PhD Thesis Acceptance Report
Research Discipline Council of Biological Sciences
Jagiellonian University in Kraków

Candidate's name and surname: mgr Kamila Stefania Zając

PhD Thesis Title: Phylogeny and phylogeography of selected species of terrestrial gastropods

Thesis Supervisor: dr hab. Paulina Kramarz, prof. UJ

Assistant Supervisor / Second Supervisor/ Co-supervisor (if applicable): -

Reviewer: dr hab. Piotr Skórka, prof. IOP PAN

THESIS EVALUATION

1. **Scientific merit of the thesis**
   a. Originality of the research *(25-200 words):*

The presented thesis is formed from three published papers in good subject-specific journals. However, there is no single solid research problem that is solved in the doctoral thesis. Topics of these papers are loosely connected. The only common feature is that the model organisms are terrestrial gastropods. This may be perceived as a feeble part of the thesis. On the other hand, this indicates that the PhD candidate can undertake and solve various research topics. Moreover, I must admit that the scientific techniques used in the thesis present a high-quality level.

b. Scientific merit of the chapters / articles *(25-200 words):*

The first publication states a clear research problem which is to identify the origin of invasive *Arion vulgaris* basing on molecular analyses. The scientific level seems to be high and PhD candidate was able to compile data from published sources and her own. The publication contributes to solving the mystery of the origin of invasive alien slug *Arion vulgaris* and disentangles its taxonomical status. In the second publication, the hypotheses are very weakly elaborated. There are many analyses but I am not sure why they were done. However, these analyses are highly advanced and PhD candidate was able to combine various molecular techniques, species distribution modelling and climatic models to understand the phylogeography of the mountainous *Faustina faustina* slug. The third publication is a typical taxonomic paper that aims to discriminate three closely related *Deroceras* species by using molecular markers and investigation of species morphology. Analytical methods used in all papers seem correct (but see my critical notes below).

2. **Substantial merit of the thesis**
   (ability to introduce the research topic and clarity of research hypotheses, the choice of research methods and statistical tools for data analysis, presentation and critical analysis of the research data, the ability to discuss research data and the theoretical background, clarity and quality of the conclusions) (25-200 words):

As I mentioned above, there is no single uniform topic of the thesis. Each publication represents a different research problem. Scientific hypotheses could have been much more elaborated and more deeply set in scientific theories. However, research methods and statistical tools are very good. My major concern is why the phylogenetic trees in the two first publications were constructed only with COI molecular markers despite other markers were investigated as well. Discussion is correct however it lacks a wider ecological or phylogeographical context. I would like to know how these papers contribute to the general science of invasion ecology, phylogeography, biogeography or taxonomy.

3. **Layout and register**
(layout, register and the clarity of the language, the quality of the visual material etc.) (25-200 words):

The thesis is generally well organized and has a standard structure that includes: general introduction, three published publications with authors' declarations of contributions, general discussion, references and supplementary materials. Language is good I found only a few mistakes (but I am not a native Englishman). Moreover, the figures are good although the font could have been larger, especially in phylogenetic trees. General discussion is mostly repetitive and seems to be an extended summary of findings presented in a published paper. Only the last paragraph present prospects for future research and identify gaps in knowledge. I also think the supplementary files are an integral part of publications and could have been placed together with publications that would make reading the dissertation much easier.

4. **Critical notes**

Below I present my critical points, divided into four parts: a detailed review of presented publications and general comments.

Publication #1

In this paper, the authors analyzed the phylogeography of the *Arion vulgaris*, a pest slug in Europe. Using complex genetic methods they concluded that the slug is probably native in central Europe and then colonized eastern Europe but the evidence is not very convincing. Generally, this is a nice piece of science but I have several critical comments on the analysis, presentation of results and interpretation. My major concern is on dividing Europe into four subregions (North, West, Central, East). Why did not authors include southern Europe and ignore the potential impact of geographic latitude on genetic diversity? This division seems to be arbitrary and might have an impact on the obtained results. First, a quick look at the map (Figure 1 in original publication) reveals that the division is artificial. For example, localities in Denmark are close to each other but were classified into two parts of the continent. They are closer to each other than with other localities in predefined parts. This is even more confusing when Supplementary Tables 1-3 are checked. Almost all Denmark localities are classified as Central Europe, none North Europe. The same problem occurs for localities in south-eastern Germany. Every locality in Germany is classified into Central or West Europe, none to eastern Europe despite the map shows something different. This may indicate that the division shown in the map does not fully correspond with the real classification presented in tables. My confusion was deepened by several statements in the text. For example, in Results authors state that "All studied populations of *A. vulgaris* were partitioned into four, approximately equally **strong** regional groups (West, Central, East and North) based on their geographical coordinates". What does the "strong"

mean here? I believe the way the authors' divided localities is the weakest part of the publication, probably wrongly determined statistical analyses and finally interpretation of results. For example, the authors used Mann-Whitney test to compare median haplotype diversity among parts of Europe. This test is thus biased because it assumes that data are independent. This assumption is rarely found in spatial data and violation of this assumption is actually confirmed by authors themselves by showing spatial dependence via Mantel's test performed within regions.

How thus authors should have coped with this problem? In my opinion, they should have abandon partitioning the continent into parts and treat data points as continuous variables. It would make analyses clearer and perhaps easier to interpret. Instead of using Mann-Whitney test they could have used general additive models to reveal a spatial pattern in haplotype diversity. Good software is "mgcv" R package. By introducing an interaction term between splines of geographic coordinates one could have received both great maps and statistical significance of the spatial pattern in haplotype diversity. Also, plotting a pruned phylogenetic tree (a tree restricted to samples of *A. vulgaris*) on the map of Europe would be very helpful. It is easy to create with phylo.to.map() function in "phytools" R package. Such analyses would have improved the reading of the Results in this paper because the description of spatial patterns in haplotype frequency and diversity is rather long and tedious. However, I must admit that authors plotted data in a spatial scatterplot but this is not a map, nevertheless more trustworthy than Mann-Witney tests.

When it comes to spatial scatterplots. I am curious if there is a correlation (corrected for spatial dependence) between haplotype diversity calculated in CO I and Zink finger markers. This test would help in data interpretation and would indicate if results from these two markers are consistent (Mann-Whitney test suggests they are not).

I also have some difficulties in understanding how many sequences were used in analyses. At the end of subsection "DNA extraction, amplification and sequencing" it is stated that 307 COI and 285 ZF sequences were obtained, respectively. However, in the next section entitled "Comparative and phylogenetic analyses" it is given 427 COI and 371 ZF sequences were analyzed including published ones. However, I could not recalculate the sample size of sequences obtained by authors from tissue samples. How can this be explained?

In Results, subsection "Phylogenetic analyses" it is stated that *Arion* species investigated were represented by 21 strongly supported clades. However, I could not find these clades on the phylogenetic tree presented in Figure 6. It seems the authors meant "nodes" rather than "clades", or mixed these terms. But even if we assume these nodes are clades, then we have subclades within clades. This is very unclear. In the same section, the authors stated the following: "All obtained haplotypes for *A. vulgaris* clustered into one clade. This clade clustered together with sequences defined in GenBank as *A. lusitanicus*, which probably are in fact sequences of *A. vulgaris.*" I believe that the latter sentence is unjustified. On what basis do authors suggest these sequences are from *A. vulgaris*? One may say simply that *A. lusitanicus* and *A. vulgaris* are the same species. These *A. lusitanicus* sequences should be depicted in the phylogenetic tree as well as included in the discussion.

Discussion of the paper is interesting however there are statements that are disputable. Authors state that "The Iberian Peninsula is unlikely to be the place of the origin of *A. vulgaris* due to the low densities recorded in the area". This is hardly acceptable inference because most invasive species have a very low abundance or population density in the native area. Invasive species become super-abundant in newly colonized areas. This is a kind of paradigm in the invasion ecology.

I also think if there may be a more convincing way of identification of the (native) region for the invasive species. PhD candidate assumes that haplotype diversity is the highest in the native range (but this is

not rooted in the theoretical background). This is not always true, especially in invasive species. I wonder if it was possible to perform ancestral state reconstruction for haplotypes. Having identified the ancestral haplotypes one could check where (in which regions or countries) these haplotypes are the most frequent and thus indicating the putative native area for the species. Having closer look at the phylogenetic tree presented in Figure 2 (in the Supplementary material) one can notice that the oldest haplotypes within the *Arion vulgaris* clade seem to be numbered 23, 24, 27 (I am not sure if I am correct, though). All of them were identified in the specimens from France. I would like to know the opinion of the PhD candidate on this problem.

I also found no discussion about results from haplotype median joining into the network. Figures 2 and 3 present interesting patterns in the haplotype frequency (however quite different between markers used) together with the number of mutations. This result was not discussed. I wonder if it was possible to calculate the mutation rate in each population and relate it to the geography of Europe. That would also help in identifying the native region of this species.

Publication #2

The introduction is unclear. First, the authors describe the role of Pleistocene glaciation and its role in species distribution. They state that during the last glaciation species retreated to the south and survived in glacial refugia. After the retreat of the glacier, the species began recolonizing Europe northwards. Then, all of a sudden, they start describing Carpatian refugia. It is slightly confusing because these refugia also enable the survival of cold-adapted species nowadays. Also, the introduction does not create a scientific problem. The promising beginning of the introduction ends with the description of the studied species but this is not clear why studying this particular species is important and what hypotheses were tested. Thus, the introduction is disappointing and it seems that this is a paper describing genetic diversity using modern methods combined with ecological niche modelling but how this is grounded in theory and how advances ecology, phylogeography or taxonomy remains vague.

Description of methods is detailed and tools used seem appropriate. Authors quoted work by Groenenberg et al. (2016) who used H3, COI, 16S and CytB markers. I wonder why were not these markers used in the current study? The PhD candidate would have had a much larger sample size for analyses.

Authors use so many abbreviations (besides those standards as DNA, AUC etc.) that reading the paper becomes kind of tiresome. A few of the abbreviations are not fully explained. For example, three past-climate models were used named: CCSM-4, MICRO-ESM, and EPI-EM-P. There is no single sentence explaining how these models differ from each other. By the way, it would be interesting to model potential species distribution by using existing projections of future climate change. This would enable to predict which areas can become refugia and which haplotypes are prone to climate change. Hence, it would allow estimating the effect of climate change on the genetic diversity in this species.

In species distribution modelling authors state they identified areas with long-term stable conditions where the species may have persisted across time from the Last Glacial Maximum to the present day. It was done by summing averaged layers of models for LGM, MH and the present day. I do not fully understand why the sum of predicted suitability at the site was used in this analysis. To explain my concern please imagine the following possible scenario. Let's say the SDM produce following suitability values for three periods: LGM = 0.9, MH = 0.6, present day = 0.3 with total sum equalling 1.8. Then, imagine another possible values: LGM = 0.6, MH = 0.6, present day = 0.6 with total sum also equalling

1.8. However, it is clear that climatic suitability values were very variable in the first scenario (standard deviation = 0.3) but very stable and still high in the second (standard deviation = 0). Thus, using the sum of values for these three periods, in my opinion, is a not perfect method. I would use mean values somehow weighted by the standard deviation. This would take variability into the account and would not produce probably spurious results.

When it comes to results. Authors calculated different population genetic parameters (S, P, Hd), but they did not discuss what these statistics mean. What can be inferred from these parameters? Figure 5 is slightly misleading because of the strange numbering of clades. There should be a reference to Figure 1 where clades are explained and numbered the first time. By the way, the font on the phylogenetic tree (Figure 1) is pretty small and the reading is obscured by dark colours used to show rectangles on the clades. Also, I do not understand why the phylogenetic tree is unrooted (PhD candidate use rooted trees in two other publications).

Discussion is mostly focused on the geographical distribution of haplotypes and the presence of refugia for this species. I miss a more detailed discussion explaining the observed differences in distribution among clades and haplotypes, e.g. more details on the role of dispersal and movements of specimens would be very helpful. Also, this would give some light to the possible recolonization mechanisms. Needless to say, the association of shell colouration with clades and haplotypes should have been better elaborated. In land snails the shell colouration may play a key role in thermal regulation thus including this in discussion seems natural, especially when considering glaciation and possibilities of survival in refugia. Truly, saying the result showing that species could survive in climatic refugia in situ is, in my opinion, one of the most important findings in this work, but slightly counterintuitive. I am curious if any studies are showing the vertical distribution of this species in mountains. Does this vertical distribution change with geographical latitude? Is it possible that this species shifted its vertical distribution towards lowland forest during the glaciation? Also, there should be a paragraph explaining the differences in predictions of glacial climatic refugia between the three climate models used. Throughout the entire paper, especially Discussion, authors use the term "probability of species occurrence". This is not entirely correct. Species distribution modelling produces the suitability index, here climatic suitability. This may not be linked with real occurrence (authors themselves wrongly suggest that Alps appear to have a relatively high probability of species occurrence). This is a simplification which especially dangerous when predictive modelling use pseudo-absences as it was done in this paper. Also, that is a pity that the authors did not use other available environmental GIS layers such as the Numerical Terrain Model which seems the first choice in species distribution modelling of mountainous organisms.

Finally, this pretty long work would have gained importance from a summary or conclusions at the end of this Discussion.

Publication # 3

This is a taxonomic study devoted to the identification of three related *Deroceras* species. I have a few critical comments on methodology and data presentation.

In the description of methods, there are some unclear sentences. In the paragraph "DNA extraction, amplification and sequencing" the authors state that "Sequencing products were cleaned.....and sequenced in both directions in Genomed company." This wording suggests that genetic markers were sequenced twice, which I believe is not true.

The authors also stated that they blasted COI sequences in NCBI BLAST. Why were not sequences of other markers blasted as well?

In Results, the second paragraph, authors state that they obtained "Good quality sequences". What does this mean? Moreover, I think the detailed description of sequences is unnecessary here (although I usually do not read many taxonomic papers and do not know if it is an accepted standard).

I also have some doubts about statistical support for the constructed phylogenetic tree presented in Figure 3. I believe that exact values from Bayesian inference and Maximum Likelihood method should be presented even non-significant ones (a bold font may be used for statistically significant values and normal font for these non-significant). I wonder why the second node of the tree (that separating three studied species from other *Deroceras* species) has BI support equalling 1 but ML is non-significant/not-shown. May this indicate that the basic structure of the tree is uncertain?

Other parts of the thesis

PhD candidate contributed to each paper substantially (declared contribution always above 60 %). However, I am very curious how so detailed a value as 63% was calculated. Declaration of scientific input is required in a doctoral dissertation, however, I doubt if any research team is able to declare individual contribution to exact one percent resolution. In my opinion, so detailed declaration is at most hilarious and distorts science. We are all aware that advanced research is being done with collaboration and everyone within the team wants to solve a scientific problem and no one really cares much what is her/his percentage contribution. What I would like to see in these declarations is a detailed description of what was done by the PhD candidate and other co-authors during the research and manuscript writing.

5. **Final grade** (justification *25-200 words):*

In summary, I believe that the dissertation is a valuable contribution to malacology and taxonomy, despite my numerous concerns. It combines different analytical methods to infer the phylogeny of chosen species, predict native areas of occurrence and solve specific taxonomical problems.

I, hereby, declare that the reviewed PhD thesis by **Kamila Stefania Zając** meets the criteria pursuant to art. 13.1 of Act of 14 March 2003 on Academic Degrees and Academic Title and Title in the Arts (O.J. no 65 item 595 as amended) and request that the Research Discipline Council of Biological Sciences of the Jagiellonian University in Kraków accepts **Kamila Stefania Zając** for further stages of doctoral proceedings.

YES

I, hereby, request that the thesis is accepted with distinctions. Justification *(25-200 words)*

NO

.....
Date: 13.08.2021

.....
Reviewer's signature

INFORMATION FOR THE REVIEWER:

1. Information on requirements concerning PhD thesis structure: http://www.wb.uj.edu.pl/en_GB/stopnie-tytuly/doktoraty
2. A digital copy should be sent to:
nauki.biologiczne@uj.edu.pl

A duly signed original should be sent to:

Rada Dyscypliny Nauki biologiczne
Dziekanat Wydziału Biologii
Uniwersytet Jagielloński w Krakowie
ul. Gronostajowa 7
30-387 Kraków