Reply to reviewers round 2

Guillaume Achaz

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.

As far as we can see, the only other force that could cause the pattern we see are biased mutation rates. We are accounting for that possibility by, for example, maintaining substitution rates and excluding potentially APOBEC affected sites, and now added a paragraph in the discussion that explains this issue in more detail. We also provide additional explanations to reviewer 2(3) for why we believe that his concerns do not apply to our approach.

- 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.

After rereading this paragraph we can see how this might be confusing. We rewrote the paragraph and hope it presents the logic of our argument more clearly. We rewrote these paragraphs to say the following:

"Private mutations are mutations that we only see in a single HIV population in our dataset. Ideally these should be a sample of all mutations that occur during the replication of HIV. However, as the sequences replicate for more than one generation selection will play a role (see measuring mutation rates literature Mansky 1996). Selection will act because mutations that for example introduce stop codons in the middle of an essential gene will be lost in the next generation. Those mutations we expect to be underrepresented in our set of 775 env mutations. The distribution of those 775 mutations across the env gene does not seem particularly biased, as clustering approaches have shown (Figure 2).

Other than the location of mutations across the gene we can also measure nucleotide diversity at each site in the env gene. We measure nucleotide diversity in an alignment of all 95 different consensus sequences. For these measures we do not take the mutations that we have observed in the Keele and Li data into account. For comparison to a neutral model, we randomly distribute mutations across the entire env gene and measure the mean diversity across all 775 positions. Interestingly the mean diversity of randomly distributed mutations in 1000 independent simulations was always lower than the mean diversity at the positions, at which the Keele and Li mutations occurred. This means that mutations do not occur at low diversity regions in the Keele and Li data. We propose that this pattern is caused by purifying selection, i.e. when mutations occur at low diversity sites they cause strongly deleterious or lethal phenotypes that will not leave offspring in the viral population and hence will be underrepresented in the Keele and Li dataset."

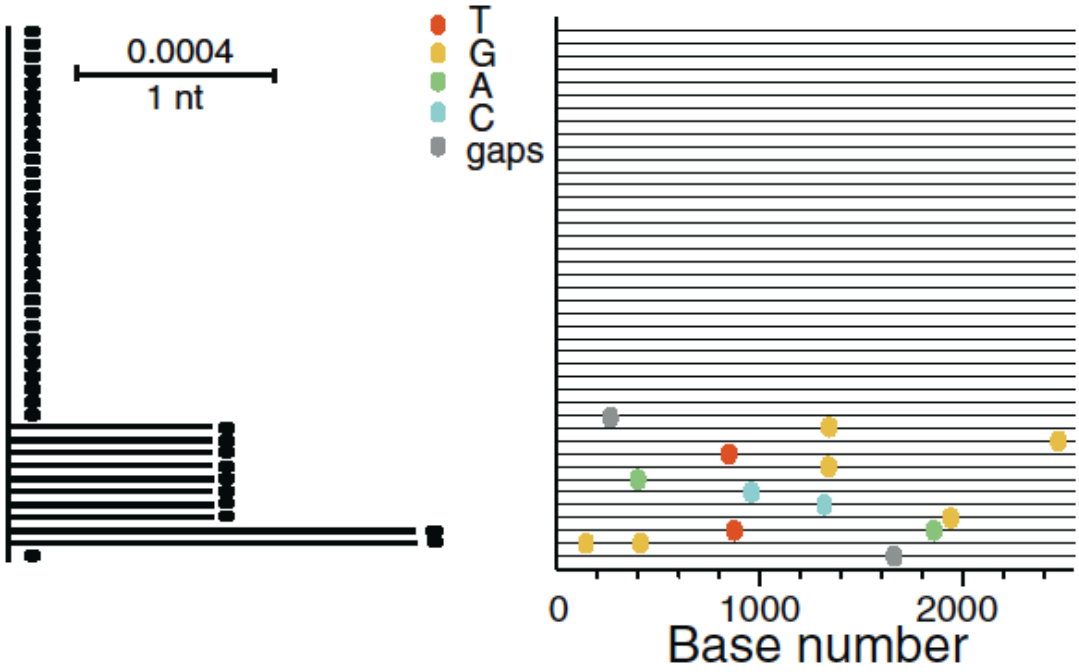
After recalculating the data (due to a small mistake in which sequences we use in our data, previously we left a few hypermutated sequences in the datasets) there was no significant difference to the neutral model for convergent mutations. Nevertheless, if we had observed a difference then this would have meant that convergent mutations also occur preferentially at high diversity sites. The occurrence of convergent mutations at high diversity sites in the env gene cannot be explained with purifying selection. Purifying selection is the elimination of lethal or deleterious mutations from the population. Lethal or deleterious mutations are unlikely to occur in parallel in multiple independent populations. Even neutral mutations are not expected to occur in multiple populations in parallel (see Figure 1 comparison to neutral model). Hence observing convergent mutations at high diversity sites simply means that these sites are not under purifying selection and hence amenable for adaptive changes.

Personal suggestion for improvement:

- To assess the independence between the mutations (current rev 2), the authors could first test for recombination (using 4-gamates like test or decay in LD or any \rho 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).

Because in our current dataset there is a maximum of two mutations per sequence, it is not feasible to detect recombination. Additionally, even if there were recombination, it would not invalidate our point (as discussed in response to comments by reviewer 2). MRCA analyses were performed in the original study that provided the data we analyze by Keele et al. The MRCA was concluded to be the consensus sequence in each of the 95 population samples. In each of the 95 populations, a single virus founded the HIV-1 population.. Below is a figure from the Keele et al. 2008 paper to show an example of a typical sequence set from the 95 env alignments. The left shows a phylogeny of the sequence alignment on the right. As you can see most sequence carry zero mutations, some carry one

mutation and very few two mutations. Furthermore, mutations almost always occur only once in the alignment.



- 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?

  One out of 10 mutations is synonymous (after rerunning the analysis it turned out that there was a mistake and that one of the mutations (G8285A) only occurred in three instead of four HIV populations in parallel). The synonymous mutation is an A to G mutation (in the previous version there were four, three of the four were synonymous in env but non-synonymous in rev). Below are the nucleotide substitution frequencies observed in our data:

AC	AG	AT	CA	CG	CT	GA	GC	GT	TA	TC	TG
0.03	0.2	0.03	0.03	0.01	0.13	0.36	0.006	0.01	0.02	0.1	0.06

  As you can see G to A mutations are the most frequent followed by A to G and C to T. Observing 9 G to A or A to G mutations in a sample of 10 is quite unlikely (we observed 217 in 10,000 randomizations of the data). It is possible that our data is still affected by APOBEC, which further increases G to A mutation rates in certain motifs.

  In order to test what the effect of APOBEC on our data might be we performed our analysis only on sequences that are likely hypermutated, i.e. sequences that contain 4 or more mutations. The following table shows the substitution rates for those sequences.

AC	AG	AT	CA	CG	CT	GA	GC	GT	TA	TC	TG
0.03	0.08	0.02	0.02	0.01	0.06	0.65	0.003	0.01	0.02	0.05	0.04

The table shows that we indeed found that G to A mutations were highly overrepresented (65% of all mutations compared to 36% without hypermutated sequences). Only 30 patients contain sequences with four or more mutations. Yet, mutations are shared in this dataset. The most common mutation is also the most common mutation in our current dataset (G7668A) occurring nine times. The next two mutations occur in four patients in parallel, one of them also occurs in our dataset (G8311A). Of the 10 convergent mutations we identified six that were also found in the hypermutated dataset. Of these six, three occurred in the hypermutated dataset in more than one patient. If we assume that these three mutations are the result of (low) APOBEC activity and not just mutations that occurred in a hypermutated sequence background then we end up with a total of seven convergent mutations. If we now randomly choose mutations with the measured substitution probabilities then we end up with six or more G->A or A->G mutations in 981 times out of 10,000 trials. Hence, G to A or A to G mutations are not significantly overrepresented anymore after excluding potential APOBEC induced mutations (which we did in our main analysis).

We added a paragraph that presents these results (L140).

- 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?

*This point is well taken, and a common issue arising from sparsely populated outcome spaces. The way to deal with this issue is to group outcomes. That is why we say on line 120 that the difference is particularly large for mutations that occur in five **or more** populations in parallel. To make this clear we have added **Supplementary Figure 1**, where we grouped mutations that occur in more than three populations in parallel.*

There is one synonymous convergent mutations, which occurs in five populations in parallel. We have added this information to the manuscript (L132).

- The paragraph L382-L388 needs to clarified.

We have rewritten this paragraph.

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.

Done.

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

Done.

To conclude, I think this ms is evolving in a right direction although it still deserves some extra work. I 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.

Reviewer 1
Jeffrey Townsend

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.

This would indeed be a very interesting analysis. Unfortunately, MACML did not terminate within two days for the small and large datasets, which is when we decided to terminate the program.

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

Done.

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

Done.

Reviewer 2 (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 it in principle allows identifying positively selected synonymous mutations. Furthermore, the authors argue that private mutations -- i.e., mutations

only found in a single HIVpopulation -- are an indicator of purifying selection. Overall the authors conclude that the extent of convergent evolution can be a good predictor of positive selection.
****
GENERAL OPINION AND MAJOR POINTS
I have read and reviewed this paper now for a second time and I still find it a bit peculiar. I might be repeating myself, but here is why: First, of course when doing experiments one tries to replicate the findings to show that the outcome of those experiments are non-random. However, unlike in this paper, there are controlled conditions, a specific hypotheses to be tested, and maybe already candidates (for selection) identified. Here, the logic is very simple. What is unusually common must be selection. Similarly, for the "private mutations". What is unusually uncommon must be (purifying) selection. So everything is selection in the end. Intuitively one would agree with this logic (as this is a requirement for many other experimental approaches as stated earlier), though I feel this approach is ignoring a lot of things (or making some explicit assumptions that are never discussed nor formulated).
If it was selection acting, selection pressures need to be almost identical between populations in order to get convergent evolution.

> *In our study, we present a clear null hypothesis: a neutral model. This null model, which captures convergence by chance, is rejected. Thus, not "everything is selection" in our approach, but we provide evidence for non-neutrality. The non-neutrality we have observed could be due to two factors. Either mutation rates are significantly biased so that mutations preferentially occur at certain positions in the gene or selective pressures are similar in different infected individuals during early HIV infection. We further show that convergent mutations are not evenly distributed across the env gene but instead are clustered towards gp41, the part of the env gene responsible for fusion with the host cell. The concentration of convergent mutations in gp41 is a good indicator that convergent mutations are selected for rather than the result of mutation biases.*

Next, one would have to know what is actually selected for (which phenotype; what is the selective pressure). Just observing changes is not enough to impose selection.

> *We disagree with this opinion. Many studies determine the nature of selection without identifying the selected phenotype, for example, every study that uses a classical dN/dS ratio analysis. With her/his opinion, the reviewer invalidates all these studies. To infer (not "impose") selection we do not only rely on observed patterns of sharing but statistically test for sharing that exceeds the expectation under a neutral model.*

Furthermore, this approach makes a strong "independence" assumption on various levels that are not made explicit, but could be crucial. First, each site/nucleotide is considered an independent target of selection, neglecting any effects due to (physical) linkage and/or epistasis.

> *This is true, we never consider linkage between sites, but considering linkage in this approach is not necessary because only beneficial mutations will rise in frequency in multiple individuals. The other mutations that are potentially*

linked to a beneficial mutation should again be randomly distributed across the genome (behave according to our null model), and hence not be present in multiple individuals. We would like to emphasize that our approach simply looks at repeated occurrence of mutations between populations. We do not consider mutations that occur multiple times within a population (this could happen via linkage or due to a benefit or epistasis etc.). Such mutations will be treated the same as those that only occurs once. To improve the approach one could consider taking mutation frequency data into account however this is not what we have done and would be riddled with problems as the reviewer rightly suggests. We have added a paragraph to the discussion that highlights this point (L362).

Selection could act on a very different gene, and the changes observed are just correcting for the negative side-effects of the true target of selection.

The arguments laid out to the point above apply here, too: by focusing on mutations that are rising in multiple individuals, we identify the true targets of selection. Moreover, the data we analyzed comprises virus populations that evolved for approximately 20 generations in the hosts. During transmission and early infection env is the most relevant gene, because it encodes the proteins involved in target cell entry and fusion. For these reasons "side-effects" are not very likely. One would have to consider incredible fitness effects for such an event to play out in such a short time frame. But even if this scenario were possible I do not see why this matters. We do not claim that the mutations we identified as positively selected for are not compensatory mutations.

Similarly, how do you control for false positives in your approach? When a single selected nucleotide is sweeping through the population it is expected to drag its surrounding nucleotides with it, which would inflate the number of "convergent mutations".

As outlined above, our approach involves a neutral model, and very explicitly tests for "sharing by chance". Any mutation linked to a beneficial mutation should be randomly distributed across the genome. Randomly distributed mutations will not occur in multiple independent populations in parallel. However, there is the possibility that mutations occur in parallel as a result of biased mutation rates. These could be the result of the effect of APOBEC, which causes mutations at certain motifs in the sequence. This is why in the current version of the MS we excluded all sequences that are likely affected by APOBEC (sequences that contain three or more mutations). We also added another paragraph and analysis about this issue in the results section(see above and L140).

Maybe when correcting for this effect, the number of convergent mutations found is not different from that of a neutral model any longer. Given the short time span after infection, it is most likely that there was not enough time for recombination to break up these associations... Crucially though, you do not account for these effects in your "neutral model" when randomizing mutations (because here each site is an independent realization of a mutational event).

We do not model mutational events we simply redistribute the mutations we observed in a dataset, which is exactly how we expect neutral mutations to be distributed. Whether they are linked to a beneficial mutation or not does not matter.

Second, what is the phylogenetic/geneologic relationship between the strains? If say 10 strains have been transmitted from a single strain than seeing some mutations (as defined to a consensus/reference strain) coming up 10 times is not so surprising ("no independent replicates").

The described scenario would be ideal for our approach. Our approach is the most powerful when both virus and host are identical. The more different the virus and host are between replicates the less powerful our approach is. On the one hand this is because the fitness landscape of the virus will change for different hosts and viruses. But there is also a technical reason as there will be fewer and fewer sites that are identical between viruses. Our approach requires that both consensus nucleotides mutate into an identical mutation. Hence for divergent viruses our window for detecting convergent evolution gets smaller.

It is also difficult to say that mutations are enriched in the gp41 region as compared to the rest, when the rest is an identified "low mutation density region" as you write. So it could also very well be that the number of mutations found in the gp41 part is "normal"... To check one would have to compare against another gene (that is neither an identified mutational hotspot or "coldspot").

Mutations are more or less evenly distributed across the env gene (blue line Figure 2). The uneven distribution is only observed for convergent mutations occurring in three or more independent HIV-1 populations. The "low mutation density region" is an effect of positive selection. We changed the wording of the sentence.

Finally, during infection/transmission viruses in general and HIV in particular undergo strong bottlenecks increasing the effects of chance events (genetic drift etc).

The bottle-neck size of all our HIV-1 populations we used was one (see Keele et al. 2008). Hence any chance effect will be present in the entire population and will be part of the consensus sequence.

During the exponential phase, reproductive skew (i.e., the random chance of a single individual to contribute the majority of the offspring in the next generation) can lead to selection like signals that are just caused by random-effects (see for instance Irwin et al. 2016). Also, if there are strains/types able to grow faster during the exponential phase, when there is no competition or any kind of selective pressure, is it really justified to call these selected sites (as you state in line 335)?

During the exponential phase virus strains compete for target cells even if they are not limiting. But the growth kinetics play a minor role for our approach because we do not identify mutations through measuring their increase in frequency. We simply test whether they occur in parallel in independent HIV-1 populations more than expected for randomly distributed mutations.

Regarding the private mutations: I did not at all get why these are a sign of purifying selection? What if these where just sequencing/misorientation errors?

> *For purifying selection we cannot identify individual sites that are under purifying selection as our dataset is not large enough. Nevertheless, we do find that private mutations occur in regions that are more diverse than one would expect them to be if they were randomly distributed. The diversity measures itself stem from the consensus sequences. So sequencing errors should not play a role. If the private mutations are affected by sequencing errors that are not of any biological significance then there is no reason why they should not also accumulate in regions of the env gene that are highly conserved.*

These are all points I feel need to be addressed.

With indirect evidence only -- and I would consider the evidence presented here indirect -- selection is one of many explanations.

> *The only alternative explanation that we can imagine is that mutation probabilities differ extremely across sites as we have discussed above. Also, indirect evidence is quite common in molecular evolution and epidemiology, and we do not share the reviewers rejection of such approaches.*

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

####
ABSTRACT
L19
"convergent mutations provide a selective advantage and hence are positively selected for" Tautologie.

> *We used this wording as a means for greater emphasis and decided not to change it.*

L20
"mutations that are only found in an HIV-1 population of a single individual are significantly affected by purifying selection"
Or just a snapshot of a transient mutation. Or a mutation positively selected for in that genetic background (epistasis).

> *Here we refer to the biased pattern of private mutations that we observed in our data. None of the mentioned mechanisms could account for that.*

####
INTRODUCTION
L36
"Well known examples..."

This is a huge difference! you need to define the target size / scale of ‚Äúconvergent evolution‚Äù. Similarly the above phenotypic examples are very different in their underlying genetics or where they evolved from. I know this is the Introduction, but this is crucial as you can make anything convergent otherwise.

By considering a neutral null model, we correct for "convergence by chance. We therefore strongly believe that "we are not making anything convergent". The examples we give are standard for convergent evolution, and serve the purpose of introducing the topic. They differ from the genetics in our study as much as they differ among themselves. The diversity in genetics across these examples is actually helpful in conveying the abstract concept of convergent evolution.

L41
"virus genomes"
-> viral genomes

Done.

L55
"in line with these findings"
The reference gets lost here. The main information and what you are referring to is actually in brackets. Now it reads as modelling would imply findings...

We changed this so the sentence starts now with "Similarly".

L59
"viral phenotypes"
Make these phenotypes explicit here already ("such as the set point of viral load ...). Note that these should also be defined as this term is probably not clear to a non-expert audience. Since this paper claims to target a general population-genetic audience, it should be defined/explained.

We added: "Typical phenotypes considered include the set-point viral load, which designates the average level of the virus during the chronic disease stage, or disease progression."

L61
"Similarly ..."
So what does that mean? High rates impose fast progression: why is that not a sign of strong selection? The selection pressure is not clear at all. Strong selection to evade the immune response could similarly lead to fast disease progression (once evaded)? Without any detail this information is highly misleading or at least does not explain itself.

We agree with the reviewer that these correlations are not obvious. The data suggests that strong immune selection (which is some type of continuous selection) constrains viral diversification and hence leads to low evolutionary rates. The inverse seems to also be true. We added the following sentence as explanation: "Hence the data suggests that strong immune selection constrains

viral diversification and hence leads to low evolutionary rates, which in turn leads to lower set point viral loads and therefore lower levels of disease progression."

L65
"Selection ..."
I guess the reference method for detecting selection for time-sampled data is currently Foll et al. 2014.

Done.

L 74
"Here ..."
I am very sorry, but yours is not a method. It is exploratory at best. For a methods you should make simulations, assess the power of the method and compare it to established methods.

The sentence the reviewer refers to does not contain the word "Method". Thus, we think any disagreement of what constitutes a method is irrelevant here.

L82
"... selected and accidental convergence ..." So in your case convergence implies selection. What's with all the nucleotides that do not mutate?

In our approach we test for sharing by chance. Not mutated sites are not considered shared. Only the same mutation (i.e. from the same original nucleotide to the same target nucleotide, e.g. G8552A) is considered in our analysis.

L86
"random null model" This model is crucial here. If the null is wrong, the alternative does not necessarily imply selection.

The only alternative explanation for more extreme patterns of sharing than expected under a neutral model is, in our opinion that mutations rates differ across sites. We took the utmost care to ensure to exclude this alternative possibility.

L90
"This biased distribution..."
What about the number of false positives? What if all these are linked and the effective number of convergent muts is way lower and indistinguishable from a neutral process?

See explanation above.

L92
"In contrast ..."
I dont get your point here (the argument why, not what you are implying).

I hope we made it clear in our comments above.

####
RESULTS
L121
"For example, there ..."
Can you rule out correlations (e.g., same founder populations; relatedness of the virus between infected individuals)?

See explanation above.

L130
"We expect about 12..."
So what is the prob of observing 19 then?

The probability of observing 19 or more mutations that occur in three or more individuals is ~ 0.035. We changed the sentence to: "In 35 out of 1000 neutral models we observed 19 or more mutations occurring in parallel. Hence, probably some of the 19 mutations are selected for the exact identity and number of these mutations is more difficult to determine."

L210
"A single ..."
I guess you want to say that these share a most recent common ancestor some time ago. Not that all individuals have been infected with the same virus?

That's true. We changed the sentence to say: "Each of the 95 HIV-1 populations was founded by one individual virus"

L225
"Absence of mutations..."
I dont get your point here.

Mutations are rare in low diversity regions of the env gene. We measure the diversity for each site in the env gene by determining the Shannon entropy for each nucleotide site from a multiple sequence alignment of the 95 consensus sequences of the 95 individual HIV-1 populations.

####
MATERIALS AND METHODS
Generally, the underlying assumptions and their implications need to be discussed somewhere.
L300
"those sequences"
typo -> sequence

The reviewer might have shifted to a different version of the MS. We already implemented the suggested changes mentioned here and below as well as answered the reviewer's questions.

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

References

Mansky LM. 1996. Forward mutation rate of human immunodeficiency virus type 1 in a T lymphoid cell line. AIDS Res. Hum. Retroviruses 12:307–314.