Skoltech
Skolkovo Institute of Science and Technology

Thesis Changes Log

Name of Candidate: Valentina Burskaia

PhD Program: Life sciences

Title of Thesis: Positive selection in parallel evolution

Supervisor: Prof. Georgii Bazykin

I wish to thank all reviewers for the time they took reading my manuscript and for their kind and helpful comments. The thesis document includes the following changes in answer to the external review process.

Reviewer: Konstantin Severinov

This thesis concerns with analysis of two examples of positive selection driving parallel evolution, one in Lake Baikal amphipods, another - in mitochondrial genes of some birds. The results are presented in two separate chapters of the thesis, each with their won Introduction, Materials and methods, Results and Discussion subsections. Work on amphipods has recently been published, with Ms. Burskaia being the first (and corresponding) author.

Comment 1:

As presented in the thesis, this entire part as a word for word copy of the published work. I assume the other chapter is similarly a preprint of work submitted for publication. I think it is worth considering how appropriate this is. Even more time could have been saved when making this thesis by simply including the pdf of the published work, for example.

Answer:

Indeed, Chapter 3 has been published in Genome Biology and Evolution, and Chapter 4 has been submitted to the same journal (and a preprint of Chapter 4 has been deposited onto bioRXiv). In the revision, I have introduced changes and added new text throughout the manuscript, thus diverging the thesis content from that of the corresponding papers (see below). I also rewrote introductions to Chapter 3 and Chapter 4, emphasizing how they represent a part of a coherent study. New data was also added in Chapter 3, section 3.3.8 “Contribution of homoplasy and hemiplasy in parallel evolution”.

Comment 2:

On the other hand, since the published paper contains 8 authors and in the Personal Contribution section of the thesis it is stated that Valentina did “most of the bioinformatics and conceptual work”, while Prof. Bazykin “…participated in writing”, I can’t help but wonder how much of the text of the thesis was actually written by its sole author.

Answer:

In the revision, I have expanded the Personal Contribution section, explicitly spelling out the roles of all coauthors. As explained in this section, I performed all analyses except those mentioned. I wrote all the text of both papers that constitute Chapters 3 and 4, and the remainder of the thesis text. Georgii Bazykin edited the text, with minor contributions from other coauthors.

Comment 3:

At the very least, I would like to see, for example, that Figures which were supplementary in the published work (and remained so in the thesis) were moved to the results part of the thesis and were more thoroughly discussed. If nothing else, this would make the text more accessible to readers outside of the immediate field.

Answer:

As suggested, all figures from Appendix A and Appendix B are now moved to the results parts of Chapter 3 and Chapter 4. I have now expanded description of Figures 3.4, 3.5, 3.7 (former Figures A2, A3, A4). I now also discuss them in the text in more detail, mainly as the new section 3.3.7. "Methodological factors that could affect P test values". Former Figure A5 was merged with Figure 3.1 by including bootstrap values from figure A5 into figure 3.1.

Comment 4:

I believe a better job needs to be done to show how the two parts of the work are related to each other, making a single thesis united by a common avenue of inquiry.

Answer:

I added the "Goals and objectives" section to demonstrate the unity of direction of the research and its objectives. I also added new text to chapter introductions to explain how they constitute parts of a common research agenda.

Comment 5:

Further, I would like to see a structured statement of the scientific problem being attacked followed by the aims, scope, and goals of the work presented.

Answer:

I rewrote and expanded the "Relevance and significance of the work" section to describe the scientific problem more explicitly.

Comment 6:

I would then like to see conclusions written not as a free narrative, as it is done now, but as an itemized list of statements that allow someone who is not an evolutionary biologist to understand what exactly was discovered during the work and how the field was advanced.

Answer:

I reformulated "Conclusions" section to shorten the main ideas and added discussion of the significance of discoveries made. As suggested, I also add an itemized list of conclusions, making sure that no jargon is used to make it as accessible to non-specialist audience as possible.

Comment 7:

A statement about possible practical implications of obtained results would have also been useful.

Answer:

I rewrote the "Implications of this work" section in Introduction, elaborating on the applied significance.

==========

Reviewer: Matthew Hahn

This is a very nice dissertation consisting of three substantial chapters. After a brief introduction, the first chapter is a review of the concepts and ideas studied in the dissertation, the second chapter is a study of parallel molecular evolution within the whole genomes of three clades, and the third chapter focuses on parallel molecular evolution within the mitochondrial genome of birds. The ideas being addressed are at the forefront of questions in evolutionary genomics right now, and the methods used to answer these questions are appropriate and up-to-date. There is a lot of excitement about convergence and parallelism, especially with the growing ability to identify the processes leading to them. The number of available whole genomes makes all of this possible, and the dissertation leverages such data to answer important questions.

Comment 1:

All chapters are well-written and up to the standard expected at my own university and other universities around the world. The scientific results presented are of high quality, with chapter 3 recently having been published in a top journal in our field. I foresee little trouble publishing chapter 4, though if chapter 2 is to be published as a stand-alone review paper I think it would need more work. Right now, chapter 2 ends without much of an overall conclusion or roadmap for methods that do work, two things that would be expected in a good review paper.

Answer:

Thank you for the comment. I added a short conclusion, as actually the review was meant to be an introduction into the relevant literature for chapters 3 and 4 rather than a stand-alone work. Roadmap for modern methods definitely implies more thoughtful work.

Given all of this, I do have some scientific issues that I would like to see the candidate address in their formal defense, as well as some minor typographical and other errors that should be changed in the thesis.

Comment 2:

One issue that is raised in the thesis is the idea that hemiplasy can be driven by adaptive natural selection. The example given in the text is the Eda gene in sticklebacks, where repeated selection on an ancestral variant has led to fixation independently in multiple lakes. I very much agree with this formulation, as it separates the mutational process (hemiplasy/homoplasy) from the selection process (neutral/adaptive). The line between these becomes a little less clear in cases of balancing selection, where selection is maintaining ancestral variation, but this possibility is not a focus of the thesis. However, given this distinction, I wondered why in the Discussion of chapter 3 hemiplasy is dismissed so readily. Given the stickleback example above, it seems quite straightforward for selection to favor ancestral nonsynonymous variants over synonymous variants. Indeed, this is one of the findings of Pease et al. (2016, PLoS Biology). The conclusion here therefore requires much more justification.

Answer:

Thank you for this important comment. First, I wish to stress that the mode of origin of the mutation (hemiplasy/homoplasy) does not affect our conclusion that the results of Chapter 3 cannot be explained without invoking adaptation.

Having said that, I agree that the obtained results do not allow to establish the origin of that variation completely unambiguously. Homoplasy is needed as the explanation for distant amphipods because the maximum distance between path I and path II in species quartets exceeds 10%, and we observe P>1 even for that remote species. A neutral polymorphism would be unlikely to survive for that long prior to it becoming adaptive. Of course, if the mode of selection that operates is in fact balancing, then polymorphisms can survive for very long; but we provide arguments against balancing selection. Hybridization also is quite unlikely for species of 10% divergence.

For closer groups of gammarids hemiplasy seems to be a good source for discovered parallel adaptation. I believe that it easily can be a case. And parallelisms in cichlids seems to be made of ancestral variation almost entirely, as I now described in the end of new section 3.3.8. "Contribution of homoplasy and hemiplasy in parallel evolution". Discussion of Chapter 3 was also slightly changed to mention hemiplasy as a mechanism of parallel adaptation.

Comment 3:

It was unclear to me what a "robust" phylogeny of amphipods means, though it is likely referring to the bootstrap support on each branch. Much more important for inferences about hemiplasy, however, are concordance factor values on each branch. It is reported that the cichlid phylogeny has a lot of incongruence, but not information is given on the amphipod clade. A comparison of these values between phylogenies may help to reveal why patterns of parallelism are so different between them. As it stands, there is little biological or technical explanation as to the difference between cichlids and amphipods.

Answer:

Indeed, the “robustness” of the phylogeny mentioned in the text refers to high bootstrap values.
I absolutely agree: tree concordance should be measured to estimate the phylogeny. Unfortunately, in our transcriptomic data there are few genes of sufficient length. It makes estimation of incongruence unreliable at regions of phylogeny with fast branching. Now the results of incongruence analysis are described in the beginning of new section 3.3.8. “Contribution of homoplasy and hemiplasy in parallel evolution”. This section also discusses the differences between cichlids and amphipods.

Comment 4:

On a related note, I did not understand where the colonization history of Lake Baikal came from. Certain sub-clades are listed as having colonized first or second, but it was not clear how this was inferred. I do not think this information can be gleaned from the phylogeny shown in Figure 3.1. Some further explanation is therefore needed.

Answer:

Sorry I did not mention that in the text: two independent invasions were discovered in previous papers (Including Naumenko et al. 2017). Gammarus lacustris, which separates two clades, was sampled in a small lake adjacent, but not connected to Baikal. It proves that invasions were independent at our phylogeny (too). Now I added missing description to the Figure 3.1.

Comment 5:

In section 3.3.2, results on parallelism for nonsynonymous and synonymous substitutions are given, with a higher rate for nonsynonymous than synonymous. But the text concludes " in strength of negative selection among sites," and goes on to introduce the P-test. I do not understand how a difference in slopes could be due to varying negative selection among sites, nor how the P-test would control for such an effect if present. Perhaps this can be explained more clearly.

Answer:

Speaking about “differences in strength of negative selection” among sites I meant that negative selection constraints allow only few amino acid positions in each site, while in most synonymous sites all mutations allowed. I believe it is main explanation of differences in slopes at Figures 3.1B,C. This is now rephrased for clarity.

The P test uses normalisation, which should not cause P>1 when nonsynonymous substitutions are neutral (even when they are limited by constraints). Indeed, if N(A,B) nonsynonymous substitutions occurred along path I (no matter what happens in path II), the probability of a parallel (A,B) mutation in path II will be either equal to neutral (and will result in *M* parallel mutations), or (if negative selection constraints prevent that mutation) it will be less than neutral (and will result in L<M parallel substitutions).

The first case will give dNp = M / (M+N), which is equal to dSp and results in P=1.
The second case will give dNp = L / (L+N), which is lower than dSp and results in P<1.

Therefore, no combination of constraints can result in P>1.

More minor issues:

++The "GWAS" approach in chapter 4 seems a lot like the "phyloGWAS" approach laid out in Pease et al. (2016, PLoS Biology).

Indeed, we use the same idea, I added a reference.

++In chapter 2 the text says that McGee et al. (2020) found that hybridization led to a shared predatory phenotype. But I do not believe that this paper showed any evidence for hybridization.

Yes, the evidence is inconclusive as only in small frame it works and could be a random fluctuation; removed.

++Avise and Robinson (2008, Systematic Biology) created the term "hemiplasy," not Hahn and Nakhleh (2016).

Yes, but I need to use the term for both ILS and Hybridisation, while Avise and Robinson use it only for ILS.

++I do not understand the effect ascribed to bottlenecks on p. 19 of the thesis. Also, there is a typo in "Lefebure et al" at the same location.

Thank you. I mean that if in some position particular mutation type is more probable than others, bottlenecks can rise probability of random fixation of those mutations.

++"Hahn" is misspelled on pages 20, 21, and 66 ("Hanh").

I am really sorry for that, now fixed.

++"withinspecies" needs a space between the two words (p. 26).

Fixed.

++"Figure" is misspelled twice on p. 42 ("Figire").

Fixed.

++"McDonald" is misspelled on p. 44 ("MacDonald").

Fixed.

==========

Reviewer: Fyodor Kondrashov

This is a solid thesis that details the efforts of two directions of research: to study the frequency of parallel evolution in protein sequences (using three groups of animals) and to look for a pattern matching amino acid changes in mitochondrial genomes with phenotypic changes in birds. The first venue of research revealed very interesting and exiting results,

the second was more emblematic of the usual outcome of research projects, yielding mostly negative results, and, therefore, was quite fitting for a PhD thesis.

I found the questions to be interesting and the methodology appropriate. I especially complement the candidate on the extensive efforts to take into account various confounding factors in the first part of the work (alignment, errors, trees, etc) – it is a very solid piece of work.

The pattern of more nonsynonymous parallel substitutions (than synonymous ones) in amphipods is a great discovery. The candidate is leaning on positive selection as the underlying mechanism, in part seemingly convinced by the pattern that polymorphisms at those sites seem to have lower frequency than synonymous polymorphisms.

Comment 1:

Here is a thought process that I would like the candidate to refute during the defense. Suppose that a protein sequence starting in one point in genotype space (so just a single sequence) can only evolve one amino acid at a time. When one substitution happens then another one opens up for evolution, etc. Essentially, between two points in sequence space, A (the starting point) and B (some distant point) there is only a single path to traverse. Suppose then while this path is being traversed, the sequence randomly duplicates and the duplicates evolve independently (these duplications would be speciation events). If synonymous sites in this sequence can do what they will, wouldn't one see the same access of parallel nonsynonymous substitutions as has been observed in amphipods?

Answer:

The scenario proposed by the Reviewer is the extreme case of site-specific negative selection: at each time point, just a single amino acid substitution is permitted. This scenario will of course lead to an increase in the level of parallelism (relative to divergent substitutions) at nonsynonymous sites, compared to synonymous sites, similar to that observed in Fig. 3.1B-C. The rate at which the (only permitted) nonsynonymous substitution occurs will then depend on selection in its favor. If the substitution into it is neutral (e.g., in the neutral network-type fitness landscapes) or deleterious, it will result in $P \leq 1$ (see also response to Prof. Hahn's comment 5). $P > 1$ has to imply that this substitution is advantageous, i.e., that the observed parallelism is adaptive.

Comment 2:

Methods/Concepts question: was the phylogeny reconstructed using nucleotide sequences or amino acid sequences? If the latter, then perhaps synonymous sites created hemiplasies of amino acids. It would me good to see that the phylogenies are not dependent on whether nucleotides or amino acids are used for phylogeny reconstruction.

Answer:

Only nucleotide sequences were used, because short internal branches make amino acid reconstruction quite inappropriate. I agree that amino acid tree would merge linages with convergent evolution, as it happens in Castoe *et al*. (2009) study. In the revised version, the sensitivity of our analysis to phylogenetic reconstruction is considered in more detail; in particular, we show that our key result ($P > 1$) holds for the subset of gammarid lineages for which this reconstruction is unambiguous.

Minor issues:

Start of Section 3.1. I do not think that the candidate really means this here: "a substitution between a pair of species can only be neutral, and if the fitness landscape is invariant, an identical substitution at this site between other two species is also expected to occur at the neutral rate." I think what the candidate means is that if selection is invariant.

This phrase starts with "Under purely neutral evolution, a site that has experienced a substitution between a pair of species can only be neutral". This is correct as stated, by definition of purely neutral evolution. The distinction made by the Reviewer is unclear: seems like "fitness landscape" and "selection" is the same in this context?

Section 1.2, implications of the work, revealed Russian-thesis-style modesty. Starting it with methods is rather self-deprecating. I think that the work brings more to our knowledge than the candidate revealed there.

We have now expanded this section, elaborating on our main result – the large extent of genome-wide parallelism.

The thesis was well-written, but in parts was not grammatically or stylistically perfect, or native-speaker level. This is something to work on in the future.

==========

Reviewer: Dmitry Ivankov

In the presented PhD thesis "Positive selection in parallel evolution" Valentina Burskaia analyzed identical nonsynonymous substitutions in different species and adaptation as a possible reason of the identified substitutions.

The thesis consists of two parts.

First, Valentina analyzed repeated non-synonymous substitutions in different groups of species. For the analysis, she chose the following three groups: (i) lake Baikal amphipods, (ii) lake Malawi cichlids, and (iii) vertebrates. Valentina found that amphipods demonstrate high rate of parallel non-synonymous substitutions, significantly higher than expected. The effect demonstrated by the cichlids was not so prominent while the rate showed by the vertebrate group was lower than expected. Valentina reasonably accounted the found results to adaptations in the amphipods and negative selection in the vertebrates.

In the second part, Valentina tests a reasonable hypothesis that birds experiencing lower level of oxygen should demonstrate adaptations in the genes responsible for breathing. The lower oxygen level could arise due to different reasons like flying at high altitudes or diving.

The results of the presented work are scientifically significant and comply with the international level and current state of the art. The work is perspective for future applications and fundamental research. The publication is of high quality, the number of publications suits the requirements for the PhD thesis.

The dissertation conforms to high international standards. It has a clear structure; the topic corresponds to the actual content.

The remarks that I have concern only the text of the thesis. For example, the first reference to Fig. 3 is located in the page 27 of the thesis while Fig. 3 itself is situated in the page 36, i.e., nine pages later, which is not convenient for readers. In the page 28, Valentina writes "…we used our approach developed previously (Bazykin et al. 2007)." However, Valentina is not a co-author of the 2007 paper, and, therefore, the use of the word "our" is not relevant here (contrary to the paper Burskaia et al., 2020, where the use of the word "our" is completely acceptable since Bazykin is one of the co-authors). There are also a few number of misspellings; the corresponding list is sent to the dissertant.

I agree that approach, described as P test was invented in Bazykin et al. 2007, I only implemented it. Now corrected.
Thank you for all the numerous corrections of the text: it is now fixed.

To summarize, I rate the PhD thesis of Valentina Burskaia as very important, of high quality and scientifically significant.

==========

Reviewer: Konstantin Popadin

I was enjoying a lot by reading the PhD work by Valentina Burskaia: very deep and detail introduction, clear definitions of all terms, main problems and approaches; interesting results and detailed discussions. The dissertation is very well written with a straightforward logical flow.
The topic of the dissertation is relevant to its content, methods are adequate and fresh. Results of the work significantly improve the understanding of the scale and targets of parallel evolution. High quality of scientific research is supported by two first-author publications: one is in GBE (doi: 10.1093/gbe/evaa138) and one is uploaded to BioRxiv (doi: https://doi.org/10.1101/2020.09.15.298117).

Without any doubts, I recommend that the candidate should defend the thesis by means of a formal thesis defence.

I have two comments related to chapters 3 and 4 and would be glad to know the opinion of Valentina.

Comment 1:

Chapter 3.
Intensive whole-genome parallel evolution in amphipods is very interesting. I would like to discuss the possibility of the parallel relaxation of selection in different species, which can at least partially contribute to the expected results. Taking into account that the main mode of selection is purifying and the diversification of species may be associated with a parallel decrease in Ne (as compared to the common ancestor) it is possible to assume, that pronounced decrease in Ne in some but not in other species of the quartets may drive the parallel molecular changes - accumulation of suboptimal slightly-deleterious substitutions. The main problem here is that it is difficult to approximate Ne and it is possible to use either some life-history traits, for example, the dimensionality of the environment (2D or

3D) or resource-poor/rich environments (see example here doi 10.1101/gr.212589.116) or genetic correlates of Ne - for example genome size (doi 10.1101/gr.212589.116: longer genomes on average mark less effective selection due to low Ne). As I see from chapter 3 authors used in analyzes many interesting traits, but none of them approximate Ne.

I would like to know the opinion of Valentina on the suggested scenario - which results can vote for or against the proposed scenario?

Answer:

I think that negative selection relaxation can never cause P test values higher than one. See also the response to Prof. Hahn's comment 5. Now the explanation is also added to the text, at the end of 3.2.2. section "Calculation of P statistic for a species quartet".

Given that, I believe that effect of high population size could rather lead to growth of probability of independent adaptive mutations appearance, as it happens during experimental evolution of bacterial populations.

Comment 2:

Chapter 4.

Mitochondrial genome of all vertebrates is extremely highly constrained and the NULL hypothesis is that the purifying selection is the only mode of selection, acting on mtDNA of vertebrates. The vast majority of studies, claiming positive selection in mtDNA of vertebrates is questionable. Additionally, birds are among the most constrained vertebrates because of high temperature and correspondingly high level of aerobic metabolism. From these points of view - the attempt to search for events of parallel adaptive evolution in mtDNA of birds is a low-success task from the very beginning. And, despite this, Valentina was able to find several candidate sites evolving in parallel. From my point of view, this is the amazing and unexpected result (I would expect zero sites evolving in bird mtDNA under parallel positive selection).

Another peculiarity of mtDNA and especially mtDNA of birds is very strong nonuniformity in mutagenesis, which, according to my opinion, may affect results/interpretation of the results. Two of the most common mtDNA substitutions, totally explaining ~ 80% of the whole mutational spectrum in mtDNA, are C>T and A>G substitutions on heavy chain (deamination of cytosine and deamination of adenine on the heavy chain, which spends a long time in single-stranded mode during mtDNA replication). Moreover, the fraction of A>G has been recently associated with oxidative damage in organisms with a high level of aerobic metabolism (doi: 10.1101/2020.07.25.221184). A>G substitution on heavy chain corresponds to T>C substitution on the light chain which is a reference strand in all public mtDNA databases. Thus, from mutagenesis point of view, we expect a strong mutational bias towards T>C (light chain notation) in species with high oxidative damage and no bias ("T>C" ~ "C>T" or even "T>C" < "C>T") in species with low oxidative damage. Assuming that environmental hypoxia (in high-altitude birds) may be associated with cellular level hypoxia we can expect that these species will have decreased frequency of mutations driven by oxidative damage: i.e. decreased fraction of T>C on the light chain, which automatically means an increased fraction of C>T. Altogether, we expect higher C>T fraction in high-altitude birds versus low-altitude birds.

One of the strongest results is based on four Histidin to Tyrosine substitutions in position 57

of the ND5 gene, which are going in parallel with switches from background to a high-altitude environment. The one-step trajectory from His to Tyr is possible due to C>T substitutions in the first codon position. This is exactly what we expect if the pressure of oxidative damage is decreasing in these species and the probability of C>T substitution is increasing.
So, one additional interpretation of the results from chapter 4 is that changes in the environment (high-altitude => hypoxia => decreased fraction of A>G on heavy chain == decreased fraction of T>C on light chain == increased fraction of C>T "which is opposite") may change rules of mutagenesis, allowing new categories of substitutions (C>T) becomes relatively more common. So, this is yet a parallel evolution, however, most likely due to neutral changes associated with mutagenesis.

It is not a very common situation when mutational spectrum drives amino-acid changes, however, it has been shown recently on SARS-CoV-2 evolution (doi: 10.1101/2020.05.01.072330) and also it has been recently shown for human mtDNA (unpublished data from my Kaliningrad laboratory).

I would greatly appreciate having comments of Valentina to the above-described scenario. More precisely the question might be rephrased like this: if TreeWAS takes into account different types of nucleotide substitutions during simulation of the "null" distribution?

Answer:

I agree that parallel substitutions in bird mitochondrial genomes most likely are not adaptive and could be caused by neutral changes associated with specific mutagenesis.

I also agree that the main problem of TreeWAS approach is absence of adequate evolutionary model: it does not account for different mutation types and for typical gamma distribution of mutation rate variation. The method can be significantly upgraded in that area. However, generation of "null" distribution for each particular site will cause other problems. Particularly, the form and parameters of that distributions are subject of many discussions.

Valentina, thank you for the interesting work!