Kraków, 9th July 2021

prof. dr hab. Wiesław Babik
Institute of Environmental Sciences
Jagiellonian University
Gronostajowa 7
30-387 Kraków
email: wieslaw.babik@uj.edu.pl
phone: 12 664 51 71

Evaluation of PhD thesis of Oleksii Bryzghalov, MSc

"Application of the transcriptomic data and evolutionary conservation in search and characterization of long non-coding RNAs in animals"

Long non-coding RNAs (lncRNAs) are ubiquitous in Eukaryotes. Some of the lncRNAs play important and well-understood functions. In the vast majority of cases, however, we know very little beyond the simple fact that a given lncRNA exists in certain tissues at certain developmental stages. An important challenge in studies of lncRNA is poor conservation of their primary sequence between species that makes the establishment of orthology using standard methods is extremely difficult or even impossible. Therefore, various characteristics of lncRNAs need to be considered to understand their evolutionary conservation, and comprehensive databases of lncRNA sequences from various taxa are essential for a better understanding of lncRNA biology. A large scale characterization of lncRNAs in humans and other primates is the topic of Oleksii Bryzghalov's PhD dissertation. The thesis, supervised by professor Michał Szcześniak, was prepared at the Institue of Human Biology and Evolution, Faculty of Biology, Adam Mickiewicz University in Poznań.

The dissertation consists of a short summary, a general introduction and three research articles published in respectable journals listed in Journal Citation Reports. The candidate is the first author of all three papers and he estimated his contribution to articles 1 and 2 as 40%, and to article 3 as 70%. In each case Mr Bryzghalov was involved in the preparation of the manuscript, however, I did not find information about his leading role in the writing of any of the papers or his contribution to the study design. The dissertation contains also declarations of co-authors describing their contribution to the published articles. I have only one comment regarding the formal side of the dissertation: the summary is extremely concise, the content of the thesis is described in a single short paragraph and article 1 is not mentioned in the summary at all. The summary makes an impression of compiled in haste and does not really allow a reader to obtain more than a vague idea about the content of the thesis.

The general introduction contains some background information about lncRNAs that basically repeats the information contained in introductions to the papers, and a brief description of the three articles that form the core of the thesis. Although the introduction is short, it provides a sufficient overview of the topic and a useful context for the papers that follow. In the introduction, Mr Bryzghalov writes that lncRNAs are found across all major groups of Eukaryotes, and then mentions fungi, plants and animals. I would like to ask what is known about lncRNAs in unicellular Eukaryotes from major supergroups (Stramenophila, Alveolata etc.)?

Paper 1, published in *Acta Biochimica Polonica*, investigated retroposition as a possible source of long non-coding RNAs. RNAs that transcribed in antisense from retrocopies may regulate the function of their parental genes, possibly by base-pairing interactions. In humans, the authors identified 35 pairs of lncRNA-retrocopy overlaps. Using *ab initio* assemblies of multiple chimpanzee RNAseq libraries they identified 23 such pairs in this species. The authors found no conservation of antisense transcripts of retrocopies even at close evolutionary distances, as demonstrated by the lack of overlap between the sets of lncRNA-retrocopy from humans and chimpanzees. Interestingly, an instance of apparently independent origin of lncRNA transcribed in antisense from the same retrogene was found in mouse and human. Analysing 153 RNAseq libraries from the Encode project the authors found that 27 of 35 human lncRNA were coexpressed with their parental genes, and three pairs showed significant expression correlation across libraries (two – positive, one – negative). I did not find information about correction of *p*-values for multiple tests, which would be needed in this case as 27 tests were performed. Still, the results of the correlation analysis, although not conclusive, are suggestive. A substantial part of the paper provides a detailed discussion of three out of 10 cases where the analysis of RNA:RNA duplexes suggested regulatory effect of lncRNA on the parental gene – these three cases were those in which significant correlation in expression was detected. Overall, paper 1 provides an interesting, thorough and largely convincing analysis, indicating that lncRNAs derived from retrocopies may play a role in gene regulation in humans. The paper is well written, except for some redundancy between paragraphs 1 and 2 of Results and Discussion. Article 1 was published in 2016, and I would like to ask the candidate about the progress in the understanding of retroposition as a source of antisense lncRNA over the last five years.

Article 2, published in *Nucleic Acid Research Database Issue* describes SyntDB – a database of orthologous lncRNAs in humans and 11 non-human primates. The paper clearly outlines the problem and describes state of the art, emphasizing that information about lncRNAs, in particular about their orthologous relationships between species is scarce. The authors describe their *ab initio* transcriptome assembly pipeline as well as the procedure they used for the identification of lncRNAs in transcriptomes. The limitations of the existing databases are discussed and both features and content of SyntDB are described. Orthologues between species are identified with the help of whole genome alignments and four levels of lncRNA conservation are distinguished, based on the phylogenetic depth of orthology retention: i) human-specific, ii) great-ape specific, iii) conserved, and iv) ultraconserved lncRNA. Three types of orthology evidence are distinguished: a) exon, indicating that exon-intron structure is retained, b) locus, indicating lncRNA sequence similarity without conservation of splicing pattern, c) syntenic, indicating that lncRNAs in two sepcies are transcribed from the corresposding genomic regions without evidence for either sequence or splicing pattern conservation. SyntDB reports also detailed expression data for human lncRNAs and their orthologues in different species. The paper presents the details of database implementation and illustrates its typical uses. The web database interface is visually appealing, useful and has a modern touch. An illustrative example helps in familiarizing the user with the database. The thing I have been missing is the lack of database statistics in a tabular form – all the information is there, but accessible via point-and-click, which is not convenient if one needs just a quick overview. Overall SyntDB is an important resource that should be of considerable importance in comparative research that has the potential of illuminating lncRNA biology. The contrast between over 78,000 lncRNAs identified in humans and 2,054 – 18,226 identified in other primate species illustrates how little we know about lncRNAs even in intensively studied non-human taxa. The paper is carefully written and edited, the only comment regarding the presentation I have is that on p. D242, in addition to *p*-values, it would be useful to see the average values of expression level.

In relation to Article 2, I wanted to ask Mr Bryzghalov what is known about the relationship between lncRNA conservation and expression level and whether such a relationship could bias the view of lncRNA landscape as presented in SyntDB, as a consequence of differences in the amount of RNAseq data available for various species. It looks like SyntDB has not been updated for some time. Do you plan to update it regularly, having at your disposal the pipeline described in publication 3? I would also like to know the opinion of the candidate on how natural selection affects the evolution of lncRNAs that do not show primary sequence conservation. Do you think this is just the matter of finding appropriate signatures of conservation that go beyond primary sequence, or do we need some more fundamental change in our approach to studying the evolution of lncRNAs?

Article 3, published in *BMC Bioinformatics*, presents lncEvo, an integrated pipeline for the identification and evolutionary conservation analysis of lncRNAs that is an extension of the methodology used to create SyntDB. The pipeline consists of three parts: 1) *ab initio* transcriptome assembly from RNAseq data, 2) prediction of lncRNAs, 3) analysis of conservation in a pairwise manner. The pipeline is containerized using Docker and can be run on the Amazon Web Services (AWS) cloud. A post-pipeline data analysis can also be performed on AWS in an interactive query service Athena that allows using SQL queries without the need of setting up the actual database. To illustrate the practical use of lncEvo the authors use RNAseq data from several species (human, chimp, horse, dog, mouse) to identify lncRNAs. In all these species the majority of identified lncRNAs were novel intergenic transcripts, which illustrates the generally poor quality of the existing lncRNA annotations. The paper is written very clearly, and describes the conceptual underpinning and implementation of the pipeline in sufficient, but not excessive, detail. The choices the authors made with respect to the development environment and software used are well explained. The paper contains an informative analysis of the computational performance. It was not entirely clear for me whether data in Table 1 refer to the entire pipeline (as the text would suggest) or without the conservation part (as table legend would suggest). Fig. 7c is not easy to interpret because of the large number of colours.

Regarding Article 3, I would like to ask about syntenic homologues (syntologs). Do you think that syntologs arise mostly via parallel evolution or rather that they are typically of single origin but diverge in both sequence and exon structure so much, that they retain only the genomic position? How would you test such alternative hypotheses? The authors present lncEvo as a pipeline dedicated for mammals and currently support only species with genomes available in ENSEMBL genomes. I am curious whether this pipeline can be applied to other taxa with available genome assembly and RNAseq data, and if so, would this require any substantial modifications of the pipeline?

As can be seen from the assessment above, I have not found serious flaws or errors in the papers forming the core of the thesis. Most of the points raised above refer either to the formal aspects of the dissertation or are just comments and questions. This is perhaps not surprising, as the published papers already went through a rigorous editorial and peer-review process. The overall quality of the thesis is high and it has been a pleasure to read and evaluate it. The candidate demonstrated an excellent command of an impressive array of bioinformatics tools and techniques, and an ability to combine these skills to address an important and novel research question in comparative genomics. I suppose that the work on this thesis was facilitated by a stimulating intellectual atmosphere and expert advice provided by professor Makałowska's genomics group.

To sum up, the evaluated dissertation reports the results of original research that addresses a well-defined scientific problem, demonstrates the candidate's general theoretical knowledge and his capability of conducting independent research. Therefore I conclude that the presented dissertation fulfils the requirements of the Polish Act on Academic Degrees and Academic Title and Degrees and Title in Art dated 14 March 2003 with subsequent changes (Dz. U. 2017, pos. 1789) and I recommend the Scientific Discipline Board for Biological Sciences at Adam Mickiewicz University in Poznań to admit Oleksii Bryzghalov to the further stages of the doctoral procedure.

Wiesław Babik

Elektronicznie
podpisany przez
Wiesław Babik
Data: 2021.07.09
17:16:05 +02'00'