Skoltech

Skolkovo Institute of Science and Technology

Thesis Changes Log

Name of Candidate: Aleksandra Bezmenova

PhD Program: Life Sciences

Title of Thesis: Evolutionary processes in hypervariable fungus *Schizophyllum commune*

Supervisor: Prof. Georgii Bazykin

The thesis document includes the following changes in answer to the external review process.

Prof. Gelfand

Some of negative results described in Ch. 4 and 5 are not reflected in the Conclusions section; I think they are interesting enough (even if not final) to warrant being mentioned.

I have now reflected results of Chapter 5 and negative results of Chapter 4 in the Conclusions section.

Prof. Anderson

…I made a series of comments by means of strikethroughs, sticky notes, and insertions in the PDF file; of course, it is up to the candidate and supervisors about how to address each suggestion, or not.

Thank you for the meticulous editing, I have revised the manuscript accordingly.

"Thus, one can hypothesize that the mutation rate in S. commune may be elevated given less heterozygous genome segments." Are you suggesting this may be due to the ineficiency of homologous repair process in the divergent regions? Try to be more specific here.

I assume that if recombination rate is higher in more conserved regions, and recombination itself may cause locally elevated mutation rate, than the mutation rate in more conserved regions may be higher.

"Some of these offsprings were back crossed with parents, and one crossing that produced sufficient amount of offsprings was selected for further analysis." Was this due to spore inviability in the others?

Mostly yes.

"I hypothesize that extensive vegetative growth and lack of mechanisms of preserving the genetic material during this growth…" Do you mean sclerotia or other hardedn structures enabling perennation?

No, I rather mean some mechanisms that preserve the accumulation of mutations during the growth.

I found the Introduction and literature review to be OK. The essential topics were covered, but not much more. I especially thought that the candidate could have gone deeper into the background on recombination (more on this below) and she could have been more specific and insightful about how her contributions fit into the broader picture.

In setting the stage for this part, I also thought that there should be more background context on recombination with different levels of nucleotide sequence diverge. What about the yeast (Saccharomyces) example? With high levels of divergence, as in inter-species yeast hybrids, the mismatch repair machinery goes into overdrive and recombination is entirely blocked, leading to mis – segregation of entire chromosomes, resulting in widespread inviability of offspring. Indeed, this postzygotic mechanism is the main means of reproduction isolation between species. Might something similar be operating in S. commune in the highly divergent regions of the genome and less so in the regions with low divergence? It seems that the introduction should not ignore the history in Saccharomyces.

This is now discussed in Chapter 2 and in the discussion in Chapter 6.

In the thesis, the candidate implies that the haploid monkaryotic phase of S. commune is as prevalent in nature as the dikaryotic phase (thesis page 58, Chapter 4). This is not the case. Monokaryons tend to be fertilized rapidly by spores or hyphae and so the monokaryotic phase is extremely ephemeral in nature. (As an illustration, take a monokaryotic mycelium in a petri dish into the outdoors or indoors, remove the lid for 24 hours, and then sample the resident mycelium. It will by then have become dikaryotized by spores in the air!). Haploid monokaryons are extremely avid to take on fertilizing elements.

I have now corrected the wording and do not state that monokaryons are as prevalent as dikaryons. However, our results from Chapter 4 show that sometimes monokaryons can occupy territory.

The two-fold difference in rate actually might actually some sense. The monokaryons have one candidate nuclear type for mutation, while the dikaryon has two.

The substitution accumulation rate estimated in Chapter 4 was normalized by the diploid genome length, thus it should stay in line with our estimates from Chapter 3; however, as was pointed out by another reviewer, it is perfectly in line with our estimate for the Russian mycelium from Chapter 3, for which the rate was higher than for the USA mycelia (samples from Chapter 4 were also collected in Russia). This is now reflected in the manuscript.

My other point concerns the idea that some filamentous organisms might show a decreasing rate of mutation the longer they grow vegetatively. I see no mechanistic reason why this should be the case under conditions of steady-state growth. And I interpret the Armillaria data (page 54) very differently than the candidate did. The more likely explanation for the data in Fig. 16 of the thesis is that the individual of Armillaria spread rapidly after its birth to fill up its present environment and then basically "ran in place" over the years, exploiting new food sources as they became available in the immediate locality. This means that isolates that were collected from points relatively close together in space may have almost as many cell divisions separating them from their common ancestor cell as isolates taken from points much further away from one another. In other words, the short distances drastically under- represent the actual number of cell divisions. The phylogeny of changes in (Fig. 2 in Anderson 2018) is consistent with this possibility – the terminal branches are all long compared to the internal, phylogenetically informative changes.

This is undoubtedly true. However, all my statements were in regard to the mutation accumulation, not mutation emergence rate (I have now clearly stated that in the Fig.16 description). And the scenario described by the reviewer does not contradict my statements: it's just one of the explanations why the mutation accumulation rate may decrease with distance.

Chapter 5. I am not sure why it is not possible to filter out the mutations happening during the growth of the dikaryon before fruitbody formation and then for the analysis to proceed with the mutations that were unique. Of course, it's best to minimize propagation of the dikaryon before fruitbody formation. But this growth phase is impossible to eliminate entirely. Even within the fruitbody, there are mitotic divisions of the dikaryon that precede basidium formation and meiosis.

Because it is impossible to determine whether all the unique mutations appeared after the formation of the fruit body.

Conclusions Chapter. The end came rather abruptly. I've seen this many times in Ph.D. theses. The candidate has worked hard on experiments and writing and is maybe tired at this point. But please, dig a little deeper and give us more to go on for the future research in this area.

I have expanded the Conclusions section.

Prof. Agrawal

On p. 44 it says: "At these positions, we called variants that had the following properties: (i) at least in one sample, coverage in the 10-90% range and non-reference variant frequency >30%, or coverage in the 15-85% range and non-reference variant frequency >20% (13962 variants); (ii) not supported by any read in the reference sequence (289 variants). For these variants, we assessed their frequencies in all samples." It is unclear to me what the numbers in parentheses indicate (i.e., 13962 variants and 289 variants). It sounds like 13962 variants are identified at one stage but in the results there are only 289. Does this mean that 98% of the variants are excluded because of reads in the reference sequence?

Yes.

Have you considered how this issue will affect your estimate of the rate? In other words, if a TRUE mutation occurs at a random site, what is the chance you would then exclude that variant because that site would have a supporting read in the reference sequence. (By your procedure many sites may not be fully callable because there is an errant read supporting an alternative base.)

Given the estimate of the HighSeq 2000 error rate (0.3%) and the mean sequencing depth (135x), the probability of an error read supporting the mutation variant is 1-(1-0.003/3)^135 = 12%. Thus, we have probably lost about 1500 substitutions. This number is quite huge, and if considered may can increase our estimate of the mutation rate several times. However, we can also make the opposite mistake – not sequencing a very low-frequency ancestral variant. Overall, our results are probably underestimated, but this can only further emphasize out conclusions.

I would like to have seen an estimate of the "population size" (i.e., number of hyphal strands) in narrow and wide tubes.

Unfortunately, we only have very uncertain assumptions of the following orders of magnitude for the "population sizes": $10^1 – 10^2$ in narrow tubes; 10^4 and higher in thick tubes.

How important are the number divisions when grown in liquid culture for sequencing?

The growth length in liquid media was 1-2 cm at most, so not so important.

Table 3.2 is very useful but I would also like to see the what proportion of callable sites are in each category so it would be easy to see if nonsynonymous mutations are underrepresented.

We calculated dN/dS ratio for that; nonsynonymous mutations were underrepresented in thick tubes, but not biased in narrow tubes.

The estimated mutation rate is described as similar to that observed in Chapter 3. However, the one Russian sample in Chapter 3 had a much higher mutation rate and the estimates of these Russian samples in Chapter 4. This should be acknowledged.

Thank you, this is now acknowledged!

Much effort goes into measuring the former yet there is massive (and unknown) certainty in the cell divisions per generation.

This is indeed true. However, I would like to raise two points. First, in Chapter 4 we sequence fruit bodies, thus, we obtain the sequences at the state immediately prior to the meiosis and the formation of the offsprings. Second, our somatic mutation estimates are in line with the generational rates obtained in previous studies (Baranova et al 2015) – we see that most likely the somatic mutations are the main source of generational mutations and the distance between meioses determine the generational mutation rate. We discuss it in Chapter 3.

If "generation" is defined as sexual events, then you might be able to gain insight by relating mutation rate to recombination rate by comparing theta = 4 Ne*u to rho = 4 Ne*r.

Thank you, this is an area for further research.

Table 5.1 reports the "Callable length" but the text below the table says it is "impossible to estimate the callable length of the genome", which seems to suggest that Table 5.1 has accomplished the impossible.

I have corrected the wording: **"…**it was impossible to estimate the callable length of the genomes that were targets at the time when the mutational events happened."

Can you simply use a single sample per fruiting body to estimate the mutation rates (averaging over the outcome of using different samples)? This should allow you to avoid the problem of "cluster" mutations.

I suppose I cannot because I do not know the number of target genomes at the time when the mutational events happened.

**This chapter reports a very striking result – a mutation rate 2-3X higher in homozygous than heterozygous regions. However, I was disappointed by the lack of discussion of this result. What are plausible explanations? Presumably, most of these mutations occur while it is growing as a dikaryon so the homologous chromosomes are isolated in different nuclei, correct? Is there any bioinformatic (i.e., technical artifact) reason that could lead to this pattern. In particular, I wonder about this step of variant calling (p. 69): "ii) not a
single read in both parental mappings supported non-reference nucleotide in case of 'both'
genotype, or not a single read supported non-reference nucleotide in parent 1(2) in**

case of 'parent 1(2)' genotype." Could that filtering step make it more likely to exclude variants from heterozygous regions than homozygous regions?

I believe the answer is connected to the previous answer – from this data we do not actually know the *de novo* mutation rates.

This is a clever experiment. However, I have one serious concern. There is a very striking difference in recombination between your two crosses. Those crosses differ in heterozygosity of the focal region but also in genotype at many parts of the genome. With these data alone, it is seems impossible to make a strong inference that the difference in recombination is due to heterozygosity or just genetic background. If you think heterozygosity is directly affecting recombination, can you speculate on mechanism?

Unfortunately, we indeed cannot exclude the effect of the genetic background. However, we can speculate that the rate of the recombination is affected by the MMR system that suppresses recombination given mismatches. This is now pointed out in Chapter 6.

<u>Prof. James</u>

The candidate should review all chapters and correct typographic errors and improve the language as appropriate.

This has been done.

Please have a close look at the methods of analysis and make sure that they are crystal clear. I think this is something that we can discuss as a committee. For example particular methods of mapping will greatly influence the outcome. There were times when what was done was unclear to me: "non-reference variant frequency in reference sample <= 20%; pg 60).

I have corrected the wording.

<u>Prof. Ivankov</u>

The typos and language were corrected according to the additional file with comments provided by the reviewer.