**Response to Editor’s/Referee’s comments for:**

On the origin and structure of haplotype blocks

Daria Shipilina, Arka Pal, Sean Stankowski, Yingguang Frank Chan, Nicholas H. Barton

Editor Comments to Author:

I would like to apologize for the time it took to provide feedback on this manuscript. It experienced significant delays during review. I therefore had to move to a decision based on the feedback from only one of the initial reviewers (which was very thorough). I do thank the authors, however, for addressing all the previous comments. I believe the revised version of this manuscript is much improved. I particularly like the inclusion of the practical example based on the Heliconius dataset. I think this addition helps to better make the case with regards to the utility of implementing methods that reconstruct the ARG. Also, I commend the authors for discussing in more detail the tools available for ARG inferences, and the tradeoffs involved in terms of accuracy and computational demands. I encourage the authors to address the last (relatively minor) issues highlighted by the reviewer. I add a small number of additional comments below.

We are grateful to the Editor and the reviewer for the very positive feedback on our manuscript, we now implemented the suggested changes.

1. Was Figure 1 not included in the current revision? The figure caption is still present, although the actual figure appears to be missing.

Figure reattached

2. Line 771: the Figure 2 caption contains a typo "of thein the genealogy".

Fixed

3. Please use either "Figure" or "Fig.", throughout the manuscript. For example, line 116 lists "Fig. 2", whereas line 125 lists "Figure 2"

We now refer to figures as “Fig.”

4. Lines 126-127: this is true if one limits a STRUCTURE analysis to the "uppermost" level of genetic structuring, rather than conducting a hierarchical analysis. Presumably, multiple rounds of such analyses aiming to detect more shallow levels of structure will recover variation generated by coalescence and recombination events. Thus, consider removing or tempering this statement.

This statement is now changed

5. Two copies of Figure 3 were uploaded (one with "Blocks" and one with "Edges")

Current version of the figure attached

6. the X axis of Figure 5 spans ~3kbp (1386 -> 1389), yet the text (line 350) indicates this focal region spans 100kbp.

We now clarified this sentence: 100kbp was mentioned in a context of the location and not the span of the focal region. The sentence now reads: “Fig. 5 shows a focal region (~ 3 Mb long) located approximately 100 kbp 3’ from optix that may correspond to a distal regulatory hub controlling the distinctive wing rays…”

7. Lines 400-401: would the multiple selective sweeps not be easy to identify using just SNP data? How exactly does the reconstruction of the ARG help in this situation?

We argue that for the situations where selective sweeps are occurring in the proximity of each other (ex. flower color loci in *Antirrhinum* <https://doi.org/10.1073/pnas.1801832115>), inspecting the haplotype block structure could be beneficial. It will allow to track the origin independently. We clarified this point in the text.

8. Line 409: revise to avoid repeating "only" twice in the same sentence.

Revised

9. Line 419: "argweaver" should be changed to "ARGweaver".

Changed

Subject Editor, Molecular Ecology

Reviewer Comments to Author:

Reviewer: 1

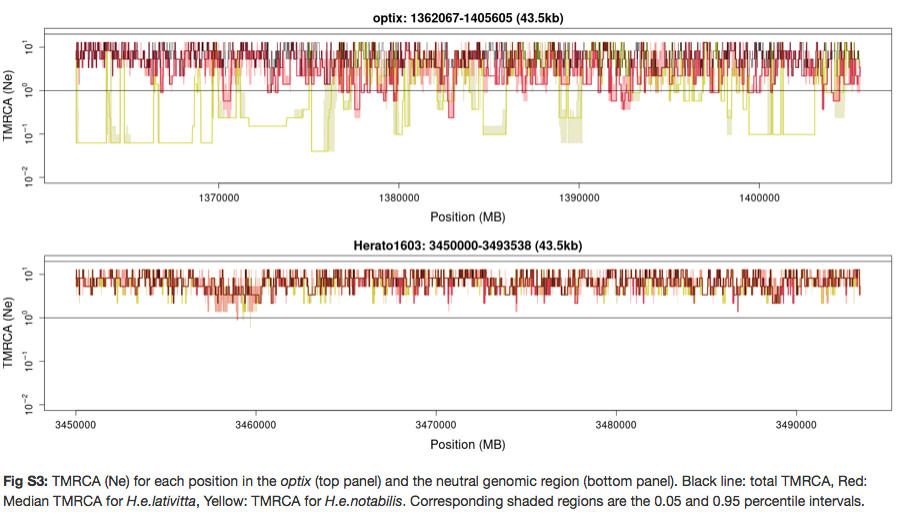
Comments to the Author

Shipilina, et al have made substantial revisions to their manuscript, in the form of clarifying terminology, additional background information about ARGs and related population genomic concepts, further explanation of the simulations conducted, and discussion of an example empirical data set. The additions further clarify the intent of the manuscript and will make it accessible to a wider audience. Box 2 serves as a useful link between methods likely to be familiar to a wide audience and the tree-based concepts described in this manuscript. The main and supplementary text about the simulations together will allow readers to much better understand how simulations were conducted and follow along in detail. Further, the empirical example provides a crucial link between the theoretical discussion of the proposed definition of haplotype blocks and its potential use in population genomic studies. The discussion of strengths and weaknesses of current ARG inference methods beginning on line 397 is also useful and thorough.

We thank the reviewer for the evaluation of the revised version of the manuscript and a positive feedback.

My only concerns with the revised manuscript relate to the new figure and empirical data analysis. I could not find the referenced figure (S3) showing ARGs sampled from a neutral locus, which would be useful against which to compare the data from Figure 5. But some text contrasting these two loci could be added to the paragraph on lines 367-396 to help readers imagine how they might apply this haplotype block concept to detect selection. Beyond this, this section contains useful further discussion of the ARG-based haplotype block concept, and importantly, how it applies to ARGs inferred from empirical data sets and its possible limitations.

We apologize for the confusion with the numbering of the supplementary figures and materials. Figure S3 is part of the Supplementary 2 “ARGweaver analysis”:

****

We explain the difference observed between neutral and selected regions at the end of the corresponding paragraph.

Other than this, I have a few specific points:

Line 349-350: It would be useful to include the reference genome coordinates of the region shown in the figure, either here or in the figure itself.

Figure 5 shows genomic coordinates of the reference genome (from ~1.386 to ~1.389 Mb on the chromosome 18. We additionally clarified this sentence (please see response to Editor).

Line 371: “largest haplotype block” – I would clarify here what is meant by largest (I think it refers to the number of tag SNPs, but one could imagine size referring to genomic span, number of SNPs, or number of genomes).

This sentence is now clarified

Line 376: “shallow TMRCA region” – this seems to refer to Fig. 5D, but this is not clear. I’m confused about which TMRCA is plotted here (see Line 888 comment below), and it seems like low-TMRCA regions correspond to the sweep, but this could be stated explicitly if the authors meant to highlight this.

Clarified

Line 378: “fixed SNPs” – it was confusing to me at first that the SNPs in panel B refer only to H. e. lativitta, and either more text clarifying this in the main text or some indication in the figure itself of which panels refer only to H. e. lativitta samples could be useful.

We highlighted population in focus in the main text and in the caption to figure 5.

Line 442-444: My understanding is that Relate also places SNPs on branches (i.e. see “.mut” file description on https://myersgroup.github.io/relate/getting\_started.html), but current wording suggests that this is not the case.

We changed the wording to reflect that it is a common feature in both softwares

Line 882 (Fig 5 caption A): how were these genomic points chosen?

The genomic points were chosen specifically to show branches on marginal trees that correspond to the coloured edges. We have included this sentence in the Figure 5 caption A.

Line 888 (Fig 5 caption D): 50-percentile TMRCA estimates – is this the same as the “relative TMRCA halflife” statistic from Rasmussen (2014)? – these look like a useful metric for highlighting where the swept blocks exist, but slightly more clarifying text would be useful.

Yes, this is the same approach, we clarified that.

Figures in the supplementary PDF appear to be cut off – possibly due to different formatting in the notebook vs. PDF rendering (e.g. see page 8).

Supplementary materials are now converted to pdf manually to avoid rendering problems