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

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?

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.

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.

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).

####
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.

L41
"virus genomes"

-> viral genomes

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...

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.

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.

L65
"Selection ..."

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

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.

L82
"... selected and accidental convergence ..."

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

L86
"random null model"

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

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?

L92
"In contrast ..."

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

####
RESULTS

L121
"For example, there ..."

Can you rule out correlations (e.g., same founder populations; relatedness of the virus between infected individuals)?

L130
"We expect about 12..."

So what is the prob of observing 19 then?

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?

L225
"Absence of mutations..."

I dont get your point here.

####
MATERIALS AND METHODS

Generally, the underlying assumptions and their implications need to be discussed somewhere.

L300
"those sequences"
typo -> sequence

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.