



Peer Community In Evolutionary Biology

Is convergence an evidence for positive selection?

Guillaume Achaz based on peer reviews by **Jeffrey Townsend** and 1 anonymous reviewer

Frederic Bertels, Karin J Metzner, Roland R Regoes (2018) Convergent evolution as an indicator for selection during acute HIV-1 infection. Missing preprint_server, ver. Missing article_version, peer-reviewed and recommended by Peer Community in Evolutionary Biology. <https://doi.org/10.1101/168260>

Submitted: 26 July 2017, Recommended: 21 November 2018

Cite this recommendation as:

Achaz, G. (2018) Is convergence an evidence for positive selection?. *Peer Community in Evolutionary Biology*, 100060. [10.24072/pci.evolbiol.100060](https://doi.org/10.24072/pci.evolbiol.100060)

Published: 21 November 2018

Copyright: This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

The preprint by Bertels *et al.* [1] reports an interesting application of the well-accepted idea that positively selected traits (here variants) can appear several times independently; think about the textbook examples of flight capacity. Hence, the authors assume that reciprocally convergence implies positive selection. The methodology becomes then, in principle, straightforward as one can simply count variants in independent datasets to detect convergent mutations. In this preprint, the authors have applied this counting strategy on 95 available sequence alignments of the env gene of HIV-1 [2,3] that corresponds to samples taken in different patients during the early phase of infection, at the very beginning of the onset of the immune system. They have compared the number and nature of the convergent mutations to a "neutral" model that assumes (a) a uniform distribution of mutations and (b) a substitution matrix estimated from the data. They show that there is an excess of convergent mutations when compared to the "neutral" expectations, especially for mutations that have arisen in 4+ patients. They also show that the gp41 gene is enriched in these convergent mutations. The authors then discuss in length the potential artifacts that could have given rise to the observed pattern. I think that this preprint is remarkable in the proposed methodology. Samples are taken in different individuals, whose viral populations were founded by a single particle. Thus, there is no need for phylogenetic reconstruction of ancestral states that is the typical first step of trait convergent analyses. It simply becomes counting variants. This simple counting procedure needs nonetheless to be compared to a "neutral" expectation (a reference model), which includes the mutational process. In this article, the poor predictions of a specifically designed reference model is interpreted as an evidence for positive selection. Whether the few mutations that are convergent in 4-7 samples out of 95 were selected or not is hard to assess with certainty. The authors have provided good evidence that they are, but only experimental validations will strengthen the

claim. Nonetheless, beyond a definitive clue to the implication of selection on these particular mutations, I found the methodological strategy and the discussions on the potential biases highly stimulating. This article is an excellent starting point for further methodological developments that could be then followed by large-scale analyses of convergence in many different organisms and case studies.

References:

[1] Bertels, F., Metzner, K. J., & Regoes R. R. (2018). Convergent evolution as an indicator for selection during acute HIV-1 infection. BioRxiv, 168260, ver. 4 peer-reviewed and recommended by PCI Evol Biol. doi: [10.1101/168260](<https://dx.doi.org/10.1101/168260>)

[2] Keele, B. F., Giorgi, E. E., Salazar-Gonzalez, J. F., Decker, J. M., Pham, K.T., Salazar, M. G., Sun, C., Grayson, T., Wang, S., Li, H. et al. (2008). Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. Proc Natl Acad Sci USA 105: 7552–7557. doi: [10.1073/pnas.0802203105](<https://dx.doi.org/10.1073/pnas.0802203105>)

[3] Li, H., Bar, K. J., Wang, S., Decker, J. M., Chen, Y., Sun, C., Salazar-Gonzalez, J.F., Salazar, M.G., Learn, G.H., Morgan, C. J. et al. (2010). High multiplicity infection by HIV-1 in men who have sex with men. PLoS Pathogens 6:e1000890. doi: [10.1371/journal.ppat.1000890](<https://dx.doi.org/10.1371/journal.ppat.1000890>)

Reviews

Evaluation round #2

DOI or URL of the preprint: [10.1101/168260](https://doi.org/10.1101/168260)

Version of the preprint: 2

Authors' reply, 01 October 2018

[Download author's reply](#)

Decision by [Guillaume Achaz](#), posted 01 October 2018

This preprint worth a revision

The revised version by Bertels et al. shows a considerable improvement when compared to the previous version. It has a better flow and is much easier to read. For this, I would like to congratulate the authors for the effort and work they have put in this revised version. This was worth it. The first reviewer has no further major comment but the second reviewer (reviewer 3 of the previous version) is still unconvinced by the conclusions. I have to confess that I am still myself unsure that the patterns reported here constitute strong support for selective effects, although they can be considered as good clues. I however found that the approach proposed here is clever and is worth delivering to the community. Thus, I think that on top of the major improvements the authors have made so far, some extra work (mostly on writing) is still needed before I can recommend this preprint.

While revising this ms, please keep in mind that:

- The indication for the implication of selection is still weak. Thus I would suggest the authors to lower the strength of their claim. Keep in mind that the *indisputable* pattern you describe here is that your null model does not fit. Rejecting H0 may have other causes than selection.

- The second reviewer rightly points at a confusing argument on the effect of purifying selection (par L215-228). The same pattern (a positive correlation with diversity) is interpreted in one hand as an effect of positive selection for the convergent mutations and at the other hand as an effect of purifying selection for the private ones. I recommend caution.

Personal suggestion for improvement:

- To assess the independence between the mutations (current rev 2), the authors could first test for recombination (using 4-gametes like test or decay in LD or any ρ estimation method) and, if no recombination, built phylogenetic trees with ancestral states reconstruction for each sample (and even use the MRCA sequence to orientate if they include an outgroup). They could then see whether convergent mutations occurred 1 or several times in the samples and eventually test if they hitchhike on each other (please take this only as a suggestion, not as mandatory extra work).
- The remark of the ex-reviewer 2 of the previous version is still valid. Why 10/11 of the non-synonymous convergent mutations are either G->A or A->G. It deserves at least to be reported in the results and discussed in the article. Do you observe the same for the synonymous convergent mutations? If you would assess the expected number of convergent mutations by types of mutations (and not globally) is this still very unlikely?
- The level-off of the decline reported for Figure 1 may be slightly overclaimed (L120). This is based on 11 mutations that cannot be below 1 (while the null model can go well below 1). What do you observe for the synonymous convergent mutations?
- The paragraph L382-L388 needs to be clarified.

On a didactic level a Black&White version of this ms is almost impossible to follow as the colors on the plots look identical. May I suggest that you use filled and empty circles and dashed, pointed and continuous lines on top of the colors (if you like colors) in all figures? Another possibility is to use dark vs light colors.

Typos: - L43: remove 'will' to change the sentence into present time - L411: positions -> position (delete the 's')

To conclude, I think this ms is evolving in a right direction although it still deserves some extra work. I am almost convinced that the next version will be ripe for recommendation. Take all the suggestions of the reviewers as constructive feedbacks (or genuine incomprehensions) and include a point by point response to all comments along with your next version.

Reviewed by [Jeffrey Townsend](#), 02 July 2018

I am satisfied by the comprehensive revisions as performed. A few minor points for consideration:

1) It looks like only the maximum likelihood "model selection" clusters from MACML have been used / displayed. Model selection (linear hot/cold cluster detection) appears to have been informative in this way, but if it was not examined already it is worth mentioning that it may be illuminating to use the computationally intensive model averaging (over "hot" and "cold" spots hierarchically detected) to provide a pseudo-continuous profile of clustering across sites. See flag -m in the MACML user manual.

2) line 36, no "," after "are"

3) line 60, needs "," after "load"

Evaluation round #1

DOI or URL of the preprint: [10.1101/168260](https://doi.org/10.1101/168260)

Version of the preprint: 1

Authors' reply, 22 June 2018

[Download author's reply](#)

Decision by **Guillaume Achaz**, posted 22 June 2018

Revise

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 your manuscript accordingly. A point by point response to their comments must be included along with your revised version of the ms.

Reviewed by **Jeffrey Townsend**, 28 November 2017

[Download the review](#)

Reviewed by anonymous reviewer 2, 28 November 2017

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).

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.

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.

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.

:: 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.

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.

I303-307 - 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).

I309-315 - why doing an alignment with a reference sequence ? (and not only all sequences together without the ref). This seems odd.

I348 - 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.

I220-I223 - Please clarify as it is slightly confusing as it is.