RESPONSE TO REVIEWERS

Editor

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.

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

Editor: 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: *We have used the revision to include additional analysis and improve the presentation. We hope that you and the reviewers now feel that our manuscript makes a major contribution towards understanding what aspects of influenza virus evolution are predictable, and which ones are not. Given the major public health importance or influenza and the fact that influenza virus evolution is one of the very few systems where adaptive evolution can be observed in real time, we feel that our results are of very high relevance to the community.*

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:

Response: *We address the individual points by the reviewer below.*

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.

Response: *We thank the reviewer for pointing out the work by Illingworth and Mustonen, which is indeed very relevant and complementary to our approach. We discus their relation below.*

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: *The reviewer is correct that our original manuscript did not attempt to account for nested mutations and we agree that this should have been done.*
We added a new section in the Supplementary Material where investigate the extent to which nesting affects our results: We cluster trajectories of mutations that partly appear on the same strains to various degrees and trajectories corresponding to mutations always or often appearing on the same strains are then counted as one "effective" trajectory. Our analysis shows that nesting of mutations does not change our results significantly. We therefore left the figures of main text unchanged and refer to the supplement for a discussion of this issue.

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.

Response: *Illingworth and Mustonen (I&M) presented a careful analysis to what extent trajectories can be fit using a model assuming a single fitness coefficient, or a model that differentiates different background fitnesses. Our results are consistent with their finding that a single parameter model (which only allows monotonic trajectories) does capture the majority of trajectories. Beyond that, our approach and aim differ from I&M. We investigate a priori features of mutations that have information on whether the mutation fixes, not whether the trajectories can be parameterized by a model. Fitting a model to a partial trajectory to predict fitness is not expected to work: I&M showed that subsequent interference has a substantial effect and our results indicate that the speed at which the mutation rises does not predict fitness.*

Furthermore, I&M estimate fixing mutations to have a much higher intrinisic fitness than dying mutation while we show that very little such information is available a priori. The retrospective fit to trajectories tends to assign a positive fitness effect to fixing (and hence on average rising) mutations. This is why we resort to a model free approach ask what a priori information can help to determine which mutations fix or die.

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.

Response: *In a recent model-driven effort (Huddleston et al), we investigate the power of sequence based, antigenically informed, experimentally informed, or phylogenetic (LBI) scores to predict future influenza virus populations. The LBI was among the highest performing models and we hence used this score in the original manuscript. But we took the suggestion of the referee (and referee 2) to heart and now include all highly performing models from Huddleston et al. These additional results show that only mutational load has some power to differentiate between fixing and non-fixing trajectories. We believe this addition has made the manuscript stronger and we thank both referees for the suggestion.*

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.

Response: *We thank the reviewer for this assessment and the constructive suggestions below.*

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.

Response: *A similar point was also brought up by reviewer 1. In addition to the LBI, epitopes, age, and geographic distribution we now also include the fitness scores of the best performing models in a recent machine learning effort to predict the composition of future populations (as opposed to predicting fixation). These models include antigenic, experimental, and sequence or phylogenetic information. Interestingly, all these models do similarly poorly and with the exception of mutational load. These results are now included as supplementary information and discussed.*

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?

Response: *Whether 2A or 2B are perceived as similar or not probably depends on how big a deviation from the diagonal one expected a priori. We now discuss now highlight the differences in the text and discuss potential reasons for the different behavior.*

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.

Response: *Computational prediction of which circulating strains will dominate future influenza populations have been attempted for about 2 decades and have recently become a regular input to the biannual consultations on the influenza vaccine composition. Since performance of predictive models can only be assessed on limited historical data, we believe that our model-free approach clarifies which aspect of influenza evolution are predictable and which ones are not.*

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.*

Response: *The degree to which the vaccine drives evolution is a very interesting question! In the past, global influenza vaccine uptake was unlikely sufficient to generate much pressure and fairly thorough mixing of variants across countries with very different vaccination rates have been used to argue against a strong effect. But this might have changed in recent years and the paper by Wen et al in Viruses, 2018 picks up some signal. Overall, the case for strong effect is still weak to our knowledge.*

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.*