Skoltech
Skolkovo Institute of Science and Technology

Jury Member Report – Doctor of Philosophy thesis.

Name of Candidate: Anastasiia Stoliarova

PhD Program: Life Sciences

Title of Thesis: Genomic patterns of epistasis at macro- and microevolutionary scales

Supervisor: Professor Georgii Bazykin

Name of the Reviewer: MOLLY PRZEWORSKI

I confirm the absence of any conflict of interest	**Date: 05/11/2021**

The purpose of this report is to obtain an independent review from the members of PhD defense Jury before the thesis defense. The members of PhD defense Jury are asked to submit signed copy of the report at least 30 days prior the thesis defense. The Reviewers are asked to bring a copy of the completed report to the thesis defense and to discuss the contents of each report with each other before the thesis defense.

If the reviewers have any queries about the thesis which they wish to raise in advance, please contact the Chair of the Jury.

Reviewer's Report

Reviewers report should contain the following items:

- Brief evaluation of the thesis quality and overall structure of the dissertation.
- The relevance of the topic of dissertation work to its actual content
- The relevance of the methods used in the dissertation
- The scientific significance of the results obtained and their compliance with the international level and current state of the art
- The relevance of the obtained results to applications (if applicable)
- The quality of publications

The summary of issues to be addressed before/during the thesis defense

This thesis explores the causes and consequences of epistasis among amino-acid sites within a gene through simulations and data analyses, within a highly diverse species as well as in phylogenetic data. It consists of a lengthy literature review, two first authored papers and a first authored preprint. It's an impressive collection and there is no doubt in my mind that the candidate is ready to defend.

The topic is of broad interest, and the work brings many interesting insights to the question. As someone who has worked on linkage disequilibrium (LD) and balancing selection, but never epistasis, I was grateful for the opportunity to read it. In that spirit, I'd encourage the author to think of how the literature review, which is now structured as a glossary, might be transformed into a review. In that regard, they might be interested in empirical findings by Peter Andolfatto and Joseph Thornton's groups (among others) on the predictability of the evolutionary response in invertebrates and vertebrates (e.g., Zhen et al. 2012 Science).

Below, I briefly describe the two papers, then focus mainly on the manuscript, since it is unpublished and closer to my expertise.

The first paper, published in Royal Society Open Access, is based on the ingenious idea of looking for rapid bursts of adaptation-- defined as multiple non-synonymous substitutions in the same gene--in transcriptome data from Lake Baikal amphipods and a multispecies alignment of Catarrhini, two sets of species related by short internal branches. The authors identify some intriguing and convincing examples and show that the phenomenon is not rare.

The second paper, published in Nature Communications, considers the fitness of an allele as a function of time, in the presence of epistasis and fluctuating environments. The authors point out that while epistatic interactions lead to increased fitness via "entrenchment," changing environmental pressures (and more rarely epistasis at a new allele) lead to decreased fitness ("senescence"). They also develop a method to identify the impact of these two forces on phylogenetic data. I was a bit confused about the idea of assigning a selection model (positive or negative) to a site, rather than an allele. But I found the paper thought-provoking and timely, given the many thousands of genomes now coming online.

In turn, the manuscript is motivated by the hypothesis that if the species is sufficiently diverse, epistatic interactions will be seen within species, and for that reason examines patterns of LD in *Schizophyllum commune*, a fungus with up to 20% neutral polymorphism levels (!).

Assuming the mutation rate of this species isn't incredibly high—a detail I'd encourage the author to discuss or at least speculate about--the coalescence events within this species must be extremely deep/old. In that regard, I was surprised to that learn in the Methods section that only 25% of SNP were identical by state (and presumably mostly identical by descent) between the samples from the USA and Russia. It made me want to know more about quite what it means for these two populations to be the same species. In a similar vein, rather than excluding samples from Florida based on geography, I was curious to see some representation of genetic similarity (eg a PC analysis based on a few thousand SNPs).

The Results consist in three parts:

1) Simulation results indicating that only in highly diverse populations should compensatory alleles arise before the loss of the deleterious allele, and hence only in such populations will epistasis generate linkage disequilibrium (LD). I found these results quite interesting and convincing.
2) Next comes a data analysis looking at pairwise LD in *S. commune*, *D. melanogaster* and humans. The findings are consistent with simulations in that only in *S. commune* and not in these less

diverse species is LD between non-synonymous sites higher than for synonymous ones. That's pretty cool. However, these plots also show that this measure of LD (the expectation of r2) is lower for synonymous than non-synonymous sites in these other species, highlighting its sensitivity to allele frequencies (AF). In that regard, it seemed to me that these plots should be stratified by allele frequency and not just distance.

3) The third section of results is an analysis of haplotype blocks and other properties of the data that are interpreted as evidence for balancing selection. Here I worried quite a bit about the impact of (AF) on the statistics, for instance for statements such as "*Polymorphic sites within haploblocks are characterized by higher MAF than that at sites that reside in non-haploblock regions*" (p. 83).
   I also wondered about alternative explanations e.g., when observing that LD is higher in genes with greater pn/ps, could both reflect Hill-Robertson interference?
   Similarly, the author interprets the observation of two deeply divergent haplotypes as evidence for balancing selection, which I interpret to mean any selection pressure that maintains alleles in the population substantially longer than under neutrality. In that regard, it seems to me that an alternative to consider and exclude is that recombination rates are low (as under a constant population size model and no recombination, two haplogroups are expected; eg. see Hudson's 1990 review of the coalescent). Moreover, in this species with unusually high diversity, it seems possible that there isn't always enough of a stretch of homology for recombination to occur, thereby generating diversity-dependent cold spots. Is anything known about the recombination landscape of this species or related ones?
   More generally, I was also not sure I understood some of claims about LD. For instance p. 87, the author states that synonymous sites are on average older, but wouldn't that make the LD levels lower not higher? Here too, it seemed useful to me to compare r2 for synonymous and non-synonymous sites while matching allele frequencies of the pair of sites.

On a minor note, I thought a few details were missing from the Methods, like the sequencing coverage after read mapping. Also, the author might also be interested in Callahan et al. 2011 PLoS Genetics.

In summary, I found the hypothesis stimulating and plausible, and the data analysis intriguing—though I had some suggestions of further analyses for them to consider. I was less convinced about the evidence for balancing selection, and thought it might be helpful to test an explicit neutral null model. These concerns notwithstanding, I learned a lot from reading the manuscript, and expect an impactful paper to come out of it.

Provisional Recommendation

X *I recommend that the candidate should defend the thesis by means of a formal thesis defense*

☐ *I recommend that the candidate should defend the thesis by means of a formal thesis defense only after appropriate changes would be introduced in candidate's thesis according to the recommendations of the present report*

☐ *The thesis is not acceptable and I recommend that the candidate be exempt from the formal thesis defense*