickleback. The lack of selection pressure in lncRNAs as compared to the protein-coding genes required a calibration of BLAST parameters to define a cut-off score indicative of significant conservation. Using ROC analyses we calculated the best BLAST parameters able to select regions of lnRNA conserved in vertebrates. The predicted conserved candidates were also evaluated in terms of their RNA secondary structure using the RNAfold software. Gene ontology and expression pattern enrichment of flanking protein-coding genes was performed with DAVID software.

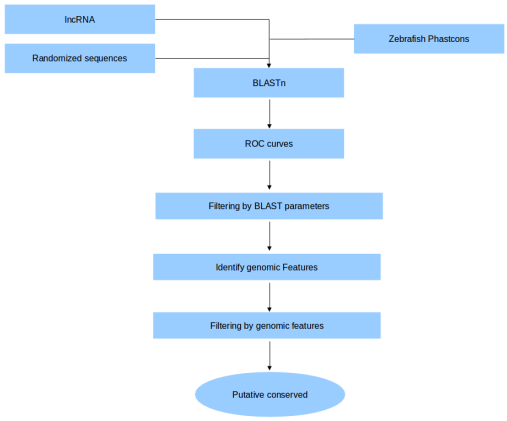

Figure 1. A schematic representation of the pipeline followed to identify putative conserved mouse long non-coding RNAs in the zebrafish phastcons elements.

Results

Our results show that the usage of the alignment length as cut-off is sufficient to distinguish the conservation of mouse lncRNAs in zebrafish as compared to conservation of random genomic regions. The RNA secondary structure prediction was not able to define any threshold for conservation. From an initial dataset of ~2,800 lncRNAs we could predict that 235 are conserved using the defined cut-off on the alignment length. Gene ontology enrichment analyses, related to the protein-coding genes proximal to the region of conservation in mouse and zebrafish, highlighted corresponding GO classes such as regulation of transcription and central nervous system development. The proximal coding genes exhibited a similar enrichment for their tissue of expression where brain was highly enriched in both mouse as well as zebrafish. Two interesting candidate regions of conservation were chosen for future experimental validation based upon the presence of ESTs overlap and the function of the proximal proteins (in this case the interest being development and functioning of the nervous system). The analysis is poised as an initial pipeline to select interesting candidate lncRNAs conserved among vertebrates.

References

1. Guttman M, Amit I, Garber M, French C, Lin MF, et al. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458: 223-227. doi:10.1038/nature07672.
2. Ponjavic J, Oliver PL, Lunter G, Ponting CP (2009) Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. PLoS Genet 5: e1000617. doi:10.1371/journal.pgen.1000617.
3. Qureshi IA, Mattick JS, Mehler MF (2010) Long non-coding RNAs in nervous system function and disease. Brain Res 1338: 20-35. doi:10.1016/j.brainres.2010.03.110.
4. Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS (2008) Specific expression of long noncoding RNAs in the mouse brain. Proc Natl Acad Sci USA 105: 716-721. doi:10.1073/pnas.0706729105.
5. Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T, et al. (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat Genet 36: 40-45. doi:10.1038/ng1285.
6. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, et al. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420: 563-573. doi:10.1038/nature01266.
7. Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, et al. (2011) Systematic identification of long non-coding RNAs expressed during zebrafish embryogenesis. Genome Research. Available:http://genome.cshlp.org/content/early/2011/11/22/gr.133009.111.abstract. Accessed 23 November 2011.
8. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, et al. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes '& Development 25: 1915 -1927. doi:10.1101/gad.17446611.
9. Khachane AN, Harrison PM (2010) Mining mammalian transcript data for functional long non-coding RNAs. PLoS ONE 5: e10316. doi:10.1371/journal.pone.0010316.
10. Jia H, Osak M, Bogu GK, Stanton LW, Johnson R, et al. (2010) Genome-wide computational identification and manual annotation of human long noncoding RNA genes. RNA 16: 1478-1487. doi:10.1261/rna.1951310.
11. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, et al. (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotech 28: 503-510. doi:10.1038/nbt.1633.