RESPONSE TO REVIEWERS

Editor

Response: *Thank you very much for handling our manuscript and providing these two constructive reviews that have helped us to improve the paper.*

Editor: The comments and recommendations from two expert reviewers are now available for your manuscript. These reviewers judged the reported discoveries to be of medium significance, and the potential scientific impact of your work to be medium/high. They found that the manuscript needs improvement in text and additional data analysis. Editors generally agree with their concerns and recommendations, which led to a designation of medium/high priority.

Note that many manuscripts receiving medium priority based on reviewer comments are not accepted by the Board of Editors. If appropriate, the board may invite a resubmission following the rejection, which is intended to enable you to improve the manuscript towards receiving a high or top priority. The authors should pay close attention to the detailed review comments and address each comment with significant improvements.

Response:

Reviewer: 1

R1: The manuscript by Barrat-Charlaix et al. discusses the problem of predictability of mutations in seasonal influenza viruses. It presents data analysis of frequency trajectories in H1N1 and H3N2 lineages and simulations of different selection scenarios. It is an interesting revisit to predictions in influenza, however I have some major concerns about the analysis and formulation of the conclusions that should be addressed:

R1: 1. The prediction problem is defined for each mutation: based on the tracked frequency trajectory, can the future (fixation, loss, or polymorphism) of the mutation be predicted. Such formulation has been previously proposed by Illingworth and Mustonen, (eg. Genetics 2011, Plos Pathogens 2012). By averaging over all amino-acid substitutions, the authors show that the fates of mutations are not determined by the value of the starting frequency.

R1: 1.1. Mutations in influenza are highly nested, with a substantial hitch-hiking, and no attempt is made to disentangle such dependencies when counting the mutations.

Response: *This is a good suggestion, and we agree with reviewer 1 that our original manuscript lacked a method to disentangle nested mutations.*
We added a new section in the Supplementary Material where we attempt to cluster together trajectories of mutations that partly appear on the same strains. Trajectories corresponding to mutations always or often appearing on the same strains are then counted as one "effective" trajectory. However, this new way of counting mutations does not significantly change our results, and as a consequence we left the figures of main text unchanged.

R1: 1.2 Despite the more general formulation in the beginning, this approach makes use only of the last time point in the trajectory, rather than the full trace (contrary to the work of Illingworth& Mustonen). I find the observed neutral-like statistics not surprising for such a limited data input, which doesn't capture past frequency dynamics. Therefore, these conclusions should be revisited and benchmarked against the more general implementation of Illingworth&Mustonen-like approach.

ResponseDraft: *We thank reviewer 1 for pointing out the work of Illingworth&Mustonen to us, which we were unaware of. However, we believe that the method used in their article is not applicable as is in our case. To explain our thinking, we first briefly summarize the approach taken in Illingworth&Mustonen: the authors fit two kind of models to fully known frequency trajectories:*

- *An "unlinked" model, where the fitness effect of a mutation fully determines its ultimate fixation or loss. The only trajectories that the model can produce are upward or downward sweeps, for beneficial or deleterious mutations respectively.*

- *A "linked" model, in which the fate of a mutation is determined both by its own fitness and by the fitness of other mutations present on the genomes it appears on. This model fits actual trajectories much more accurately, which is one of the main findings of the paper.*

Interestingly, both models depend on the same parameters that have to be fitted, namely the fitness effects of individual mutations. Linkage between mutations in the second model is introduced by using the measured frequencies of appearance of two mutations on the same genome.
Pierre: What I'm trying to say below

- *Illingworth&Mustonen don't do prediction, so there is no clear way to apply their method to our article.*
- *Fitting the initial part of the trajectory using their models doesn't make much sense.*
- *We want to stay model free: it's possible to find signs of selection without fitting (cf. H1N1/simulated data)*
- *Even if we're doing something very simple, it's surprising to see neutral-like behaviour.*

We first want to point out that this approach is fundamentally different to the one of our article. Since Illingworth&Mustonen fit models to fully known trajectories, these models cannot meaningfully be used for prediction purposes. Indeed, fitness effects of individual mutations are fitted with knowledge of the fixation or loss of the given mutation, and of the frequency of other mutations in the case of the second model. This strongly contrasts with our approach which is to find patterns of predictability of the future behaviour of frequency trajectories using past data only: e.g. *if a frequency trajectory has risen from* 0 *to a frequency* $f > 0$*, what can be said of its future shape or of its fixation? Thus, methods in Illingworth&Mustonen are designed to answer a different question than the one we are asking in our work.*
A way of reconciling the two approaches would be for us to fit the initial part of a trajectory with either model, and to use this fit to predict its future behaviour. However, we feel that this idea is not very relevant to our article for two reasons. First, neither model proposed by Illingworth&Mustonen is really practical in our setting. The "unlinked" model only generates sweep-like trajectories, which our results show to be not representative of the "typical" trajectory (in good agreement with Illingworth&Mustonen). The "linked" model can fit more complex trajectories, e.g. *ones that rise rapidly but ultimately die. However, it crucially depends on the frequencies of joint appearance of pairs mutations on genomes to introduce linkage. These frequencies vary through time, allowing the* effective *fitness of a mutation to vary, and a trajectory to rise and then fall. In our case, these frequencies would only be known for the past, making impossible any prediction of the model.*
The second reason is that we believe that our model-free approach has intrinsic benefits. [Something saying that we'd prefer not to start fitting models to data, because it's not the point. But I'm not sure how to formulate this.]
*Finally, we are aware that using only the last time point of a rising trajectory gives only limited opportunity for predicting its future. However, observing apparently-neutral statistics in the case of A/H3N2 is still surprising to us. The case of simulated populations shows that clear signs of selection can be observed by looking at basic features of trajectories (*e.g. *figure S20 of the supplementary material). This is consistent with the theory of evolving population: if the only available information about a mutation is that it has risen in frequency starting from* 0*, it has a higher chance on being beneficial than deleterious, and its fixation probability should be higher than what it would be in a neutral scenario. Hence, even with our very simple approach, we observe that predictability of A/H3N2 qualitatively differs from what models of adapting populations would suggest.*

R1: 2. The authors examined the predictive power of one predictive method, the LBI. However, they did not do a systematic comparison of the different methods that they cite (Morris et al, 2018), which differ in prediction targets and methods. Therefore, the general conclusions about predictability, eg. on page 8, in line 47, and on page 9, line 48 are too sweeping and should be made precisely for those methods looked at in detail.

R1: 3. Going through the previous and cited literature, I think the authors should cite some of these works in a more careful way. Specifically, the very related work by Illingworth and Mustonen is not mentioned. The distribution $P_delta_t(f|f_0)$ has been at the core of the method of Strelkowa and Lassig, 2012 (which is cited, but at other parts of the text), and the same paper also uses a similar simulation model, which should be acknowledged at the appropriate points of the text.

R1: It would help if the plots with red-green-blue lines were also distinguished by differing markers.

Response: *We have now added markers in most figures.*

Reviewer: 2

R2: The paper does a retrospective study of amino acid substitutions in seasonal Influenza to determine what properties of these substitutions could help predict their fate in the future. The authors find that future frequency trajectories are surprisingly unpredictable. Even predicting which mutations fix in the population is hard. The authors find that the current frequency of a mutation is the best predictor for the probability of fixation, which would be expected under neutrality but not in a model with selection. I appreciate this study, I think it will be of interest to many readers and it is generally well done and fairly easy to follow.

R2: 1. The authors focus on one feature at a time (frequency, epitope status, LBI). It is interesting to see that each of these is not very predictive of future frequency / prob of fixation. However, I think the obvious next step is to see whether a combination of many features could do a better job of predicting. I am not sure why the authors dont try to fit a model that takes into account all information they have about sites (say, type of AA change, location in the gene, current frequency etc) and see if a ML model is able to make predictions.

R2: 2. Fig 2A and 2B look quite different to me. In the text it appears to me as if these are very similar to the authors. What are the characteristics of the AAs that fix in H1N1?

R2: 3. Question: what do these results mean for vaccines? How to decide which strains to use for vaccines? This may be known to those who work on Influenza, but for a relative outsider it is not clear.

R2: 4. Question: at what point is it expected that the vaccine itself will influence frequency trajectories? See Wen et al Biorxiv 2020

ResponseDraft: *We should mention Wen* et. al. *MDPI-Viruses 2018:* Estimating Vaccine-Driven Selection in Seasonal Influenza. *The answer is that yes it's expected that vaccine influence frequency trajectories to some extent, but it is very hard to measure in practice. The above paper attempted to measure this, and did not find clear-cut results.*

R2: Fig 1B is very hard to read. Maybe it should get more space (fig 1C could work with less space).

ResponseDraft: *We have reworked this figure to allow more space for panel B.*