UNIWERSYTET IM. ADAMA MICKIEWICZA W POZNANIU
Wydział Biologii

dr hab. Michał Szcześniak, prof. UAM
Adam Mickiewicz University in Poznan
Faculty of Biology
miszcz@amu.edu.pl

Poznań, March 27th, 2025

Review of Monika Kwiatkowska's PhD thesis entitled

„Genomic characterization of long-noncoding RNAs in the zebrafish genome"

The thesis submitted for review was supervised by dr. hab. Barbara Uszczyńska-Ratajczak, prof. IBCH PAS, with dr. Silvia Carbonell Sala serving as a co-supervisor. The research was carried out at the Department of Computational Biology of Non-coding RNA, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznań, and was funded through two National Science Centre grants awarded to the Supervisor.

The thesis follows a classic structure, divided into eight numbered chapters, including: Introduction, Objectives, Materials and Methods, Results, and Discussion. In addition, the thesis includes unnumbered sections such as the Abstract, Table of Contents, Supplementary Materials, and Bibliography. Notably, the PhD candidate provided a GitHub link granting access to the thesis in electronic format (PDF), along with the code used for data analysis: https://github.com/cobRNA/MKwiatkowska_PhD_Thesis.

1. Evaluation of the thesis

1.1 Background

Zebrafish (*Danio rerio*) is a widely used and highly valuable model organism for molecular biology and developmental studies. Many discoveries made in zebrafish are later translated to humans through the study of orthologous genes. This species offers numerous advantages as a model system, including the fact that approximately 70% of human genes have at least one zebrafish counterpart. Additionally, zebrafish are relatively inexpensive and easy to maintain, and they exhibit a high fertility rate, especially when compared to traditional mammalian models such as mice and rats.

Zebrafish has repeatedly proven to be a powerful model for studying vertebrate development, gene regulation, and molecular mechanisms underlying both physiological and pathological conditions. However, the full potential of zebrafish as a model organism is still limited by incomplete genome annotations, especially concerning long non-coding RNAs (lncRNAs), whose annotation lags significantly behind that of protein-coding genes.

lncRNAs represent crucial components of the transcriptome, playing diverse roles in gene regulation, chromatin remodeling, embryonic development, cellular differentiation, and disease mechanisms such as cancer progression and neurodegenerative disorders. Given their importance, it is justified to develop more comprehensive and accurate annotations of zebrafish lncRNAs. Enhanced annotation will not only deepen our understanding of lncRNA function in zebrafish but also strengthen the translational relevance of zebrafish studies to human biology.

Keeping this in mind, the PhD candidate set up to address the following challenges in zebrafish lncRNAs annotations:

I. Identify orthologout lncRNAs with human and mouse, using a synteny-based approach

II. Eliminate annotation bias by emphasis on adult zebrafish organs, rather than developmental stages, which are relatively better characterized

III. Improve lncRNA annotation completeness in terms of determining the 5' and 3' ends of transcripts

IV. Enhance representation of lowly expressed lncRNAs

Fulfillment of those priorities was expected by the PhD candidate to strengthen zebrafish as a model organism, especially in the context of functional studies on lncRNAs.

1.2 Evaluation of the thesis

1.2.1 Achieved goals

One of the major strengths of the thesis is the optimization and application of the CapTrap-seq library preparation protocol for zebrafish, tailored to enrich full-length, 5' capped transcripts. This method was systematically benchmarked against another approach, Template Switching Oligo (TSO), and demonstrated clear superiority in several aspect. The addition of bead-based size selection (SS500) further improved transcript model lengths and polyadenylated read recovery. The candidate integrated this optimized method with the LyRic annotation pipeline, leading to the generation of an extended zebrafish transcriptome, including over 14,000 novel genes, of which more than 12,000 were classified as lncRNAs. This resulted in a fourfold

expansion of the known zebrafish lncRNA catalog and significantly increased the average number of isoforms per gene.

The study focused on transcriptomically complex adult tissues, counteracting the bias in zebrafish transcriptome annotations, which are heavily skewed toward embryonic stages, which aligned with the second goal. The thesis also tackled the issue of lowly expressed lncRNAs, which often limits their detection and annotation. The combination of CapTrap-seq with Capture Long-read Sequencing (CLS) proved effective in selectively enriching lowly expressed lncRNAs, achieving enrichment levels up to 100-fold for targeted loci.

In terms of functional analysis, the thesis employed a synteny-based approach (ConnectOR) to identify 49 positionally conserved lncRNAs between zebrafish and humans, addressing the goal of identifying orthologous lncRNAs. The candidate further examined the expression and potential regulatory functions of these lncRNAs, with a special focus on small RNA host genes, such as *snhg1*. Through RNAscope experiments and expression data analysis, the study revealed conserved and tissue-specific expression patterns, particularly in the brain and eye.

In conclusion, this thesis clearly achieves its stated goals and represents a significant contribution to the field of zebrafish transcriptomics, particularly in the area of lncRNAs annotation. The central aim of the work - to improve the quality and completeness of zebrafish lncRNA annotation - has been addressed with methodological rigor.

1.2.2 Critical points

Nonetheless, there are some aspects that are not clear enough or could possibly be improved. Below is a couple of issues I would like to raise.

I. Please discuss both the strengths and weaknesses of the ConnectOR synteny-based approach. Why were only 49 human counterparts identified - what factors limited this number? Can these lncRNAs truly be considered orthologs (some call them syntologs, for example)? Why were the majority of zebrafish lncRNAs mapped to human protein-coding genes rather than annotated lncRNAs?

II. While the developed protocols reduce some known biases, they might also introduce new ones. For example, the approach focuses on polyadenylated transcripts, excluding many biologically important non-polyadenylated lncRNAs. These non-polyadenylated lncRNAs are often unspliced, nuclear-retained, but may actually have important regulatory functions in the nucleus. Some non-polyadenylated lncRNAs are even exported to the cytoplasm via alternative mechanisms and have described functional roles (e.g., PTEN pseudogene transcripts in human). Please discuss the consequences of

focusing exclusively on polyadenylated RNA species and how this may limit the biological conclusions.

III. Additionally, antisense lncRNAs were excluded from capture probe design, yet they represent a significant fraction of the actual lncRNA population (~70% of human protein-coding genes have antisense partners). lncRNAs overlapping protein-coding genes in sense or in their close proximity were also omitted. Nevertheless, such transcripts appear later in the analysis (e.g., Figure 4.31), creating some inconsistency. The candidate should clearly explain the rationale and implications of these filtering steps and discuss possible strategies to overcome this limitation in future studies.

IV. Please clarify how capture probe target regions were selected across annotation sets: was there redundancy in the captured loci (i.e., were any loci targeted through multiple annotations)? A Venn diagram or UpSet plot illustrating overlap among annotation sets would significantly enhance this section.

V. The thesis lacks detailed information on software versions and parameter settings used throughout the computational pipeline.

VI. The data visualization is of high quality and improves the readability of the work. However, some plots lack important contextual information: lacking units (e.g., Fig. 4.36 – are values expressed in TPM?). There is also lack of basic statistics such as median, standard deviation, or statistical significance; these should be clearly indicated in the plots and/or figure legends, or in the main text.

VII. There exist many unspliced, monoexonic, and intergenic (or intronic) transcripts emerging from the analysis: are these considered likely artifacts, or are they genuine lncRNA candidates, even if lacking complete 3' and 5' ends? How does the candidate interpret their biological relevance, especially in tissues like testis?

VIII. The thesis would benefit from the inclusion of experimental validation (e.g., RT-PCR or RACE) for a subset of novel or extended transcripts. Such validation would strengthen confidence in the newly annotated transcript models and support the proposed methodology.

IX. More background should be provided on what is already known about *snhg1* and its associated snoRNAs in other organisms. How does this information align with the candidate's findings in zebrafish?

X. The Discussion section reads more like a summary of findings than a critical interpretation. It would benefit from deeper insights from existing literature, especially comparative analysis with zebrafish annotation efforts by others and recent methodologies for lncRNA detection and annotation in other model organisms.

2. Conclusions

According to the current regulations, including *art. 187 ust. 3 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce*, the review of PhD thesis is expected to answer the following three questions:

1. **Does the dissertation present and document general knowledge of its author in the research area and discipline?**
   The presented dissertation in the discipline of "Biological Sciences" definitely shows Ms Monika Kwiatkowska's broad knowledge and deep understanding of the studied subject. First of all, there is an introductory part at the beginning of the dissertation, including an explanation of the biology, methodology and technologies related to the studied phenomena. Also, the quality of writing in other parts and the performed studies themselves well document her knowledge in the area.
2. **Evaluation whether the dissertations shows its author's ability to conduct scientific research on his own**
   On the basis of the evaluated thesis, with a critical role of the PhD candidate in performing all the tasks, I have no doubt that she aquired skills that predispose her for future research on her own.
3. **Evaluation whether the presented dissertation constitutes an original solution to a scientific problem**
   The PhD candidate's findings and methodological advancements presented in the thesis and briefly characterized above, clearly demonstrate solutions to scientific problems.

3. Final remarks

I find the presented thesis very valuable and the PhD candidate's results of interest to the broad scientific community and my minor doubts and remarks raised above do not nullify my overall good impression. I therefore conclude that the dissertation meets all the requirements for the doctoral thesis and strongly recommend proceeding to the following steps of the doctoral defense.

A Polish version:

Przedstawiona do recenzji rozprawa doktorska spełnia warunki określone w Ustawie z dnia 20 lipca 2018 roku prawo o szkolnictwie wyższym i nauce (Dz.U. z 2018 r. poz. 1668 ze zm.) oraz w Sposobie postępowania w sprawie nadania stopnia doktora w Instytucie Chemii Bioorganicznej PAN w Poznaniu (uchwała Rady Naukowej ICHB PAN nr 28/2024/Internet z dnia 20 marca 2024 r.) i wnioskuję do Rady Naukowej Instytutu Chemii Bioorganicznej PAN o dopuszczenie pani mgr inż. Moniki Kwiatkowskiej do dalszych etapów postępowania o nadanie stopnia doktora.

Michał Szreniawa

ul. Uniwersytetu Poznańskiego 6, Collegium Biologicum, 61-614 Poznań
NIP 777 00 06 350, REGON 000001293
tel. 61 829 5945
www.biologia.amu.edu.pl