Dr. Martin Těšický

Telephone: +49 (0)89 2180-2680
Martin.Tesicky@lmu.de

Paleogenomics group
Faculty of Veterinary Medicine,
Institute of Paleonatomy, Domestification Research and History of Veterinary Medicine
Ludwig Maxmilian University of Munich

Kaulbachstr. 37 III
80539 Munich
Germany

https://www.animal-palaeogenomics.com/

In Munich
January 21st 2024

Review on Master thesis of Lukáš Cakl "Detection of positive selection in reproductive genes of songbirds"

Lukáš Cakl's master's thesis focuses on very actual topic of reproductive proteins in birds. Reproductive genes are a very diverse group of genes that are more or less linked with reproduction, but their molecular evolution in birds is poorly understood. With the advent of massive genome sequencing and robust bioinformatics pipelines, it is now possible to systematically screen avian genomes and uncover the signs of adaptive evolution (not only) in reproductive genes. The main aim of this thesis is to find such candidates in passerines using selection scans and then to characterise them functionally using GO term enrichment analysis.

The thesis has a standard structure. In the introduction, the author briefly outlines the biological concepts - sexual selection, evolution of sperm morphology in passerine birds and cryptic female choice - and then the methodological concepts. What I missed here, however, is the part devoted to reproductive genes – what are they? In short, what is known in other groups (e.g. mammals) and why this research is important. In birds, there are a few recent studies, e.g. Rowe et al. *Mol.Biol.Evol*. 2020 or Rowe et al. *J. Prot.* 2019, that would be worth including. Also, the chapter devoted to cryptic female choice and composition to female fluid is very short. Although chapter 1.5 is relevant, it is a relatively basic, textbook-based chapter with few cited sources.

The aims of the thesis are clearly formulated. The methods are generally described in detail and the criteria for input data are well justified. However, it is not entirely clear to the general reader why two nightingale species were included and whether the selection of species reflects the variability in the intensity of sexual selection/sperm morphology at all. Could the author explain this for both the genomic and proteomic datasets? The bioinformatic analysis and methods used are solid, complex, and mostly up to date. It is obvious that Lukáš has extensive programming skills and has mastered a wide range of bioinformatics tools which are not easy to run.

The results of thesis are novel and are well presented in several carefully elaborated figures and tables. In total, of the almost 1500 reproductive genes, only 6 genes were identified in the female reproductive fluid and 22 in the sperm cells. These included mainly genes related to sperm morphology and metabolism, which fits with the enormous phenotypic variation in passerine sperms and are relevant for the evolution of GRC chromosome. The discussion discusses some potentially interesting candidate genes and puts them in the

context of human/mouse disease/ mutation studies. I also appreciate that the methodological limitations of the study are addressed. Nevertheless, I would welcome a comparison of the molecular evolution of reproductive genes with previous evolutionary studies, e.g. including those on mammals and insects if possible.

Formal and linguistic level
The thesis is formally well prepared and clearly structured. Even if it is sometimes a matter of taste, I would summarise some sub-chapters into single chapters for a better flow (e.g. 3.1).
I appreciate that the thesis is written in relatively good English, but some sentences are relatively complex and need to be reworded or made more precise. The thesis also contains a few typos, but these do not affect the overall understanding.

Overall, this is a well-done, novel pilot study that opens the avenue for further hypothesis-driven research and provides a set of candidate genes for further testing. I enjoyed the reading and recommended it for defence.

I am looking forward to discussing the findings with the candidate in more detail during the defence.

Yours sincerely
Martin Těšický

I have several questions and comments:

1) As mentioned above, I was missing a bit of an explanation of what reproductive genes are. For example, are they all genes that are expressed in reproductive cells/tissues, including housekeeping genes? Or are there genes that are only involved in sexual selection? I would welcome the author's opinion on the categorisation of reproductive genes.

2) On page 17, the author claims: “While more involved methods of orthologue detection, such as Orthofinder [Emms and Kelly, 2019], exist their computational complexity essentially prevents their usage on the genome-wide scale.” I cannot agree with this statement because Orthofinder and other newer methods, such as TOGA have been shown to be relatively fast and have higher accuracy. For example, the standard version of OrthoFinder ran in 192 s. on the 4 fungal species (approx. 10,000 protein-coding genes per genome) and 1.8 days for the 256 species datasets (Davies et. al. *Genom. Biology* 2019). While the

reciprocal blast may work well for relatively closely related species, this may deteriorate with increasing divergence time between species.
I encourage the author to consider these methods as well. These are better suited for orthogroup assignment or finding duplication events in more deeply divergent species. According to Timetree.org, the divergence time between species used in this study is up to 25 MYA, which is a moderate level of divergence. Also, it was not entirely clear if the selection scans were always performed on orthogroups with only one gene copy per species or if orthogroups with multiple sequences from the same species were also included?

3) In the reasoning why the author only used only bio++ selection method and no other widely used dN/dS tests, he claims that "the reason for rejecting HYPHY was due to it being less conservative." Could the author elaborate on this? The HYPHY package contains several dN/dS tests that are usually quite fast (e.g. FUBAR, FEL, MEME, etc.) and different levels for posterior probabilities/p-values can be applied. Do the authors plan to compare the results of several selection methods in advance?

4) Working with non-model organisms in evolutionary analysis is always a challenge, especially when we lack well-annotated reference genomes of closely related species. Some authors used for the annotation of genes in non-model birds only identified orthologous genes between chicken/zebra finch and assign them with their human orthologous for GO term annotation. Here, the author rather used two predictive methods, Blast2GO and Interproscan. Could the author briefly describe how these methods work and evaluate their advantages and limitations compared to the first approach? Which taxon was used to perform the gene-over-representation analysis?

5) What is meant by the "coverage check" in the following sentence?

"From this 633 orthogroups were rejected by coverage checks resulting in the final number of analyzed orthogroups to be 12015, which covers approximately three quarters of the expected gene count and results in the distribution of orthogroups shown in figure."

6) In the discussion, the authors shed light on the function and significance of some candidate genes in sperm and seminal fluid. One approach is to examine whether these genes are associated with a disease phenotype, but the author could also check whether these genes were found under positive selection, e.g. in mammalian studies, if possible. What other functional categories of genes involved in reproduction could author hypothesis could to be under positive selection in birds despite not being identified here?

Other minor (not complete) comments/suggestions that may be useful for the preparation of the manuscript):

Aminoacid – better to use with space "amino acid"

p. 9 In: "Over sufficiently long time frames we can observe three types of loci, based on the prevailing type of substitution.".
Better to use "amino acid sites over loci"

p. 9, "In loci, where mutations lead to no change in fitness, we observe an equal amount of both types of substitutions and finally, in loci where non-synonymous substitutions form the majority, we can postulate that their effects must be positive."
This statement is inaccurate because even in proteins that are subject to positive selection, the majority of sites evolve under purifying selection and only a small fraction (a few %) with a specific function, e.g. ligand-binding sites, are subject to positive selection, e.g. Velová et al.Mol. Evol. Biol. 2018.

p. 12., Chapter 3.1., High-quality genomes from long-read sequencing platforms are now released in Vertebrate Genome Projects and are also available for multiple passerines (https://vertebrategenomesproject.org/).

p. 13, Figure 3.1. Phylogenetic species tree sounds like being manually compiled: "
It was then extended manually to include the Blue tit, which belongs to the same Paridae family as the Great tit [Johansson et al., 2013]."
It can be better extracted automatically from Bird tree https://birdtree.org/ and then also compiled with bootstrap values.

p. 26, Table 5.1., Also including the proportion of PSS to the gene length might be more relevant to include than just numbers.

p. 35, Another solution to deal with the multiple testing issue and statistical power when doing gene overrepresentation analysis is the reducing the number of hierarchical GO terms to a certain level – e.g. the highest terms and lowest terms, which at some point are no longer very informative, can be automatically excluded.