Response to reviews of "Mapping mutational effects along the evolutionary landscape of HIV envelope" for *eLife*

Hugh K. Haddox, Adam S. Dingens, Sarah K. Hilton, Julie Overbaugh, and Jesse D. Bloom

February 26, 2018

Below, the original comments are in blue, and our responses are in black.

SUMMARY

Information on the possible amino acid trajectories of the rapidly evolving HIV envelope protein has long been limited. In this manuscript, you extend previous work from your lab analyzing the mutational landscape of the HIV-1 Env gene during in vitro viral replication. The work focuses on two clade A transmitter founder viruses, BG505 and BF520, which differ by more than 100 mutations. A deep mutational scanning approach is used to identify the permitted mutations at each residue of Env. Of particular interest is how amino acid preferences differ for these two proteins. A set of 30 sites with clear shifts in amino acid preference are identified. These data are compared to known mutational constraints on Env and to one another, and sources for the shifts in mutant affects are proposed by eliminating many scenarios. The results offer a body of knowledge useful for identifying i) residues and networks that are critical for basic function and ii) hot spots for evolvability or immune pressure.

The analysis has been performed to a high standard (though some clarifications are necessary; see minor comments), and the openness of the data is a great service to the community. The paper represents a significant contribution towards understanding the structure-function relationship of HIV envelope. Therefore, the paper is likely to be accepted for publication if the following points are adequately addressed.

Thanks for the excellent summary of our work. We appreciate the favorable evaluation of the rigor of the work and the importance of the data. The comments below are extremely helpful, and we have revised the manuscript in line with these suggestions. These revisions have further strengthened the manuscript.

ESSENTIAL REVISIONS

1. How are mutations that strongly modulate Env's expression level taken into account?

This is an important point. Our experiments directly measure how mutations affect viral growth. This approach has many advantages over simple surface display approaches, since it is more relevant to the true function of Env. However, a caveat is that viral growth is a convolution of Env function and expression, so our experiments cannot separate these two phenotypes.

We have taken steps to limit simply finding mutations affect expression. We have done this by *not* mutagenizing the signal peptide and cytoplasmic tail, which are the most common sites of mutations that affect expression levels.

We have added text to make this point clearly, and emphasize the associated caveats. When we first describe the deep mutational scanning approach, we now say:

> *However, in our experiments we mutagenized only the ectodomain and transmembrane domain of Env, and excluded the signal peptide and cytoplasmic tail. The reason is that we measure how Env mutations affect viral growth, which is influenced both by the functionality of Env protein molecules and their expression level. Mutations in the signal peptide and cytoplasmic tail commonly affect Env expression level (Chakrabarti et al, 1989; Yuste et al, 2004; Li et al, 1994), so we excluded these regions with the goal of reducing the degree to which we simply identified mutations that affected Env expression.*

Then when we describe the results, we clearly emphasize how expression level (like Env protein stability, which was already discussed) could have general effects on the results:

> *Differences in Env's expression level might also contribute to a general broadening or narrowing of tolerance to subsequent mutations. Because our experiments select for viral growth—which is affected by both Env function and expression—it is possible that some of the shifts are due to epistatic mutational effects on expression rather than function.*

2. Only the native, pre-fusion Env trimer is used to calculate proximity of substituted and shifted residues. It is possible that substitution-shift colocalization may be detected in other conformations of Env as it undergoes structural rearrangement to mediate fusion, as opposed to acting through longer-range allostery in the pre-fusion conformation. Another well-characterized structure that should be considered is the post-fusion six-helix bundle of gp41.

This is an excellent suggestion. In the initial paper, we only analyzed the closed pre-fusion conformation of the Env trimer. We now analyze substitution-shift co-localization

in two other conformations: the open CD4-bound conformation and the post-fusion six-helix bundle. In the end, our basic conclusions remain the same, adding greater support to the hypothesis that epistasis is acting via long-range interactions.

First, we investigated the open CD4-bound conformation. Figure 7D statistically tests the hypothesis that shifts are larger at sites that have substituted or sites that contacts substitutions. Initially, this figure only showed the results of this test for the closed pre-fusion conformation. Now, it shows the results for both the closed and open conformations. Both conformations gave the same general result: there is no significant trend for shifted sites to contact sites of substitutions. We also considered both structures at once, classifying sites as contacting substitutions if they do so in at least one structure. Even then, we did not find an association (see the new right-most panel in Figure 7D). Thus, our investigation of the open conformation re-enforced our initial conclusion that many of the shifts are likely due to long-range, rather than short-range, epistatic interactions.

We also repeated the analyses in Figure 7B and Figure 7C with the open conformation. Our initial conclusions about the solvent accessibility (Figure 7B) and clustering of shifted sites (Figure 7B) also held for the open conformation.

Next, we investigated substitution-shift co-localization in the post-fusion six-helix bundle. Only ∼80 residues per Env monomer are resolved in available structures of the six-helix bundle. The structure with largest number of shifted sites (PDB: 1ENV) only contains four shifted sites and five substituted sites. These numbers are too low for the statistical tests used in Figure 7. But we have added a new figure (Figure 7 - Figure supplement 2) that qualitatively analyzes substitution-shift co-localization in the post-fusion six-helix bundle. Three of the four shifted sites cluster at one end of the helix, along with one of the five substitutions. However, this cluster is also present in the closed and open states as well (Figure 7 - Figure supplement 2). Thus, the post-fusion six-helix bundle did not reveal any new patterns of co-localization that were not also present in the other states.

We now describe the results of the analyses in the first two paragraphs of the Results subsection "Structural and evolutionary properties of shifted sites", interleaved with the original text in these paragraphs.

In the process of making these revisions, we discovered a minor bug that affected our calculations of residue-residue distances. We fixed that bug in the revised version. None of the main results change, but some of the quantitative values in the box plots and the corresponding P-values have changed slightly.

3. Overall, it is found that the shifted sites tend to cluster together, but they are not necessarily the same as the sites with substitutions between the two Envs or ones that contact them. This analysis is not pursued in great detail, however, leaving the reasons why shifted sites tend to cluster together unexplained. Given that the authors also observe some evidence of entrenchment of substitutions between the Envs, is it possible that

interactions between the shifted sites themselves are important? This question could be explored through a more detailed examination of the structure.

To address this comment, we examined the clusters of shifted sites in greater detail. Two clusters occur within conformationally dynamic regions of Env. One hypothesis as to why shifted sites cluster in these regions is that their conformationally dynamic nature allows long-range epistatic interactions to be propagated between shifted and substituted sites that are distant from one another.

We have added a new supplemental figure (Figure 7 - figure supplement 3) that provides a fine-grained structural view of each of the clusters. We have also added an entire paragraph detailing these observations at the end of the section in the Results called "Structural and evolutionary properties of shifted sites". Overall, these additions strengthen the manuscript by providing a better structural intuition for how long-range epistatic interactions might occur in Env.

MINOR POINTS

1. Readers may miss the rationale for excluding the signal peptide and cytoplasmic tail if it is only found in the Methods.

This is a good suggestion. We now also explain this rationale in the first part of the Results as discussed in the response to Major Point #1 above.

2. Page 3: When referring to the modifications made in this paper versus the previous publication, it is not clear whether the cell type difference was expression of CCR5 or use of a new cell line altogether.

We have clarified that we used the SupT1.CCR5 cell line to allow for infection with viruses bearing CCR5-tropic Envs. This cell line is a SupT1 clone that was lentiviral transduced to also express CCR5 in addition to CXCR4 and CD4, which SupT1 cells express naturally. In our original work with CXCR4-tropic LAI Env, we used SupT1 cells.

3. Page 6: The dN/dS parameter is introduced, but its meaning in the two different models is only described in the next paragraph. Noting that a description follows may help readers.

This is a good suggestion. We have re-structured the relevant text (second and third paragraph of third subsection of Results) to expand the explanation of the dN/dS model parameters and put them in a more logical order.

4. The color scheme of structures in Fig 7a, 9a is difficult to discern. Higher contrast colors please.

We have now changed the color schemes to be higher contrast in both of these figures. The one in Fig 7a now goes from white-to-red instead of grey-to-red. And the one in Fig 9a now goes from blue-to-white-to-red instead of blue-to-grey-to-red.

5. The methodology used here was previously developed in your group, which would benefit from a short description in the main text. For example, how exactly is ω defined, what is the stringency parameter, can the use of ω be motivated, how is ω_r defined and motivated etc? What are the assumptions in using these measures? We think this is necessary for a stand-alone paper aimed at a broad audience. All the details of the method should not be repeated, but an intuitive summary in the main text would be helpful.

This is good suggestion, and related to Minor Point #3 above. We have re-structured the relevant text (first few paragraphs of third subsection of Results) to substantially expand discussion of these models and parameters. We have also expanded the discussion around ω_r in the last subsection of the Results and in the figure / table legends.

6. With the current version 2.2.0 of dms_tools2 installed via pip today (January 19, 2018), the included iPython notebook throws an error. Error text is below:

```
---------------------------------------------------------------------------
ModuleNotFoundError Traceback (most recent call last)
in <module>()
36 import dms_tools2.dssp
37 import dms_tools2.compareprefs
---> 38 import dms_tools2.protstruct
39 import dms_tools2.rplot
40 print(dms_tools2.rplot.versionInfo())

ModuleNotFoundError: No module named 'dms_tools2.protstruct'
```

This is now fixed. Previously we were running the analysis with a developmental version of `dms_tools2` (version 2.2.dev1) that was not yet on `PyPI`. Now the analysis uses a stable version (2.2.4) that is on `PyPI` and installable with `pip`. Therefore, the analysis should now work if `dms_tools2` is installed with `rplot` as described here: `https://jbloomlab.github.io/dms_tools2/installation.html#installing-the-rplot-module`. Note that some other external Python packages are also needed as indicated by the imports at the beginning of the iPython notebook.

7. It might be helpful to increase the transparency on the points in Figure 3B, because at the current level it's impossible to resolve any detail in the blob near zero.

We have added a new figure supplement (Figure 3–Figure supplement 2) that shows the correlations as a KDE contour plot rather than a scatter plot. As is apparent from this new supplement, the vast majority of points are near the origin (low preference in both

replicates). This is expected since most mutations are highly deleterious. Therefore, the correlation signal is due to relatively small fraction of amino acids that have non-negligible preference—a fact that makes sesne, since most mutations are so deleterious as to be evolutionarily irrelevant. Unfortunately, given the skew of the data it is not possible to find a transparency that strikes a good balance between showing the high-preference points clearly and not over-plotting the low preference points. Therefore, we have elected to keep the original Figure 3B (scatter plot) while also adding the new contour plot in Figure 3–Figure supplement 2) to show both elements of the data.

8. In Figure 4, it's not clear why the p-value for ω greater/less than one is used for shading rather than the estimated ω itself.

We have added text in the legend of Figure 4 and the "Alignments and phylogenetic analyses of Env sequences" subsection to explain this. Briefly, for individual sites, the point estimates of dN/dS values (or ω_r) are not reliable when using a FEL framework for the reasons described in Pond et al (2015) and Murrell et al (2012). Therefore, the P-value is a better indication of the strength of the statistical evidence than the value of ω_r itself.

Here is the basic intuition. Imagine two sites: one has 3 nonsynonymous substitutions and 0 synonymous substitutions, while the other has 20 nonsynonymous substitutions and 2 synonymous substitutons. If we just calculated ω_r, we would get a value of infinity for the first site and 10 for the second site. But in reality, we think there is substantially stronger evidence for diversifying selection at the second site, since there are many more nonsynonymous substitutions. In an FEL framework, we assess the statistical evidence that $\omega_r \neq 1$, and find much more evidence for the second site. This increased evidence is manifested by the larger P-value rather than the value of ω_r itself.

9. What is the rationale for rescaling amino acid preferences by the stringency parameter before comparing them? Based on (Hilton et al., 2017), my understanding is that only the amino acid preferences are measured directly. The stringency parameter appears to be based on a phylogenetic model fit. This rescaling would emphasize differences in the basic ordering of amino acid preferences over differences of intensity, but the latter would seem to also be important.

In this comment and the one immediately below (#10), the reviewers point out the need to both justify why the preferences were re-scaled, and to clearly explain that this was done.

In the Results section where we first discuss the re-scaling, we have now added an entire paragraph that discusses this fact:

> *For the entire rest of the paper, we use the experimentally measured preferences re-scaled by the stringency parameters in Table 1. The reason we do this is*

to distinguish genuine differences between the two Envs from mere variation in the strength of selection between the two sets of experiments. Re-scaling both sets of preferences to optimally describe Env evolution in nature is a principled way to standardize the measurements; see Hilton et al (2017) and the Methods section entitled "Re-scaling the preferences" for a more detailed explanation.

At other key points (such as Figure 6 and the bottom of page 18 in the original manuscript) we now also emphasize the fact that the re-scaled preferences are used.

The detailed justification for re-scaling is provided in the new Methods section entitled "Re-scaling the preferences." Briefly, the preferences themselves are calculated from the enrichment of mutations after selection. But the extent of enrichment is a function of how the experiment was performed (e.g., how long the selection was performed) as well as the inherent effects of the mutations. For instance, let's say a given mutation is depleted 3-fold after 2 rounds of viral replication. Then if we instead allowed $2^2 = 4$ rounds of viral replication, then we would expect the mutation to be depleted $3^2 = 9$-fold. Although we did our best to standardize the experiments with the two different Envs (BG505 and BF520), it is impossible to completely standardize experiments with different viruses since they grow at different inherent rates. Therefore, we need some method to ensure that we have put the measurements for the two Envs on the same "scale." Re-scaling by the stringency parameter is a principled way to do this. See the new Methods section "Re-scaling the preferences" for a justification of why this is a principled approach.

10. In the text (p8, bottom) it is stated that the amino acid preferences are rescaled for comparison. But this is not mentioned in Figure 6 or described clearly in the Methods (p18, bottom). For the sake of clarity, it would be helpful to introduce new notation for rescaled preferences and indicate clearly where they are used.

This is an important point that we have addressed. Se our response to comment #9 immediately above for a description of the changes that we have made.

11. In the discussion it is noted that: "But although there is some entrenchment of differences between BG505 and BF520, this is not the major factor behind the shifts in amino-acid preferences: most sites that have shifted between BG505 and BF520 actually have the same wildtype amino acid in both Envs." Does wildtype refer to the amino acids in the actual BG505 and BF520 sequences? The most preferred amino acid is the same in only 12/30 cases, based on Figure 6C.

This was unclear in the original manuscript. We have updated Figure 6C to also show the wildtype amino acid at each site. This makes it clear that even though the most preferred amino acid is only the same in 12/30 cases, the wildtype amino acid is the same in 18/30 sequences. This is because although the wildtype is always among the most preferred amino acid, it is not always the most preferred one. We have updated the sentence in the Discussion to explain this, and also to say the "majority of sites" rather than "most sites"

since the former probably is a more accurate description of 18/30.

12. A body of computational work that has been done to infer the effects of epistasis on HIV proteins (including ENV) is not adequately referenced.

This is a good suggestion. Indeed, the first paper to make an in-depth attempt to parameterize an epistatic fitness landscape of Env was just published as our manuscript was under review (Louie et al, 2018). We now cite this paper as well as some related ones.

In the Introduction, we note this body of work:

> *Epistasis among a few combinations of Env mutations has been experimentally demonstrated (da Silva et al., 2010), and epistatic fitness landscapes have been computationally inferred for a variety of HIV proteins (Kouyos et al., 2012; Ferguson et al., 2013; Mann et al., 2014; Barton et al., 2015) including Env (Louie et al., 2018).*

Then in the Discussion, we raise the intriguing possibility that our data might help inform these computational approaches:

> *Our large-scale datasets of mutational effects in multiple viral strains should be useful for efforts to computationally parameterize "fitness landscapes" of Env (Kouyos et al., 2012; Ferguson et al., 2013; Mann et al., 2014; Barton et al., 2015; Louie et al., 2018).*