Skoltech
Skolkovo Institute of Science and Technology

Thesis Changes Log

Name of Candidate: Kseniia Safina

PhD Program: Life Sciences

Title of Thesis: MOLECULAR EPIDEMIOLOGY OF SOCIALLY IMPORTANT INFECTIOUS DISEASES

Supervisor: Prof. Georgii Bazykin

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

I would like to thank the reviewers for their comments and suggestions. Below are my point-by-point replies.

Prof. Kühnert

Summary of issues to be addressed before/during the thesis defense:

- *A major issue to be addressed before the thesis defense is in chapter 3 (p.57) is a "re-use" of the data set. The BEAST analyses should not be run using a prior distribution on the evolutionary rate that was obtained from the same data set. Please re-run all affected analyses with a prior obtained from a different data set (from the literature). Fortunately, this is unlikely to cause major qualitative differences to the results, as all analyses were performed with the same prior. However, quantitative differences are expected and the results should be updated accordingly.*

Thank you, this was indeed inaccurate! I've re-run all BEAST analyses with ulcd.mean priors based on the estimates obtained in [1]. In fact, the median rate for subtype A remained the same (0.0015). I've re-run the logistic growth and the BDSKY analyses for CRF63 using the estimate from [1] obtained for CRF02 (its ancestral recombinant strain; I could not find independent estimates for the relatively young CRF63 variant). The results indeed did not change noticeably. Figure 3.8, ulcd.mean in Supplementary Table A-2, and Methods (see 3.2.7 and 3.2.9) are now updated.

Minor comments:

- *I am somewhat surprised that publications 2 and 3 as listed on page 5 are not mentioned anywhere else in the thesis. I would find it useful to relate them to the rest of the work, if only in the abstract.*

I'm sorry for the confusion - PhD program at Skoltech requires two publications with at least one of those related to the thesis topic. Publications 2 and 3 were not meant to be part of the thesis, I was just asked to provide a complete list of publications.

- *Pages 57/58: multitree is not the first method to allow for this kind of analysis, although it may be more convenient than previous approaches (Novitsky et al 2015, Epidemics, Kühnert et al 2018, PLoS Path.). Please rephrase.*

I'm sorry I overlooked that! I tried some naive XML editing of ChangeTimes in BDSKY and that did not work. I've now rephrased it more neutrally (pages 59-60).

- *Please explain why the sampling proportion inferred from data set I is suited for data set II (p.58).*

As acknowledged in the Discussion (the second limitation), we consider our estimates produced for the Dataset I more reliable. I could not come up with a way to construct a proper informative expectation of the sampling proportion for the full Dataset II that carried old infections (though even when considering old infections, sampling was biased towards later years) and ended up using the prior from the Dataset I. I understand that obtaining similar R_e estimates for the two datasets doesn't indicate that the prior is suited for the Dataset II. I tried to analyze the Dataset II with the rate of becoming uninfectious fixed to estimate a three-dimensional sampling proportion together with R_e; it estimated that the sampling proportion was the highest before 2010, which is not realistic I suppose (though R_e was estimated to be 3.2).
The same limitation unfortunately applies for the BDSKY analysis that compares R_es of two separate clades, as those come from the Dataset II and not from the Dataset I; yet, the results agree with the results produced by the coalescent model which is not informed by the sampling proportion prior.

- *Why was R_e assumed to be constant? A 3-4 interval approach may have captured interesting transmission dynamics through time (p.67). Please justify.*

I have now added Supplementary Figure A-17 that shows the dynamics of R_e computed for two different interval sets; I kept constant R_e estimates in the main text as I did not find time-varying estimates informative/reliable enough. Allowing R_e to change every 5 years starting from 2005 did not show any difference in the 5-year dynamics of HIV-1 (though we have only a few lineages spanning that far back in time). Allowing R_e to change simultaneously with sampling proportion showed a moderate increase of median R_e in 2018-2019 associated with a substantial uncertainty around these estimates; this might be a hint of a growing epidemic, although this possible growth is not captured by the 5-year dynamics.

- *Supp. Fig A-9 ranges until year 2030, which may be due to a mistake in the plotting script.*

Fixed.

- *Please note that there is an option to not assume sampling to lead to becoming non-infectious using the sampled ancestors approach (Gavryushkina et al 2014, PLoS Comp. Biol.). Please remove this statement or rewrite accordingly (p.81).*

Yes, I am aware of the sampled ancestor approach! I've now specified in the text that this assumption refers specifically to the analyses we conducted, not to any birth-death models. It would be interesting to analyse multiple clades simultaneously allowing for sampled ancestors, but as I understand, this is not available in the multitree implementation that we used (at least I could not find SA operators there and did not come up with a way to add the SA package).

- *For SARS-CoV-2, page 91, why were problematic sites not masked (see e.g. de Maio et al 2020, Virological.org).*

As far as I can remember, we were already aware of this list when we prepared the paper, but the list did not seem very stable back then, and it did not affect sites that represented viral diversity in Russia early in the pandemic, so we only cropped alignment ends.

- *Why were time-varying R_e estimates obtained from EpiEStim instead of bdsky? Please justify and discuss.*

Because viral diversity in Vreden was quite low, we tried to make our model as easy as possible and only included pre- and post-lockdown R_es to check the potential difference in R_e before and after

lockdown. The EpiEstim analysis served as an independent validation which was based on another piece of data (data on patients who required a specialized treatment).

Prof. van de Vijver

- *Section 2.1.2 on modern molecular epidemiology*

Although, the theoretical background in the thesis is of a very high quality, I found section 2.1.2 a bit difficult to follow. In this section a lot of modern concepts are not very well described. For instance, the part on the Ebola outbreak of 2014-16 only lists what insights molecular epidemiology provided without giving further details. In my view, this section can also be improved by using sub-headings.

I have now added sub-heading by splitting the section into five pieces. I have also made sections 2.1.2.4-2.1.2.5 a bit more detailed.

- *Page 36 on HIV transmission in the United Kingdom*

It is mentioned that there is a higher risk of MSM transmission. This suggest that risk of HIV transmission among MSM is increasing over time, which is not the case (the number of new infections among British MSM is declining). In the next sentence it is also mentioned that there is a growing role of HIV transmission among MSM and IDU in the UK. This is not true, as HIV transmission in these groups is declining. The number of new HIV diagnosis among IDU in the UK is quite low.

I think I disagree! First, I assume different risk groups may be managed differently (e.g., in terms of their awareness or the way the UK healthcare system is involved; this actually may be a reason why the MSM route incidence is indeed decreasing in recent years in the UK), and in general, a higher risk of the MSM transmission route doesn't necessarily define temporal dynamics of the incidence, even though we can observe a faster transmission phylodynamically. Second, both works to which you are referring to were performed on HIV data prior to 2010, when the role of the HET route, but not of the MSM route, seemed to decline (please see Figures 4 and 5 in [2] cited at the end of this document) - so back then the role of MSM did increase. I admit I cannot find decent statistics on IDU - this route is much less prevalent than MSM and HET. I have now replaced the [191] with a more relevant link to [2] and specified that [193] refers to 2007-2009 years in the thesis text (page 36).

- *Discussion*

The discussion in its present form is quite brief. I would like to invite the candidate to elaborate a bit further on implications of her work for e.g. public health and for molecular epidemiology. The study in Oryol identifies transmission clusters. Could we use these clusters to reduce the size of the HIV epidemic and how? What are the implications of her work for future pandemics? Are there improvements needed in molecular epidemiology?

Thank you, I have now expanded the Conclusions section (at the end, pages 121-123).

Prof. Matsen

- *I would have loved to read a bit more about what the student would propose doing, policy-wise, in order to better control these viruses. This may seem out of scope, but at the same time it's the point of (genetic) epidemiology.*

Thank you, I have now added some thoughts on this to the Conclusions section (at the end, pages 121-123).

Prof. Pervouchine

However, below are a few points to be clarified.

- *The analysis of HIV-1 epidemic is based on sequencing of only a part of the HIV genome that includes the pol gene. The defendant should spend some time to discuss the limitations of the*

analysis that are imposed by this, especially considering the existence of a circulating recombinant form.

I have now acknowledged this on page 56 (section 3.2.6). Indeed, fragment-based subtyping is prone to yet undescribed recent recombinant events in the viral population, so some of our samples may potentially be incorrectly annotated; this can only be figured out by analysing recombination patterns in complete genomes which we don't have access to. Yet, I doubt this affects a notable part of our dataset - URFs (unique recombinant forms) are rare among all recombinants in most geographic regions [3]. Mis-subtyping of existing variants is also possible but again supposedly rare - I could not find anything similar to A6, B, or CRF63 in the *pol* region among existing recombinant forms https://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html. CRF42 or CRF51 can potentially be misclassified as B based on the *pol* fragment, but these variants seem to be rare.

- *The routes of transmission of HIV-1 and SARS-CoV-2 are drastically different: while HIV is known to be transmitted sexually and parenterally, SARS-CoV-2 is transmitted by exposure to respiratory fluids. This aspect deserves to be mentioned somewhere early in the thesis to highlight the differences between the two illnesses (unless I overlooked it).*

Thank you, this is now mentioned in the Introduction (page 14).

- *The manuscript contains multiple abbreviations, not all of which are explained. For instance SIV (simian immunodeficiency virus), IDU (intravenous drug users), CDC on p. 38. Although the list of abbreviations exists, it would be convenient for the reader to have these terms explained as they first appear in the text.*

I have now added explanations to abbreviations that were not explained upon their first occurrence.

Minor points:

- *p. 26. l.8: "As higher thermodynamic stability of the transmitted variants suggests" – it is not clear for the context how the thermodynamic stability is related to selection.*

I have now rephrased this sentence and also the initial sentence I was referring to on page 25. The authors considered the effect of the absolute change in protein stability; although most mutations decrease rather than increase the protein stability, it was inaccurate of me to say that higher stability suggests selection - deviations both ways may be deleterious for a protein and thus may be subject to selection.

- *p.31. l.8: The author could make use of footnotes to take a moment and appreciate current progress in the field. Also, the use of non-English characters at the end of the sentence is not acceptable for a phD thesis.*

Footnotes are now used!

- *p. 62 Figure 3.1. I propose converting this display item to a table because some of the low counts cannot be seen from it. For instance, the author later mentions on p. 75 that there were a few MSM samples, while from Figure 3.1B it looks that there were none.*

Figure 3.1 demonstrates the difference between the two datasets; I think it is more illustrative as a figure. It can be seen from percent values provided that HETs and IDUs don't sum up to 100%, so MSMs are not necessarily absent there, though they are indeed rather rare in our dataset.

- *p. 66. Figure 3.4. The dependence of the inferred number of singletons and transmission lineages on the number of sequences – doesn't it depend not only on the number of sequences, but also on *which* sequences were used in the analysis? I missed the point here.*

We repeated the subsampling procedure 1,000 times in order to see whether it does (gray area on Figure 3.4; now added to the legend). Our results suggest it doesn't affect our inferences that much.

1. Juan Ángel Patiño-Galindo FG-C. The substitution rate of HIV-1 subtypes: a genomic approach. Virus Evolution. 2017;3. doi:10.1093/ve/vex029

2. https://webarchive.nationalarchives.gov.uk/ukgwa/20181112133715mp_/https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/326601/HIV_annual_report_2013.pdf

3. Global and regional epidemiology of HIV-1 recombinants in 1990–2015: a systematic review and global survey. https://doi.org/10.1016/S2352-3018(20)30252-6