Skoltech
Skolkovo Institute of Science and Technology

Thesis Changes Log

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

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.

Prof. Molly Przeworski:

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).

In the revision, I expand the introduction in Chapter 3 to describe the current knowledge on mutation rates in *S. commune*, based on the papers (Bezmenova *et al*, MBE 2020) and (Baranova *et al*, MBE 2015). Genetic distances between *S. commune* samples are represented by the reconstructed phylogenetic tree on Figure 3.2. As noted by B. Charlesworth, this is not the true phylogeny since there is recombination within the sampled populations, however, it can give sense of the general population structure and genetic distance between populations. Consistent with the previous paper on the genetic diversity of *S. commune* (Baranova *et al*, MBE 2015), the star-like shape of the clades corresponding to the populations shows almost no structure within them — genetic distance between any two genotypes from the same population is approximately the same. Fst between the Russian and USA populations is approximately 0.58 (pi = 0.34), indicating their early divergence. However, fungi from different populations are still able to mate, as was demonstrated in (Seplyarskiy *et al*, MBE 2014).

Two Florida samples (FL and s1514) form a clade external to the other USA samples, while being still the part of the USA population (Fst = 0.11) — they were excluded from further analysis based on this observation, and not on geography. I clarify it in the revised text.

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.

I expand Figure 3.7, previously having showed LD between pairs of polymorphisms with MAF < 5%, with figures showing LD between pairs of polymorphisms stratified by allele frequencies in *S. commune*, *D. melanogaster* and *H. sapiens* (Figures 3.7 – 3.9).

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).

As discussed in the previous results section, the excess of LD_{nonsyn} over LD_{syn} in *S. commune* holds under different minor allele frequency thresholds (now shown on Figure 3.7), although it is more pronounced under high MAFs. The haploblocks are annotated by high LD between the polymorphisms within them. Indeed, with MAF being correlated with LD, higher MAFs within haploblocks can be confounded by this criteria. However, as we show on Figure 3.13a (Figure 3.10a in the old version of the thesis), the crucial distinction between haploblock and non-haploblock regions is the presence of two major haplotypes, which, as we hypothesise, may be maintained in the population by balancing selection. High MAF in the haploblocks arises from high frequency of the haplotypes, and therefore the polymorphisms comprising these haplotypes.

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?

Indeed, Hill-Robertson interference may cause positive correlation between LD and pn/ps — we also reproduced such correlation in non-epistatic simulations. However, as we suppose, it cannot explain positive correlation between LD_{nonsyn}-LD_{syn} and pn/ps observed in the data (Figure 3.13e in the revised text), and the overall excess of LD_{nonsyn}.

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?

Considering the recombination map of *S. commune*, there is a paper describing the patterns of recombination in 17 F1 hybrids of a pair of individuals sampled from USA and Russia (Seplyarskiy *et al*, MBE 2014). With such exceptionally high level of genetic diversity, it's possible to annotate cross-over events with high precision. Recombination events were shown to be more frequent in the regions of low diversity, including exons; it can be indeed caused by the fact that homological recombination can't occur if the parents genotypes are too different.

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.

In the discussed results section, we focus on the correlations between LD values between the same pairs of shared polymorphisms found in two populations. We suppose that if synonymous polymorphisms are on average older, they have higher probability to emerge before divergence of the populations. In such case, the correlation of LDs between such pairs can be confounded by their common decent (while LDs by themselves are expected to be, indeed, on average lower).

Unfortunately, stratifying shared SNPs by MAF together with other filtering criteria makes the datasets too small to analyse with the observed level of noise. I believe, it'll be possible if we increase the sample size by sequencing more *S. commune* individuals from two populations.

However, the main conclusions in this section come not from comparing synonymous and nonsynonymous SNPs, but from comparing pairs of SNPs within the same gene and from different genes, and nonsynonymous polymorphisms leading to parallel to different amino acid changes.

On a minor note, I thought a few details were missing from the Methods, like the sequencing coverage after read mapping.

In the revised text, I added the corresponding Appendix Table showing assembly statistics (including the average coverage) for the *S. commune* genomes.

<u>Prof. Molly Przeworski:</u>

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).

<u>Prof. Brian Charlesworth:</u>

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).

The haploblocks are indeed likely to emerge only in genomic regions with low recombination rate, since recombination will break and re-shuffle the haplotypes. However, we think that low recombination alone can't explain the existence of the haploblocks of such strength and abundance like the ones we observe in the data. In the simulations in the absence of epistasis and balancing selection, we weren't able to reproduce high values of LD observed within haploblocks even if the recombination rate is low (and even zero) — it was possible only in simulations under balancing selection. In the revised text, I add the panel showing the average LD produced in simulations with and without balancing selection to Figure 3.13 (old Figure 3.11), and add the corresponding comments to 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.

Thank you for these comments. In the revised text, I added short abstracts at the beginning of Chapters 3-5 and expanded the literature review according to the comments. Considering including of a short conclusion of the review, I tried to provide such summary in regard to the goal of the thesis in the Introduction section.

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).

Thank you for the detailed comments. Additional references were included in the revised text, minor corrections were fixed, and corresponding comments were added. The "evolutionary Stoke's shift" (here and in the comments to Chapter 5) is another term for entrenchment, or the increase of the allele's fitness with time, firstly introduced in (Pollock *et al*, PNAS 2012). It was indeed inspired by the term used to describe shift between absorption and emission energy in spectroscopy.

Chapter 3

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.*

In the revised text, I extended the introduction in Chapter 3 with general information about *S. commune* genome, estimation of its mutation and recombination rates. I believe genetic maps of other fungi (such as *Pleurotus tuollersis*) should be used very carefully, since *S. commune* is shown to have unique recombination patterns: in contrast to less polymorphic species, crossing-over events in *S. commune* are more frequently observed in relatively less diverse genomic regions (presumably, because homological recombination is impeded if the parental genotypes are too different), e.g. coding regions (Seplyarskiy *et*

al, MBE 2014). Nevertheless, general features of recombination such as centromeric suppression and the existence of recombination hot- and cold-spots should indeed be present in *S. commune*.

p.75 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?

I added the plots showing polarised LD between nonsynonymous and synonymous pairs of SNPs (Figure 3.7d,f or the revised text), and elaborated the discussion on the factors potentially leading to the observed excess of positive LD between nonsynonymous polymorphisms.

We also simulated ongoing selective sweeps, and didn't observe the excess of LDnonsyn over LDsyn in these simulations (these results weren't included in the thesis text). Additionally, the observed regions of high LD (haploblocks) with the largest excess of LDnonsyn are unlikely to be caused by sweeps: they aren't characterised by low genetic diversity, even within two major haplotypes constituting the haploblock.

p.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.

In the revised text, I rewrote the discussion on different patterns of LD observed between polymorphisms with low and high allele frequencies, and elaborated on how Hill-Roberson interference can affect these patterns (incl. Figure 3.10, showing simulations of HRI).

p.80 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.

Thank you for this important comment, surely, recombination rate in *S. commune* varies along the genome (as was discussed above), presumably including centromeric and telomeric suppression, low recombination in the mating-type controlling loci and other hot- and cold-spots of recombination. We indeed expect to observe regions of high LD in the genomic regions with low recombination rate. However, the observed haploblocks occupy ~10% of the genome, are short (typically < 1000 nt) and are distributed more or less uniformly along the genome, so it's unlikely they are caused by the suppression of recombination in the mating-type controlling loci. I also presume that centromeric and telomeric regions are underrepresented in our analysis since they are harder to assemble de novo and will correspond to the regions of poor alignment - however, we didn't check it specifically.

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).

Thank you for this important comment. We inspected a fraction of the most pronounced haploblocks manually, and found no evidence of them to be the result of inversions. Additionally, as shown on Figure 3.13 (Figure 3.9 in the old version of the text), the boundaries of the haploblocks are usually not distinct: recombination gradually breaks LD on the edges of the haploblocks; we can even see smaller haploblocks with high LD "nested" within a larger haploblock of not so high LD. The observed abundant haploblocks, with two major haplotypes with substantial levels of diversity both between and within haplotypes, should

in any case be maintained by some kind of balancing selection (e.g. heterozygotes advantage), even if they may be caused by inversions.

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.

p. 78 §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.

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

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

p.82 §3 l.1 '*pn/ps*' is not defined, nor it explained how it's been estimated. Also, the rationale for looking at this should be explained: why should higher *pn/ps* 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.

In the revised text, I added corresponding comments and fixed minor errors.

Chapter 4

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.

Thank you for the comment, indeed, dS for the genes experienced bursts on the corresponding branch is typically very low (Table 4.1, 0.5-2 synonymous substitutions per gene). On the studied phylogenies of close species, containing internal edges shorter than 0.005 dS (calculated using concatenated sequences of all genes), this is consistent with the expected number of synonymous substitutions per gene. Therefore, we conclude that the high dN/dS of the bursts isn't caused by low dS, but by high dN (Table 4.1, 6-39 nonsynonymous substitutions). Figure 4.3 represents that the exceptionally high dN/dS on the burst-containing branch doesn't expand to the neighbouring branches.

p.106 §2 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 $T \approx 2\ln(4Nes + 0.5772)/s$ 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: $8Ne\mu \ln(4Nes + 0.5772)$. This is what is relevant to HRI. It is only weakly dependent on the selection coefficient, and is only larger than $8Ne\mu$ 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, $8Ne\mu$ should be 267 x 0.01 = 2.7, if $4Ne\mu = 0.01$, so there is some scope for HRI.
The conclusion about HRI thus probably needs modifying.

Thank you for this important comment. I agree that in order to estimate the possible affect of HRI, we should consider the average number of positively selected mutations segregating simultaneously. If I understood the comment correctly, we should calculate mutation rate considering the sites constituting the burst (from 6 to 39 in our results), which makes the estimate for $8Ne\mu < 0.39$. I would be glad to discuss this question during the defense.

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.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 105 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).

Thank you for the detailed comments. In the revised text, I added corresponding comments and fixed minor errors.

Chapter 5

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?

Out method was tested to be robust to the differences in the overall substitution rate between sites and alleles — it generally compares the rate of substitutions of the selected allele on the branch with the average substitution rate on this branch, and shows whether this rate is significantly increasing or decreasing with time since the origin of the allele (which is different for different alleles/sites). Surely, the inference of ω in our datasets is far from perfect: above all, there are typically too few substitutions (particularly nonsynonymous) per site to accurately estimate dN/dS. We used estimation of ω in order to roughly stratify all sites based on their substitution rate; this is indeed a stretch to say that all sites with estimated $\omega < 1$ are under negative selection.

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.110 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

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'.

I fixed the listed errors and added corresponding comments to the thesis text.

<u>**Prof. Dmitrii Ivankov:**</u>

Page 17: Terms 'smooth function' and 'simple function' should be commented, at least. Better, they should be defined and commented.

The mentioned terms are, indeed, used without explanation. The idea of using them was to emphasise on the distinction between unidimensional and multidimensional epistasis, as in (Sailer and Harms, Genetics 2017). In the revised text, I add comments on what is meant under "simple function" and that usually the monotonic functions are used for this purpose; "smooth" here isn't related to the mathematical definition of the term and was removed from the text to avoid confusion.

On page 72, Anastasia writes "The potency of any kind of selection increases with the amount of variation; for epistatic selection, however, this increase is expected to be faster than linear, because it depends on the number of possible allele combinations." On page 34, Anastasia writes "At the same time, epistasis reduces the set of available evolutionary pathways ..." These two factors seem to have opposite influence on the rate of evolution. However, on page 72 Anastasia mentions only one factor and does not discuss the other. The influence of reduced set of evolutionary pathways on the rate of evolution (or its irrelevance) should be commented.

In the two mentioned paragraphs, the causal relationships between epistasis and the level of genetic variation (or the number of available evolutionary paths) are considered in the opposite directions and on different evolutionary scales. On page 34 of the old version of the thesis, general features of fitness landscapes and selective constraints caused by epistatic interactions between any possible (but not necessarily present in the population evolving on such landscape) variants are described. On page 72, I consider genetic variation within natural populations and the potency of selection acting on the segregating polymorphisms. In this case, the capacity of such selection indeed depends on the genetic diversity within a population, which is generally defined by its mutation rate and population size. If the neutral level of genetic variation is low, the efficiency of selection will also be low since it won't be able to eliminate deleterious variants and promote beneficial variants if they are absent. Certainly, epistatic constraints can shape patterns of genetic variation within populations, similarly to negative selection reducing variability, this is what we see in the highly polymorphic populations of *S. commune*. On the macroevolutionary scale, however, the evolutionary constraints between substitutions fixing between diverging populations may still be complex. In the revised text, I added the corresponding comments in Chapter 3.