This preprint worth a revision
by Guillaume Achaz

The ms by Bertels et al. has been reviewed by three independent experts in population genetics and molecular evolution. All three reviewers found that this ms has a good potential but also raised important points that need to be addressed before it can be recommended by PCI Evol Biol. Reviewers 1 and 2 suggested several articles that the authors must read and potentially include as references in their revised version. Reviewers 2 and 3 were convinced that the convergence approach is interesting but at the same time show some concerns on the power and the reliability of the method. I also agree with reviewer 3 that this study should not be oversold, as results are not extremely robust as they are.

Please address carefully all points raised by the reviewers and revise you manuscript accordingly. A point by point response to their comments must be included along with your revised version of the ms.

> *We would like to thank Guillaume Achaz as well as all three reviewers for the tremendous amount of work they have invested into reviewing our manuscript. We have addressed the comments as comprehensively as we could and think that the changes significantly improve our study.*

Reviewer 1 (Jeffrey Townsend)

This manuscript reports intriguing results illustrating the potential to use convergent evolution of sites in genes of HIV (namely, the env gene) as an indicator of the action of natural selection during acute HIV-1 infection. It illustrates that convergent mutations are frequent during infection of HIV-1; that they are more frequently located in the gp41 domain; that they don't occur preferentially in positions of high nucleotide diversity; and that they aren't significantly more likely to be synonymous in nature. Overall, the results are persuasive and interesting, and there is definitely utility to greater analysis of convergence as an indicator of selection in molecular evolution in general and the molecular evolution of HIV specifically. There are three major and a number of minor points that could be addressed in a revision to increase the impact of the manuscript.

First, additional contextualization of this work in comparison to other work (sometimes on other HIV genes) examining viral convergence (e.g. Xue et al. 2017) and the early evolution of HIV during acute infection (e.g. Herbeck et al. 2009, 2011a,b; Lee et al. 2009; Boutwell et al. 2010; Giorgi et al. 2013; Gounder et al. 2015; Garcia-Knight et al. 2016) is warranted. Particularly relevant to this manuscript are Yoshida et al. 2011; Henn et al. 2012; and Park et al. 2016.

> *We would like to thank Jeff for pointing these papers out to us. We have included all of them in our manuscript and in the process have rewritten the introduction.*

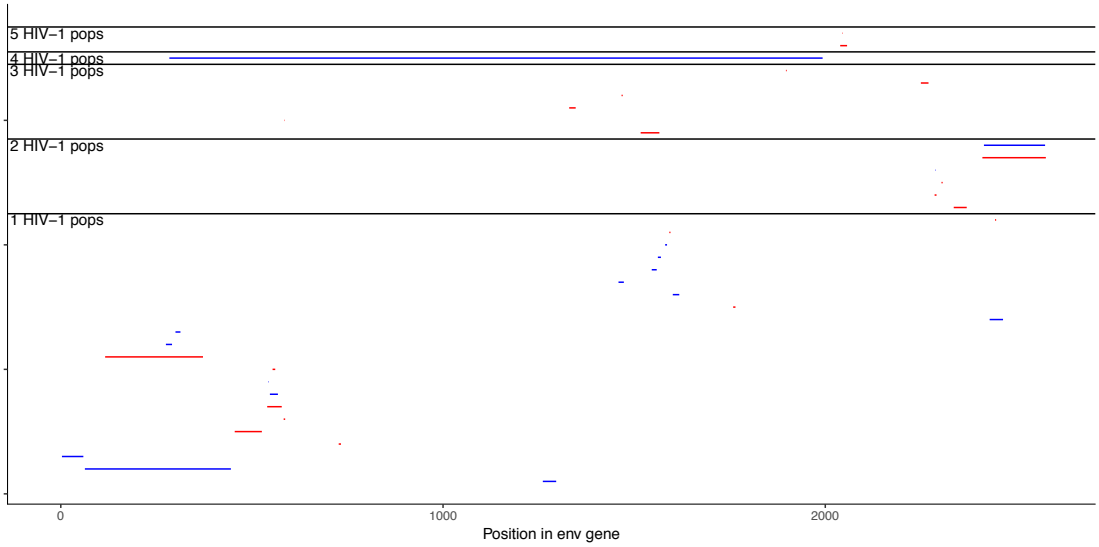
Second, adherence to two guidelines would enhance readability and persuasiveness. First, a greater adherence to the guideline of dedicating the x-axis to the independent variable and the y-axis to the dependent variable would improve the figures. In Figure 1 it is OK, but using the HIV populations on the x-axis in Figure 1 sets up a precedent that when re-used in Figures 2, 3, and 5 means that the variable that is dependent on the hypothesis ends up depicted on the x-axis, which is confusing. Switching the axes would improve their comprehensibility. Additionally, Figure 4 would be clearer if proportion nonsynonymous (amino acid replacement) were illustrated instead of proportion synonymous, because the conventional null hypothesis would be that nonsynonymous changes would be more likely to be convergently selected—at least in comparison to synonymous changes, which are often and conventionally assumed to be neutral.

> *We have switched the axes of Figure 2 and have changed Figure 4 as to Jeff's comments. However, we kept the axes as they were in Figure 3, 4 and 5. We felt that we were actually interested how the diversity etc changes for mutations that are more convergent, rather than asking the question on how convergence changes as mutations occur in more diverse regions of the genome.*

Third, for one of the main results of the manuscript (the clustering of convergent mutations within the discrete linear sequence of the env gene), there are much more powerful and persuasive approaches than are used here that can detect and illustrate the clustering of convergent mutations without a priori assumptions (e.g. Tang and Lewontin 1999; Zhang and Townsend 2009) that would represent a significant improvement over the ad hoc approach underlying Figure 2. It might be best to select an appropriate threshold of number of populations or (better) to perform clustering on all applicable thresholds (> x populations) to ensure that the result isn't a product of selecting a unique threshold and is not abrogated by saturation. Generally, it would improve the manuscript methodologically to more explicitly differentiate the utility of the methods used in this manuscript in comparison to other approaches such as molecular evolutionary models of selection (e.g. Rodrigo and Learn 2007; Kosakovsky Pond et al. 2008; Leitner 2012; Zhao et al. 2017) .

> *We have applied MACML and find it very interesting. We included an analysis in **Figure 2B.** MACML does support our observation that mutations that occur in three or more individuals are more likely to occur in the 3' part of the env gene. Additionally to running MACML, we also performed the chi square test for every possible division of the env gene for private mutations and for mutations that occur in four or more HIV-1 populations in parallel (**Figure 2B**). The minimum occurs at position 1472 (base number 7663 (HXB2)) Our MACML run for mutations that occur in three or more populations identified one significant cold cluster that overlaps with the non-gp41 part of the gene and ends at 7663 (**Figure 2**).*
>
> *However, we also came across a few issues with MACML:*

1. *MACML cannot take multiple independent mutations at the same position into account*
2. *The local approach of MACML can lead to strange artifacts that predict a hot and a cold spot for almost exactly the same region (see figure below)*



The figure shows hot spots (red) and cold spots (blue) of mutations across the env gene for mutations that occur in only a single population (private mutations "1 HIV-1 pops") up to all mutations that occur in 5 populations in parallel ("5 HIV-1 pops"). As you can see cold spots and hot spots can overlap quite significantly. The difference between overlapping hot/cold spots are the subsequence for which the model selection has been performed.

Minor points:
1. Consider the points raised in Stayton (2015) regarding research on "convergent evolution".

We added a paragraph on the definition of convergent evolution and how we identify convergence that is likely to be the result of adaptation.

2. Throughout the manuscript, for clarity, follow "this" immediatly by a noun referent.

We have adjusted all instances of "this" not followed by a noun.

3. The writing in the abstract needs work: what is "it" (line 12)? This is missing a referent (line 13). Population(s) (line 15)? Use of negative cases (not...not) in lines 21–22 is awkward and confusing.

We changed the abstract and hope it is clearer now.

4. Run-on sentence lines 62–65.

We deleted this sentence.

5. Are differences in mutation rate across sites adequately accounted for (lines 79–89)?

We did not account for potential differences in mutation rates across sites. Our simulation however maintains the original substitution rates for each dataset as well as the number of substitutions that occurred in each HIV-1 population to compute our null model.
In the new version of the manuscript we also excluded sequences that contained three or more mutations. This approach allows us to exclude all sequences that are significantly affected by APOBEC, which introduces G to A mutations at certain motifs in the HIV-1 genome.

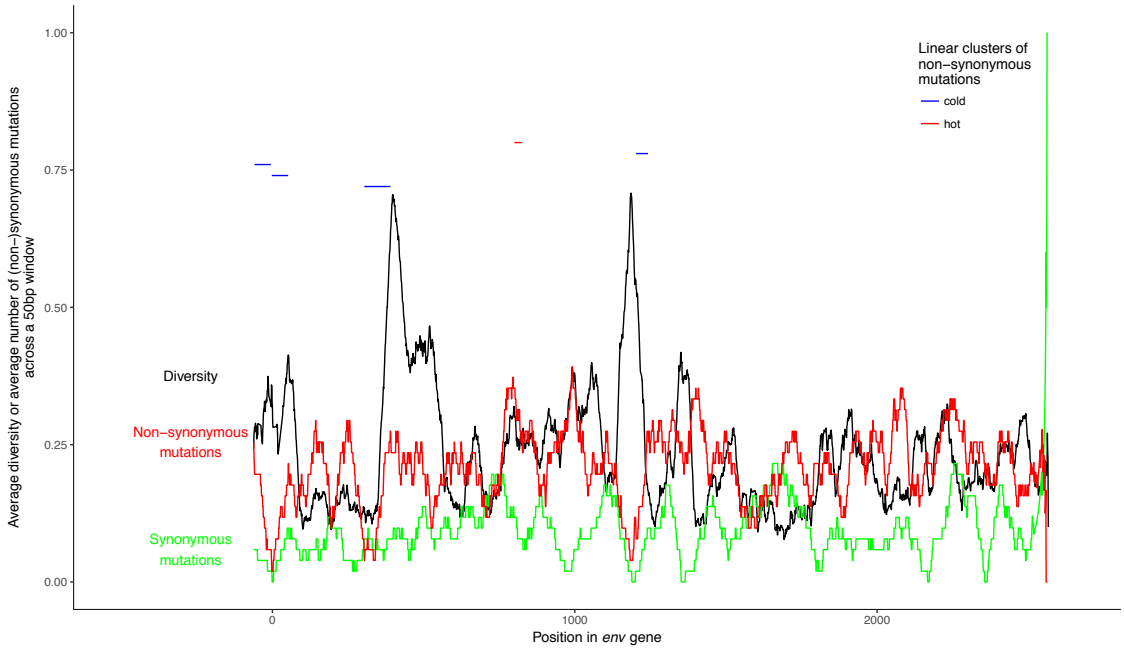
6. Awkward / confusing phrasing, lines 147–148 of Fig. 3 legend.

We changed the figure legend.

7. Confusing writing on lines 165–168; break up & clarify sentence.

We hope it is clearer in the edited version.

8. Test the assertion in line 190 by plotting diversity across sites and discrete linear clustering across sites in the same plot.



We have plotted the requested data. As you can see in the plot above there are correlations between high diversity regions and the average number of nonsynonymous substitutions in the same region. There are however only a limited number of linear clusters, of which most are cold spots. We have included this figure as Supplementary Figure 1.

9. “Is adjust” line 203.
10. Writing requires greater clarity, lines 214–222.

11. Writing requires greater clarity, lines 258–262.

We significantly rewrote the paper and hope it is clearer now.

12. On line 346, nucleotide diversity is defined as Shannon entropy, which by definition has a range from 0–1. On line 354, nucleotide diversity is stated to range from 0–2. Clarify.

To our knowledge Shannon Entropy is not defined as ranging from 0-1. As far as we know Shannon entropy is often provided in bits of information i.e. the logarithm would be to the base of 2. However, we think that the reviewer is right in that the data would be easier to interpret if the diversity ranged from 0 to 1. Hence for each position we determined the number of different bases and used this as the base of the logarithm. If there was only a single base (full conservation) the diversity was set to 0.

References cited:

Boutwell, C. L., M. M. Rolland, J. T. Herbeck, J. I. Mullins, and T. M. Allen. 2010. Viral evolution and escape during acute HIV-1 infection. J. Infect. Dis. 202 Suppl 2:S309–14.
Garcia-Knight, M. A., J. Slyker, B. L. Payne, S. L. K. Pond, T. I. de Silva, B. Chohan, B. Khasimwa, D. Mbori-Ngacha, G. John-Stewart, S. L. Rowland-Jones, and J. Esbjörnsson. 2016. Viral Evolution and Cytotoxic T Cell Restricted Selection in Acute Infant HIV-1 Infection. Sci. Rep. 6:29536.
Giorgi, E. E., B. T. Korber, A. S. Perelson, and T. Bhattacharya. 2013. Modeling sequence evolution in HIV-1 infection with recombination. J. Theor. Biol. 329:82–93.
Gounder, K., N. Padayachi, J. K. Mann, M. Radebe, M. Mokgoro, M. van der Stok, L. Mkhize, Z. Mncube, M. Jaggernath, T. Reddy, B. D. Walker, and T. Ndung’u. 2015. High frequency of transmitted HIV-1 Gag HLA class I-driven immune escape variants but minimal immune selection over the first year of clade C infection. PLoS One 10:e0119886.
Henn, M. R., C. L. Boutwell, P. Charlebois, N. J. Lennon, K. A. Power, A. R. Macalalad, A. M. Berlin, C. M. Malboeuf, E. M. Ryan, S. Gnerre, M. C. Zody, R. L. Erlich, L. M. Green, A. Berical, Y. Wang, M. Casali, H. Streeck, A. K. Bloom, T. Dudek, D. Tully, R. Newman, K. L. Axten, A. D. Gladden, L. Battis, M. Kemper, Q. Zeng, T. P. Shea, S. Gujja, C. Zedlack, O. Gasser, C. Brander, C. Hess, H. F. Günthard, Z. L. Brumme, C. J. Brumme, S. Bazner, J. Rychert, J. P. Tinsley, K. H. Mayer, E. Rosenberg, F. Pereyra, J. Z. Levin, S. K. Young, H. Jessen, M. Altfeld, B. W. Birren, B. D. Walker, and T. M. Allen. 2012. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. PLoS Pathog. 8:e1002529.
Herbeck, J. T., M. Rolland, Y. Liu, S. McLaughlin, J. McNevin, K. Diem, A. C. Collier, M. Juliana McElrath, and J. I. Mullins. 2011a. 175 HIV-1 Evolution in Primary

Infection is Affected by Stochastic Followed by Selective Processes. JAIDS Journal of Acquired Immune Deficiency Syndromes 56:73.
Herbeck, J. T., M. Rolland, Y. Liu, S. McLaughlin, J. McNevin, H. Zhao, K. Wong, J. N. Stoddard, D. Raugi, S. Sorensen, I. Genowati, B. Birditt, A. McKay, K. Diem, B. S. Maust, W. Deng, A. C. Collier, J. D. Stekler, M. J. McElrath, and J. I. Mullins. 2011b. Demographic Processes Affect HIV-1 Evolution in Primary Infection before the Onset of Selective Processes. J. Virol. 85:7523–7534.
Herbeck, J. T., M. T. Rolland, W. C. Deng, A. C. Collier, and J. I. Mullins. 2009. P07-06. HIV-1 transmission and early evolution: whole genome analysis. Retrovirology 6:P104.
Kosakovsky Pond, S. L., A. F. Y. Poon, S. Zárate, D. M. Smith, S. J. Little, S. K. Pillai, R. J. Ellis, J. K. Wong, A. J. Leigh Brown, D. D. Richman, and S. D. W. Frost. 2008. Estimating selection pressures on HIV-1 using phylogenetic likelihood models. Stat. Med. 27:4779–4789.
Lee, H. Y., E. E. Giorgi, B. F. Keele, B. Gaschen, G. S. Athreya, J. F. Salazar-Gonzalez, K. T. Pham, P. A. Goepfert, J. M. Kilby, M. S. Saag, E. L. Delwart, M. P. Busch, B. H. Hahn, G. M. Shaw, B. T. Korber, T. Bhattacharya, and A. S. Perelson. 2009. Modeling sequence evolution in acute HIV-1 infection. J. Theor. Biol. 261:341–360.
Leitner, T. 2012. The Molecular Epidemiology of Human Viruses. Springer Science & Business Media.
Park, S. Y., T. M. T. Love, A. S. Perelson, W. J. Mack, and H. Y. Lee. 2016. Molecular clock of HIV-1 envelope genes under early immune selection. Retrovirology 13:38.
Rodrigo, A. G., and G. H. Learn Jr. 2007. Computational and Evolutionary Analysis of HIV Molecular Sequences. Springer Science & Business Media.
Stayton, C. T. 2015. What does convergent evolution mean? The interpretation of convergence and its implications in the search for limits to evolution. Interface Focus 5:20150039.
Tang, H., and R. C. Lewontin. 1999. Locating regions of differential variability in DNA and protein sequences. Genetics 153:485–495.
Xue, K. S., T. Stevens-Ayers, A. P. Campbell, J. A. Englund, S. A. Pergam, M. Boeckh, and J. D. Bloom. 2017. Parallel evolution of influenza across multiple spatiotemporal scales. Elife 6.
Yoshida, I., W. Sugiura, J. Shibata, F. Ren, Z. Yang, and H. Tanaka. 2011. Change of Positive Selection Pressure on HIV-1 Envelope Gene Inferred by Early and Recent Samples. PLoS One 6:e18630.
Zhang, Z., and J. P. Townsend. 2009. Maximum-likelihood model averaging to profile clustering of site types across discrete linear sequences. PLoS Comput. Biol. 5:e1000421.
Zhao, Z.-M., M. C. Campbell, N. Li, D. S. W. Lee, Z. Zhang, and J. P. Townsend. 2017. Detection of regional variation in selection intensity within protein-coding genes using DNA sequence polymorphism and divergence. Mol. Biol. Evol., doi: 10.1093/molbev/msx213 .

Reviewer 2

The ms by Bertels et al. reports an analysis of nucleotide convergence pattern in HIV. It reads well, is mostly sound and quite easy to follow. I only however few remarks that could potentially help improving its content.

:: Major ::

Although the authors demonstrate clearly that some positions have mutated several times independently in different patients, I am not convinced this is really due to selection. One important part of the puzzle (that is never discussed) is the type of mutations the authors have found independently repeated. A summary table listing all types recurrent mutations (i.e. the type of nucleotide change) is required in the main text. As they are mostly G->A mutations (Table S1), this is suspicious as HIV has a very strong mutational bias in that direction. It would be much more convincing to find that the apparently selected mutations are not all of the same nature. If I understood Table S1, all are G->A or A->G, but the latter could simply be mis-oriented mutations (see the minor points below).

> *We are not entirely sure we interpreted the reviewer's comment correctly, but we hope that what he meant is that mutational bias (i.e. the type of a substitution) could be the reason for the independent observation of identical mutations in different HIV-populations. Because we are comparing the observations from the Keele and Li data to a neutral model, where we maintained the observed substitution rates, we can rule out that the substitution type affects our conclusions. We highlighted in relevant places that substitution rates were kept the same in the simulations.*
>
> *The fact that most of the convergent mutations are G->A is not only the result of higher G->A substitution rates but could also be an effect of APOBEC modifications. APOBEC introduces G->A mutations in the viral sequence to destroy it. In our revised version of the manuscript we excluded sequences that contain more than two mutations to the consensus sequence and hence remove most sequences affected by APOBEC.*

I am not sure how to interpret biologically the value of H. H mixes drift, selection and recurrent mutations. Some other metrics such as the number of alleles (2, 3 or 4) are more directly measuring the number of mutations at a site.

> *H is a measure of nucleotide diversity between HIV viruses from different hosts. We assume that the 95 different HIV populations (all recently founded by a single virus) are random samples from the HIV-1 population as a whole. We can measure diversity at each position of the* env *gene. The diversity at these positions can tell us something about the conservation of certain nucleotide positions. We have included a more comprehensive explanation of what exactly we mean by the diversity measurement and the purpose of the analysis.*
>
> *As per reviewer 1s request we have changed H to range from 0 to 1. This means that if there are two alleles, three alleles or four alleles at equal*

frequency at a certain position then in all three cases the nucleotide diversity will be 1.

As a general comment, I think there is room for improvement in the general flow of the article. While reading it few times, I am still confused about the statements. Casual readers could easily get lost.

We have rewritten the entire manuscript and have tried to improve the flow and highlight the points that we are trying to make.

The weak overlap between the author list of potentially selected mutations with the one from Wood et al. can suggest that the data are quite noisy and the overall power of the method(s) are simply quite weak. Although I believe this was a clever method, more discussion on this point (limitations of the method) would be welcome.

We added a paragraph in the Discussion that highlights the limitations of our method.

:: Minor ::

l142 - why did the authors chose to report only the results for >= 3 populations ? What about providing the full distribution? Can the authors also give the raw number (and not only the %). Furthermore, although this is statistically significant, it leaves 30% that are outside the gene. This cast doubts on the strength of the reported pattern.

We have added a figure (Figure 2B) to the manuscript that analyses the entire distribution for convergent mutations and shows that the difference between the distributions is highly significant.

l70 - the dN/dS strategy would also work if selection affect less dS than dN, not necessarily that dS has to be immune to selection.

This is absolutely true and we agree that this statement was to strong and hence have removed it from the manuscript. Instead we provide an overview of how selection was identified in the past: "Selection can be measured in various different ways: (1) Probably the most common way of determining selected nucleotide sites is by comparing the evolutionary rates of non-synonymous nucleotide sites to those of synonymous sites (dN/dS) on a phylogenetic tree (Kosakovsky Pond et al. 2008; Wood et al. 2009; Boutwell et al. 2010; Yoshida et al. 2011). (2) If population sequence data that spans multiple time points is available then it is possible to identify positively selected sites by assessing the change in mutant frequency over time (Henn et al. 2012). (3) More recently it has been demonstrated that it is possible to determine nucleotide sites under selection by analyzing the distribution of those sites across a gene (Zhang and Townsend 2009; Zhao et al. 2017). All

of these methods have advantages and disadvantages and all of them have helped to better understand HIV-1 evolution."

l303-307 - the ancestral sequence is not always the consensus. Mutations could simply reach high-frequency. This is true even in the standard neutral model (see expectation of the unfolded SFS). So I guess the direction of mutations may be unsure and therefore authors may want to pool symmetrical mutations (i.e. G->A with A->G and if mutations can occur in the two strands also with C<->T).

It is true that the ancestral sequence is not always the consensus. However, in their original publication Keele et al. have modeled the mutation dynamics in HIV and have come to the conclusion that in our case it is highly likely that the consensus sequence is also the founder sequence. This does not exclude the possibility that the direction for some of the mutations is incorrect. But we think that it would be wrong to leave out the direction of the mutation just because there is a small chance that for a small number of mutations the direction is incorrect.

l309-315 - why doing an alignment with a reference sequence ? (and not only all sequences together without the ref). This seems odd.

We redid our analysis and performed a standard multiple sequence alignment. All our analyses and results are based on the new alignment. Our results did not significantly change.

l348 - Did you consider using an entropy between 0 and 1 instead of [0,2] ? You would need to use log*4 instead of log*2. Eventually, you could change the base of the log depending on the number of alleles.

We changed our approach so that the diversity ranges from 0 to 1. Of course this approach has its own caveats as you assume that there are only as many possible variants as you observe in the alignment. However, because mutations do not occur with equal probability (e.g. bias towards G to A) we feel that the suggested approach has more advantages than disadvantages and changed our manuscript accordingly.

l220-l223 - Please clarify as it is slightly confusing as it is.

We rewrote this sentence.

Reviewer 3

Summary

By re-analyzing 95 independent full-length HIV env genes this study relates the extent of (genotypic) convergent evolution to the presence of selection acting on

specific mutations. The authors find an excess of convergent mutations in the gp41 region of the env gene when compared to a neutral model, supporting the view of positive selection acting on these mutations as has previously (partially) been found by Wood et al. 2009 using dN/dS approaches. One advantage of their approach to the former though is that is allows identifying positively selected synonymous mutations. Furthermore, the authors found "convergent mutations" are not more likely to be non-synonymous than under a neutral model. Overall the authors conclude that the extent of convergent evolution can be a good predictor of positive selection.

GENERAL OPINION AND MAJOR POINTS

I am a bit split on this paper, and I think the reason is that it tries to convey an easy to grasp story ("Convergent muts are under pos select and private muts under neg selection during acute infection of HIV-1" as one of the authors put it on Twitter), while exactly this simplification makes it sometimes sound like a slight oversell.

First, this paper does not introduce a real statistical method to detect selection. To be fair though this is not what the authors claim (or what they are trying to do). However, if the authors were thinking about making it a "real" method, it would most certainly require intensive simulations of genomes under selection (along independent replicates) to assess the power of the method, and under which situations it is able to detect it. Despite not being a method paper though, the latter point needs to be addressed more thoroughly I feel. There are a couple of crucial assumptions that probably for the sake of conciseness have not been discussed: For this method to work the viruses (or the system to be studied) need to be "unconditionally selected", i.e., selection pressures/environmental conditions and mutational effects need to be (close to) identical. This, however, implies that epistasis does not affect the evolutionary dynamics, and/or that the genomic basis / mutational target size for this trait is small. For instance, if a trait with a highly polygenic basis and thus a large mutational target size was under selection, this approach will probably not perform well (note that selection strength is potentially equally crucial here). Thus, viruses might really be the perfect if not only system where the authors' approach can be applied. Thus, I think the real advantage of this approach comes out when relating and "benchmarking" it to the results by Wood et al. which focus on the same gene (env; and its gp41 part), but relies on a dN/dS approach, and thus cannot identify synonymous mutations and lacks power when applied to population genetic data (as the authors rightfully cite Kryazhimskiy & Plotkin 2008 here). Invoking selection on a part of the genome that has been shown to be under selection is furthermore not very exciting unless the results are related to an earlier study as proposed. Furthermore, the idea of using convergent evolution to detect targets of selection seems a bit trivial when having some background in experimental evolution, where independent experimental replicates are a requirement for invoking selection on a genetic variant (and getting the result published). Note that Andreas Futschik has recently developed an just uploaded a paper on arxiv called "An omnibus test for testing global null hypotheses.", where he derives a statistical framework for inferring sites under selection where that have not

been selected uniformly across all replicates (i.e., where not all experimental lines show convergent genotypes). However, when seen as a "proof-of-principle" paper though and results are related to the Wood et al findings, and the underlying assumptions are discussed a bit more in detail, this manuscript can turn into a very nice and stimulating paper.

First and foremost we would like to thank the reviewer for putting a tremendous amount of work into reviewing this paper. We really appreciate it!

It was not our intention to oversell our research on Twitter. We simply stated the two main observations (positive and negative selection) we made in our study, which is supported by evidence that we present in our manuscript (we provide a link to the biorxiv version in the tweet).

Our analysis shows through comparisons to an appropriate null model that there are sites that occur in too many viral populations in parallel to be explained by neutral evolution. We have bolstered this observation by showing that convergent mutations are not evenly distributed across the env gene, which indicates that these mutations are adaptive. Similarly we showed that private mutations show a very biased distribution towards sites of high diversity, something we would not expect for randomly distributed mutations (positive selection). Instead of purifying selection this result could also be an effect of mutational hotspots however given that synonymous mutations are distributed as expected under a random model we feel that the biased distribution of private mutations is the result of purifying selection, i.e. we do not see mutations in low diversity regions because they are likely to have a large fitness cost (negative selection).

The reviewer is right in that if we wanted to publish a method we would have to approach this paper in a completely different way. However, here we are trying to test whether we can find signatures of selection in the given dataset by looking for convergent evolution (something as the reviewer rightly points out probably only works in viral genomes). This dataset is particularly interesting because the samples were taken before the adaptive immune system recognizes the virus and before drug therapy has been started. Hence, naively one would not expect selection to affect the evolution of these populations, which is not what is happening.

More specific, minor points (line L; referring to the authors' line numbers) separated by section:

#

GENERAL COMMENTS

Thank you for the line numbers. Please check your spelling of "re-analyzed vs. reanalyzed" and be coherent.

Done.

ABSTRACT

L23 "highlighting the highly..." But this statement strongly depends on the effect-size of the selected sites. Also shouldn't it be non-synonymous?

> *We felt this part of the abstract and the paper was rather confusing and maybe overemphasized the fact that we found evidence for non-neutral synonymous mutations. We changed the abstract to focus on what we think are the main signals we identify in our analysis. Namely, that over-represented convergent mutations are evidence for positive selection and that private mutations are affected by purifying selection. The strongest piece of evidence that supports the latter statement is the biased distribution of non-synonymous mutations towards regions of the* env *gene that are highly diverse. In the revised version we only briefly mention that we also observe convergent synonymous mutations, which indicates that synonymous mutations (as observed previously) are not selectively neutral.*

INTRODUCTION

L41 "small genomes, high mutation rates,..." The mentioning of high mutation rates is a bit misleading here since you have not defined on which scale you consider genotypic convergent evolution. If you were considering haplotypes or entire genomes, high mutation rates probably won't facilitate convergent evolution.

> *We agree with the reviewer that high mutation rates per se should not affect rates of convergence and have deleted that part of the sentence.*

L54 "Convergence is also..." I thought that convergence is an outcome rather than an initial state. Do you mean reduced diversity here?

> *This is true. We rewrote this paragraph.*

L62 "These mutations seem to be..." This paragraph is way too long and complicated. Please consider splitting into two.

> *We also rewrote this paragraph as well as most of the introduction. We hope that this makes the manuscript easier to read.*

L72 "is less reliable when..." Consider adding the point that synonymous mutations cannot be detected by this approach either.

> *We have added this to the discussion as we changed removed this paragraph from the introduction.*

L83 "convergent mutations" I think it was OK for the abstract, but you should give a formal definition of what you actually mean by that here. E.g., you could introduce the term more formally in line 80 at the end of the first sentence.

We have added this paragraph to the manuscript: "Here we will focus on yet another way of determining nucleotide sites that are under selection: measuring the frequency of convergent mutations across different HIV-1 populations from different hosts. Convergent mutations are mutations that occur in independent HIV-1 populations in parallel. More specifically convergence requires that two populations that share the same nucleotide at a specific position in the genome acquire the same mutation. However, due to the small HIV-1 genome it is possible that most convergent mutations are the result of chance and not selection. Appropriate null models can be exceedingly helpful to distinguish adaptive from neutral mutations (Stayton 2015). Once we have made the distinction between selected and accidental convergence we reserve the term convergent mutation for mutations that occur in parallel in different populations more often than one would expect it under a neutral model."

L85 "private mutations" Please add formal definition here.

We added this in brackets: "(mutations that are found in a single HIV-1 population only)"

RESULTS

Figure 1 On the y-axis: You probably mean "Mean Number ..." right? Otherwise non-integer numbers (10^-2) are a bit weird/unclear. What are the dots (means?) and the lines (standard deviations?). Please add in legend.

Done.

L106 "Whereas ..." To me it rather seems that 1-3 mutations are more or less equal between the two groups.

It is true that the data looks like this, however, this is an effect of the log scale of the y-axis. There is actually a significant difference between model and Keele and Li data. But this is not true anymore after we constrained the dataset to sequences without APOBEC mutations and hence we omitted this part of the sentence from the manuscript.

L118 "In total ..." Is there a list of these 19 candidates? If so please add reference.

There is and we have added the reference to ***Supplementary Table 1****.*

Figure 2 I would switch the two axis such that it rather looks like a Manhattan-Plot. You could then add a line where HIV populations = 5 above which you consider all mutations as potential candidates of selection. I dont think that the black line is needed here. Also, I do not understand the meaning of the red lines here? What is meant by "within one category" (L127)

We have switched the axes and added the following sentence to the caption: "Red bars indicate the mean of all positions of the mutations that occur in the same number of HIV-1 populations"

L127 "position" typo -> positions

Done.

L140 "single population mutations" Do you mean mutations that only appear in a single population?

Yes. We have rewritten the paragraph.

L142 "in only a single..." Do you really mean individual here? It's a bit misleading after you have talked about populations the entire time.

We meant HIV population from a single infected individual (host). The paragraph was rewritten to incorporate the results from ***Figure 2B****.*

L143 "p-Value" typo -> p-value

L164 "Interestingly, we found..." 1. Change in tenses 2. Where is this result shown?

We were talking about the results shown previously in ***Figure 3****. In the new manuscript version the figure changed to* ***Figure 4****. The figure shows that private mutations occur at env positions that are more diverse than we would expect in a random simulation.*

L167 "are under positive selection" But this would also require different selection pressures that you somehow ruled out, right?

We are not sure what the reviewer means with this question. The data we obtained is from HIV-1 virus populations shortly after they infected a new host. In previous studies the evolution from infection to the point where the viruses were sampled (estimated to be less than 50 days) was modeled neutrally. However, during these 50 days one could imagine all kinds of selection to occur. We also do not rule out that some private mutations are also selected for as we now point out in the Discussion of the revised version of the manuscript: "Third, it is possible that some non-convergent mutations are also positively selected. These mutations may be beneficial only in a particular host and hence it would be impossible for our analysis to identify such mutations."

Figure 4 What is happening for HIV Pop = 6? Is this actually discussed somewhere in the text?

Here the synonymous mutations occurred in the part of the env gene that overlaps with the rev gene. Hence, the synonymous mutations in the env

*gene are in this case necessarily non-synonymous mutations in the rev gene, which is why we excluded this region of the env gene in **Figure 4B**. In the revised version of the manuscript we only show the data for mutations that occur in non-overlapping reading frames (**Figure 3**).*

L178 "We obtain a significant ..." Are these results shown somewhere and what does "significant" mean here (which test; p-value)?

*We have now changed the manuscript to show and explain the significance of the data. See the new **Figure 3**.*

L188 "On the other hand..." See comment above.

See above.

Figure 5 I don't think it's necessary to show that wide range (y-axis). Could you zoom in and add more ticks/lines?

Done.

DISCUSSION

L270 "higher than the median..." How meaningful are these diversity values if the distribution seem to span almost the entire range?

We removed the discussion of mutational hotspots, as it is difficult to reliably identify mutational hotspots with the given data.

L279 "Nevertheless ..." Without any idea about epistasis or mutations at other genes -- since these mutations could also just be compensatory mutations for "true" underlying adaptative mutations with pleiotropic side effects, I don't think you can make any statement about effect sizes. Actually, I am also not quite sure that without prior knowledge about potential targets of selection you would be able to make any statements whether convergent mutations are primary mutations (i.e., mutations that actually rise in frequency due to the selective pressure applied such as a special drug treatment) or secondary mutations that compensate for the deleterious side effects of the primary mutations (as often observed for resistance mutations).

It is certainly difficult to make inferences about effect sizes when analyzing convergent mutations without time series. Considering the short amount of time (about 25 viral generations) that has passed since the establishment of the HIV-1 population it is likely that any mutation that rises in frequency to the point that we see it in multiple hosts in parallel in such a short time is a mutation that provides a significant fitness benefit. The reviewer's comment about epistasis and compensatory mutations may not be relevant for the specific dataset we analyzed. It is incredibly unlikely that after a maximum of 50 days or 25 viral generations of evolution (where each population was started from a single virus) that not only another mutation has already

fixed in a different part of the genome (the env gene is usually the first gene to acquire substitutions) but also that the mutations we see are already compensating for a negative effect such a mutation could have had. Furthermore all of the virus populations were sampled before the start of drug therapy. Hence we maintain that considering the short time span of evolution it is plausible that only large effect mutations will be seen in multiple independent populations.

L283 "Finally..." First, and just to point out the obvious: To make use of this approach you need a lot of (independent) data. While it is intriguing that a frequency of 15% of convergent mutations is enough to make statements about potential targets of selection, you clearly would not be able to make these statements if you observe a mutation in one out of 7 populations (i.e., the power rather comes from absolute than from relative numbers). Thus, I would make the first point and also rather state absolute numbers as these relative numbers seem misleading.

We added a statement to the discussion that a large number of independent samples is needed to identify convergent mutations during early infection.

Second, how independent are your "independent samples". Is there a chance that some patients might have been cross-infected which could introduce correlations/relatedness between samples? Relatedly, are those 15% of the strains carrying the most frequent convergent mutation more similar to one another (e.g., in terms of diversity) than the rest (which could maybe be checked with some permutation test)? If so that could be indicative of the mutation being genotype dependent (i.e., epistasis).

The samples are completely independent and all populations have been founded by a single virus during infection. We added this information to the very beginning of the discussion.

MATERIALS AND METHODS

L300 "founder strain" Full stop missing

L300 "those sequences" typo -> sequence

L312 "alignments were performed" Please specify the options used for reasons of reproducibility.

We used the standard setting, and now state this in the revised manuscript.

L322 "Neutral mutation distribution model"

I have several related questions on your neutral model. First, regarding your definition of a convergent mutation. Since you are using different consensus sequences, what if one of the consensus sequences was already carrying the "convergent mutation" (or rather nucleotide)? Would it still be called (and

counted) as a convergent mutation? Or in other words, is a mutational event required for calling something convergent mutation?

> *It is only a convergent mutation if a mutation from the consensus sequence happened. I.e. for it to be a convergent mutation two consensus sequences have to have the same wild type at the same position and this position has to mutate to the same base.*

Second, is your model truly neutral? I was wondering whether using the (upscaled) empirical transition rates in the substitution matrix isn't actually a mutation bias, since you only observe those mutations in your sequences that are not selected against (or filteder by selection for that matter)? Wouldn't equal mutation rates be more congruent with the neutrality assumption?

> *As we have described in our manuscript we think that some of the data is affected by selection. However, we find it unlikely that there is selection against certain types of mutations. Mutation rates are thought to differ due to biochemical constrains. Using the empirically-derived mutation rate matrix leads to a more conservative test for selection. If we were to include equal mutation rates, the result would be much higher rates of convergent evolution because equal mutation rates would increase the effective size of the genome, i.e. in this neutral model it would be much less likely to observe convergent mutations and hence the difference to the Keele and Li data would be very large. As this is unrealistic we inferred mutation rates from the data.*

Finally, maybe a statement about the (simulation) program/software that has been used would be good, as well as putting a file for re-doing the simulations in the SI.

> *Done.*

#

REFERENCES

L378 I am not sure the editor is usually mentioned.

> *Deleted.*

SUPPORTING INFORMATION

SI txt: Maybe some additional text explaining the content of this file would be nice.