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: Prof. Brian Charlesworth

I confirm the absence of any conflict of interest 	**Date: 09-11-2021**

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

The thesis presents the results of three research projects, plus a literature review and concluding summary. It is well-organised and generally clearly written, especially given that English is not the candidate's native language. The work is of high scientific quality and is internationally competitive. The results are highly relevant to the general scientific question involved. The work is of relevance to our understanding of evolutionary processes at the molecular level, and has no immediate practical applications. Two chapters of results have already been published in well-respected scientific journals, and represent significant contributions. Chapter 3 has not yet been published in a peer-reviewed journal.

My detailed comments on the thesis are presented below. Some of these are minor corrections, others are points that probably require discussion at the defence, and others are scientific points (e.g. additional references or corrections of errors) that the candidate should probably include as revisions without discussion.

One general comment is that the reader of the thesis would have been helped by summaries of the major chapters (2 to 4) at the beginning of each chapter; these would have made it easier to digest the often quite complex material. If such summaries are allowed, it might be useful for them to be added to the thesis.

Chapters 1 and 2

These introduce the subject of the thesis, and focus on the evolutionary role of epistatic fitness interactions. Ms Stolyarova has clearly done a thorough job of reading the literature on this subject, and presents a comprehensive review of much recent work, which is well organised and generally clearly written. The wealth of material makes it hard, however, to appreciate any general principles that may have emerged from this work. A brief summary or set of main conclusions would thus be helpful.

There are also some omissions that are a little surprising, and there is a general tendency to cite quite recent references for results that were discovered long ago; this would be appropriate if these were to standard textbooks or review papers, but often they seem to be to somewhat arbitrarily chosen research papers. There is, for example, no mention of Wright's shifting balance theory of evolution, which is probably the main example of a way in which epistasis could provide a different mode of adaptive evolution from selection acting within a population. Indeed, the concept of adaptive valley crossing is attributed to Gavrilets (2004), although it originated with S. Wright (p.28). There is also no mention of the fact that R.A. Fisher (1930) discussed both the selection pressure to reduce recombination among epistatically interacting loci, and the advantage of recombination in reducing selective interference among loci subject to directional selection. While fitness landscapes are frequently mentioned, Lande's important use of the derivative of mean fitness with respect to mean trait value in models of quantitative trait evolution is not cited. There is a discussion of theory on the interaction between recombination and selection (pp.37-41), but no mention of relevant empirical evidence, e.g. the relation between genetic diversity and recombination rate. Similarly, there is no mention of supergenes or inversions in relation to epistasis between polymorphic loci.

Some more detailed comments follow.

p.12 l.3 A reference to Fisher (1918, Trans Roy Soc Edinburgh 52:499) should be provided.

p.14 As discussed in Provine's 1986 biography of Wright (pp.307-317), Wright was inconsistent about what he meant by a fitness landscape. For studying evolutionary dynamics within populations, it involves the relation between population mean fitness and genotype frequencies, not between individual fitnesss and phenotype or genotype. It might be worth mentioning this distinction; the thesis treats landscapes purely in terms of individual or genotypic fitnesses.

p.16 l.6 I think it's confusing to describe dominance as a form of epistasis; there is a real biological distinction.

l.15 'Monotonic' is the maths term; 'monotonous' mean 'boring'.

p.21 Is there really a distinction between 'holey' and 'BDM' landscapes?

p.26 l.8-10 There are many qualifications to the statement that "Under selection only, the average fitness of a population always increases." Fisher himself would have strongly disagreed with this statement (see his 1941 paper, Annals of Eugenics 11:31-38, where he gives the example of the spread of a mutation that alters the selfing rate). It is not true with frequency dependent selection. Linkage and strong epistasis also can cause an equilibrium population to depart from a fitness peak, even with constant fitnesses.

p.28-29, p.34 It should be made clearer that these experiments are not representative of natural evolution in most cases, since there is no recombination and they start with a genetically uniform population. Malmberg (1977, Genetics 86:607) pioneered this type of experiment, and pointed out that more epistasis is expected (and found) with clonal reproduction than when recombination is allowed.

p.31 l.4 from end. This statement is too extreme (see comments re p.26).

p.37 last § l.1 The work of Fisher (1930) and Muller (1932) long pre-dates these people.

p.39 §2 l.5 This reference is wrong: it should be Maynard Smith and Haigh (1974, Genetical Research 23:23-35).

l.7 'loci' not 'locus'

l.11 'associative' not 'associate'

l.1-3 from the end. Interference also applies to positively selected loci, which is what Hill and Robertson studied.

p.40 A formal analysis of mutation-selection equilibrium under these two types of selection was first presented by Charlesworth (1990, Genet Res 55:199).

p.41 Figure legend 'Fitness flux' is not defined.

p.42 Frequency dependence was (again) first studied by Fisher (1930) in relation to mimicry, and by Wright (1939, Genetics 24:538) in relation to self-incompatibility.

p.44 last §, l.2 Heterozygote advantage was discovered by Fisher (1922, Proc Roy Soc Edinburgh 42:321).

p.45 §1, l.5 from end. What is meant by 'allelic preferences?

p.46 Figure caption. What is meant by 'propensities'? What's the relation between 111 and 168?

l.5 The term 'Stoke shift' doesn't convey anything in relation to evolution; I thought it had something to do with emission and absorption spectra.

p.49 Figure 2.21 is really hard to understand.

p.50 §1, l.8 You can have stable equilibria under mutation and selection with multiplicative fitnesses, even without recombination, so this statement is inaccurate (see Kimura and Maruyama 1966 Genetics 54:1303).

p.52 §1, l.8 How can a paper not have a date?

p.55 §2 l.2. Site and allele should be defined precisely.

l.3-5. This is hard to understand; surely, neutrally evolving sites will accumulate different 'alleles'.

p.57 §2, l.10 What is DCA?

p.60 §1, l.4 You don't need GWAS for this; simple analysis of trait variance into components tells you about this- it's been known for a long time that most quantitative traits lack evidence for much non-additive variance (including dominance)- see standard textbooks on quantitative genetics.

§2, l.2 This goes back to Fisher 1930, and is well reviewed in the standard textbooks on population genetics.

l.7 Independent segregation means the recombination fraction = 0.5, not absence of LD.

l.8 Shouldn't this be that conditions for LD are restrictive? Again, this is in standard textbooks.

p.61 Figure 2.26 'Repulsion' should be 'Repulsion LD'.

p.62 §1 l.2 The comma should come after 'negative'.

§2 l.1-2 This was discovered long before these references (see standard textbooks).

Chapter 3

This chapter describes an analysis of a set of whole genome sequences of the mushroom *Schizophyllum commune*, adding new sequences to those used by Baranova *et al.* (2015). This species is especially interesting, since it has the highest level of natural DNA sequence variability ever described. The work appears to have been done very competently, and is mostly described clearly. The most interesting finding is the existence of numerous tracts of high levels of linkage disequilibrium (haploblocks). It is proposed that these reflect widespread balancing selection involving epistatic fitness interactions between polymorphic variants. It is undoubtedly a useful contribution to a scientific field of considerable current interest.

I have some general comments, which the candidate should consider.

First, it would be helpful to have had some more background information about the system, e.g., the genome size, the number of chromosomes, the density of coding sequences, the typical gene structure, the mating type loci of this species, whether there is any information about divergence from a related species, etc. I realise that much of this information can be obtained elsewhere, but the reader of a thesis should not have to go to the trouble of looking it up.

Second, nothing is said about recombination rates or a genetic map. It seems that a map is not available at present, although Seplyarskiy *et al.* (2014) discussed the effect of diversity on crossing over rates, and the use of parent-offspring trios to measure the mutation rate was described by Baranova *et al.* (2015). It is extremely hard to make rigorous interpretations of polymorphism patterns and LD patterns without information about recombination rates and their relation to genomic location. This, of course, is beyond the control of the candidate, but should have been at least alluded to when discussing the interpretations of the results. Another mushroom species (*Pleurotus tuollersis*) has a detailed genetic map, revealing what looks like extensive centromeric suppression of crossing over and recombination hotspots (Gao et al. 2018 *BMC Genomics* 19:18). I would presume that these general features might well apply to *S. commune.*

Third, when discussing the possible role of epistasis in creating the haploblocks, it would be worth making clear that the maintenance of LD among polymorphic loci requires departure from additivity of fitness effects, in contrast to mutation-selection balance, where departure from multiplicativity is

required. Thus, with low recombination LD can be maintained with purely multiplicative fitnesses; this is the basis for the "crystallization" of the genome in Franklin & Lewontin (1970).

Some more detailed comments follow.

p.63 l.2-3 This statement is too strong; additivity seems to describe quantitative traits rather well. I suggest adding 'often' before 'engage'. Smith (1970) should be Maynard Smith. Why no mention of Wright, the early enthusiast for epistasis?

l.9-15 The point made by several of these papers (Hill et al. 2008, Crow 2010 and Maki-Tanila & Hill 2014) is that you can have quite a bit of dominance and epistasis at the level of the control of the phenotype, but these effect does not create non-additive variance because of the statistics of genotype frequencies. As already mentioned, the role of epistatic selection in creating LD goes back to Fisher (1930), and is described in standard textbooks.

l.16-17 I find the distinction between macro- and microscopic confusing; surely, the components of variance in a population are macroscopic not microscopic.

l.4-7 from end. This seems to ignore the problem of how an unfit allele combination (i.e., one which is an adaptive valley) can persist in a population, especially if N is very large so that the efficacy of selection versus drift is high.

p.64 l.2 'to' is misspelt.

p.66 §1 l.2 from end. How was diversity at synonymous and nonsynonymous sites calculated (there are several different algorithms)?

§2 l.1 This is not a true phylogeny, since recombination is occurring within populations, and phylogenetic reconstruction assumes no recombination. It can only serve as a guide to overall sequence similarity.

p.67 Caption to Fig. 3.1 The sample sizes should be given here.

p.68 Caption to Fig. 3.2. More detail about distances among samples with USA and Russia would be helpful.

p.69 The title of this section is misleading, as recombination rates were not estimated.

§1, l.2 What was done about multi-allelic sites?

§3, last l. It's not entirely clear what is meant by physical distance. It seems that separation along the protein sequence is ignored, and that this is the Euclidean distance in 3D space.

p.70 §1, l.3 This is not a genetic distance, which is measured by recombination frequency.

§1, last l. A reference to BH should be given.

p.71 §2, l.1-2 What's the rationale for this choice of software rather than the popular SLim? Are these simulations also done with no recombination?

l.3-4 Changing the mutation rate rather than *N* of course obscures possible effects of different *N* on the efficacy of selection relative to drift. A brief discussion of whether high mutation rates or high *N* is involved in the high diversity would be useful.

p.72 §2, l.3 from end. If there is no epistasis, how can there be compensation?

§3, l.1 The model is not entirely clear; were reverse mutations allowed at individual sites. If not, how can there be an equilibrium?

l.5-6 I am puzzled by this. Deterministic theory shows that, with sex but no recombination, mean fitness under mutation-selection balance is increased by negative epistasis and decreased by positive epistasis (e.g., Charlesworth 1990 Genet Res 55:199)- you seem to claim to see the opposite. How can this happen?

p.74 §1, l.3 Cutter *et al* . (2013 Mol Ecol 22: 2074) on hyperdiversity might be cited.

§2, l.1. I believe the sample size was 32 for the USA population. A MAF of 0.05 means an expected number of 1.6, so only singletons seem to be excluded. This should be clarified.

p.75 Caption to Fig. 3.4. d-e. The picture of the landscape is rather vague; what are the X and Y axes supposed to represent. The interpretation seems rather handwaving. It could simply be that positive epistasis reduces the efficacy of selection, so such alleles reach higher frequencies.

Last § Why isn't the evidence for excess positive LD shown? This seems quite important for the interpretation. Langley et al. (2012, *Genetic*s 192:593) claimed to have evidence for positive LD with respect to the more frequent variants, which they interpreted as evidence for selective sweep effects. Could these differentially affect nonsynomymous and synonymous variants?

l.78 §2 I find this confusing. As already noted, negative epistasis under mudtation and selection should lead to lower allele frequencies, the opposite of what is said here. It's also not clear what the relevance of Barton (2017) is to molecular data.

§3 l.5 'The' is missing from the beginning of the sentence.

p.80 Caption to Fig.3.8. b-d It would be helpful for the meaning of the triangle plots to be explained.

§1 'The' is missing from the beginning of the sentence.

l.2 What is known about recombination in relation to physical location on the chromosome? Centromeric and telomeric suppression of recombination is widely observed; also, there are two complex loci controlling mating types A and B, which presumably have suppressed recombination.

Also, could there be inversion polymorphisms, creating local regions of high LD, with differentiation at both NS and S sites? Small inversions are known in the mimicry genes of butterflies (e.g., Joron *et al.* 2011).

§2 l.4 'of' missing after 'thousands' .

p.81 Caption to Fig.3.9. Please show the physical scales.

p.82 §3 l.1 'p_n/p_s' is not defined, nor it explained how it's been estimated. Also, the rationale for looking at this should be explained: why should higher p_n/p_s be associated with an excess of NS LD?

§4 l.1 Where is this shown? It should be made explicit.

p.83 Caption to Fig.3.9. a. It's not clear what the x and y axes mean.

b. What is the expected MAF under neutrality (I make it approx. 0.2457)

§1 The observation of this large-scale LD at first sight suggests epistatic selection of the type first proposed by Fisher (1930), and analysed in detail in subsequent work by Feldman, Karlin, Kimura, Lewontin and others (see the standard textbooks). I think the more precise population genetics

framework is more helpful than the rather vague statements about fitness peaks, especially as these ignore the interplay between recombination and selection.

p.85 §1 The possible role of small inversions should probably be mentioned.

p.86 §1, l.3-4 Why does this follow? Also, with a high product of N and mutation rate, recurrent mutation could cause allele sharing under neutrality.

p.90 §2, l.1 'excess attraction' presumably means LD that involves excess combinations of pairs of common and pairs of rare variants.

l.3 from end. Presumptive examples of this are known in mimicry genes (Joron *et al.* 2011; Kunte *et al.* 2014 *Nature* 509:229).

§2 Some discussion of the plausibility of AOD in relation the presumably large Ne of this species would be good. The final sentence does not seem to have a firm theoretical basis.

p.91 The possibility that the recombination landscape of this species may be involved in creating these unusual patterns should be discussed.

Chapter 4

This is an interesting and fairly straightforward analysis of patterns of protein sequence evolution, and has already been published in a leading journal. I therefore have no major comments on this chapter, other than one concerning the possibility of Hill-Roberson interference (HRI) during bursts of evolution (see my comments re p.106).

Some detailed comments follow.

p.92 §1 l.3 This seems to equate unequal evolutionary rates with punctuated equilibrium. I think this is inaccurate; but unequal evolutionary rates were described by G.G. Simpson and others long before punctuated equilibrium was proposed, which was dressed up (misguidely in my view) as presenting a challenge to neo-Darwinism.

p.93 §1 Perhaps Gillespie's claims for non-constant rates of protein sequence evolution should be mentioned here.

§2, l.1 It would be helpful to indicate what time-scale is involved here; we know that beneficial mutations can spread over the course of 100 generations or so if they are sufficiently strongly selected, but molecular evolution usually involves much longer periods.

p.94 §1, l.1 'primate' not 'primates'.

p.95 §1, l.1-2 This is too simplistic you can have much positive selection and still have dN/dS > 1 over the whole sequence. There are also conditions under which purifying selection gives dN/dS > 1 at individual sites (Lawrie et al. 2011 Genome Biol. Evol. 3:383).

§2, l.2 This should be qualified; it assumes absence of selection on codon usage.

l.7 from end. 'of' not 'on'.

p.98 Caption to Fig.4.1, l.4. 'in units of dS'.

p.101 Caption to Fig.4.3, l.3. How reliable are ostensibly high values of dN/DS; if dS is low, dN/dS could be high by chance.

p.102 §2, l.1 ‘macaque’.

p.104 §2-3 There are, of course, examples from other systems, e.g. Drosophila (Presgraves group work on the meiosis genes *mei217/218*), and for non-coding sequences (e.g. the Pollard group work on human lineage specific regulatory changes, HARs).

§3 l.1 The brackets around the citations are incorrectly formatted.

l.3 ‘biased gene conversion’

p.105 §2 l.9 it would useful to say how many generations is involved. With a mutation rate of 10^{-8}, this would be 10^5 generations, and longer if selection on codon usage is taken into account. It could be twice this, with a Drosophila-type mutation rate.

p.106 §2, l.11-12 It should be specified that s is the selective advantage to a heterozygous mutation. The substitution rate formula can be found in Kimura (1983, p.48).

l.14 This calculation isn’t very informative about whether HRI is occurring- you also need to know the mean duration of each fixation, which is $T \approx 2\ln(4N_e s + 0.5772)/s$ generations, assuming no dominance (Hermisson & Pennings 2015 Genetics 169:2335). The mean number of simultaneously segregating mutations in a given generation is the product of T and the substitution rate: $8N_e\mu \ln(4N_e s + 0.5772)$. This is what is relevant to HRI. It is only weakly dependent on the selection coefficient, and is only larger than $8N_e\mu$ by a relatively small factor. But the mutation rate that is relevant is that for the whole protein, not the individual site; if 200 amino-acids are assumed, and 2/3 of sites are nonsynonymous, $8N_e\mu$ should be 267 x 0.01 = 2.7, if $4N_e\mu = 0.01$, so there is some scope for HRI.

The conclusion about HRI thus probably needs modifying.

Chapter 5

This is an interesting but quite complicated analysis of patterns of protein sequence evolution, and has already been published in a leading journal. I do not have any major comments.

Some detailed comments follow.

p.109 §1 l.2 Maynard Smith not Smith.

§2 l.1 ‘causing’ not ‘entailing’

Last l. The mysterious ‘Stokes shift’ makes another appearance.

p.110 §3 I am not sure that ‘fluctuating’ is the right term; this implies periodic or stochastic reversals of the direction of selection. However, ongoing positive selection could be caused by a steady change in the state of the environment or by arms races with parasites or predators.

p.111 l.7 Isn’t there a danger of bias when you use dN for branch length and are also looking at amino-acid changes?

l.8-10 This method seems to have low power to detect positively selected amino-acid changes, as it consistently gives much lower estimates of the proportion of positively selected fixations in proteins compared with McDonald-Kreitman type approaches.

p.113 §1.6 This is a little confusing, as only a limited number of amino-acids can be accessed by single mutations from a given codon.

l.7 What is the rationale for using log fitness?

l.115 §2, l.6 I think ‘substitution rate’ is meant here.

§3 l.8 The relevance of ‘variance’ is not clear. It has not been mentioned previously in this context.

p.118 §1, l.1-2. As with all such validation methods, it only tests what happens under the assumed model- it doesn't test robustness to deviations from the assumptions.

§2, l.2 Confusion matrices should perhaps be explained.

§3, l.4 from the end. The meaning of 0.08 is not clear; is this a relative or an absolute value?

p.120 §4, l.2 What does ‘quenched’ mean?

p.121 l.2 It seems odd not to mention Gillespie’s work in this context. Kimura’s paper dealt with allele frequencies not substitutions; anyway, it contains a mathematical error, which was pointed out by Gillespie (1973, Theor Pop Biol 4:193).

p.123 §2, l.4 ‘less frequently’ not ‘rarer’.

p.126 §1, l.6 ‘allows us’

p.128 §1, l.2. l.6 ‘allows us’

l.129 §2, l.4 Strictly, it’s a fixation probability not a rate.

p.131 §1, l.4-5 Note comment re p.111, l.8-10. This is also relevant to the statement on p.132, l.1; it’s likely that many sites with $w < 1$ are actually under positive selection.

p.135 §1, l.3-4 I don’t see that epistasis is needed; you could simply be climbing towards an adaptive peak.

§3, l.4 How can negative selection favour something?

l.3 from end. Please explain what a ‘stairway to heaven is’.

p.136 §1, l.4 Can you really say that senescence causes anything? It’s just a descriptor of the pattern of evolution.

Chapter 6

This is a concise summary of the main conclusions.

p.137 §1, l.6 Perhaps qualify to say ‘possibly indicative of’

l.10 ‘evidence for’ not ‘evident on’.

§2, l.2 ‘at’ not ‘of’

l.5 delete 1st comma.

l.6-8 My impression was that the LD involving combinations of nonsynomymous variants was mostly in the haploblocks, possibly involving balancing selection rather than deleterious variants.

p.138 §1, l.1 ‘the high…’

§2, l.2 from end ‘conserving’.

§3, l.2 Maybe qualify by ‘can accumulate’; most of them don’t.

l.4 ‘evidence’ not ‘evident’

l.5-9 It's not clear to me how you can distinguish an environmentally caused change in the direction of selection from epistatic effects.

§4 l.1 'evolutionarily', 'in' not 'of'.

p.139 §2, l.2 'it's derivative' not 'derivative of'.

l.3-4 It's the sites that are negatively selected not the alleles. You don't know if the variants that get substituted are positively selected, neutral or weakly negatively selected.

§3, l.1 'show' not 'bear'.

l.4 from end 'draw conclusion about' not 'conclude on'.

Last l. 'landscapes' not 'landscape'.

Provisional Recommendation

☒ *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*