Answer to Recommender and Reviewers

Dear authors,

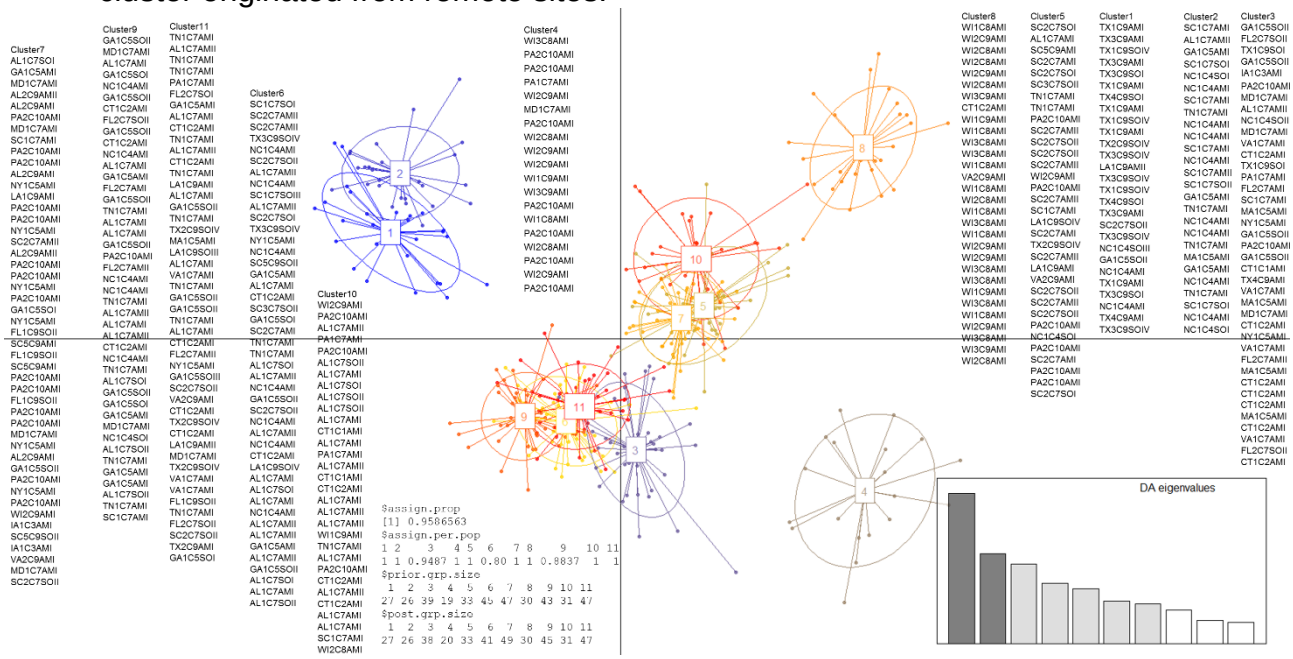
You have addressed satisfactorily most points from the first two reviewers, and I am pleased to say that the paper has gained in clarity. The Figure 6 has proven to be very effective in summarizing the steps of data curation. However, as you want to stay with the format of a short communication, the study still reads quite narrow in focus and appears as a specific problem arising from this particular dataset.

The lack of generality of the study is highlighted by the new reviewer 2. This reviewer has suggestions which would require work beyond the scope of a short communication namely 1) to conduct in depth study of spatial structure/past demographic history (emphasize the biological results), or 2) perform a simulation study (emphasize the methodological results). As you have rebuked my suggestion to perform such additional work after the first round of review, I will not insist.

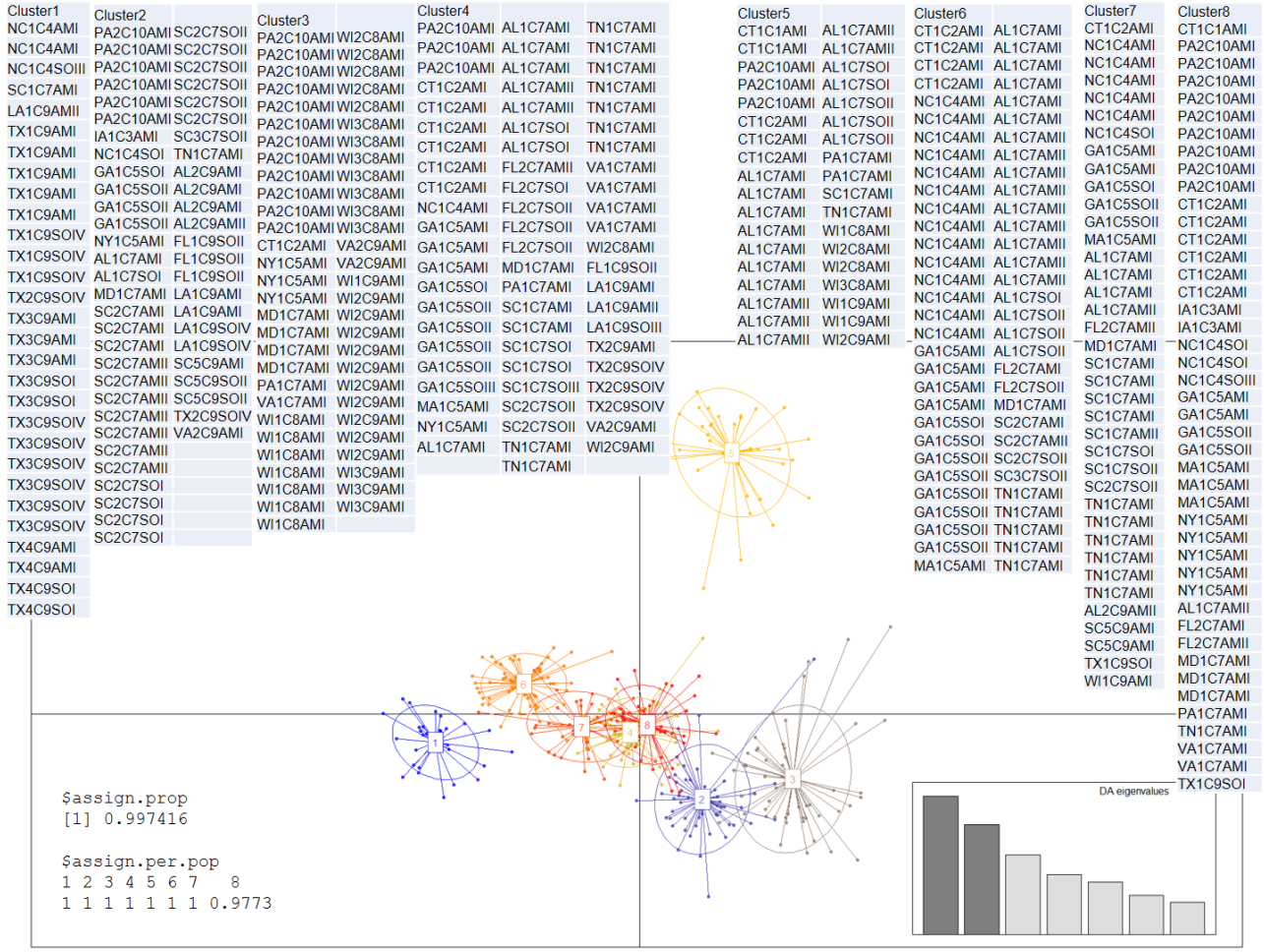
I nevertheless recommend to add one paragraph and one figure of population structure analysis with one of the classically used software as suggested by the new reviewer 2. These new results and the comparison to the Fst/Fis computed values can thus be discussed and provide additional evidence for the strong population structure in this species. This would reinforce and clarify the biological conclusion of the paper. Such addition would be also valuable to enlarge the conclusion of the paper, for example as a warning/word of caution on the influence of data curation on results obtained by classic methods (structure,...). To avoid the multiplication of figures, a possibility for a short communication article could be to group Figure 2, 3 and 5 in a single multiple panel figure.

Answer: As we have discussed this (see below), we do not see what Bayesian clustering approaches, with all their known problems, would bring to the F_{IS}/F_{ST} analyses we already undertook. Clustering techniques are fine to detect a Wahlund effect, which is not the case here. Structure can be very helpful to estimate the race or species assignment of different individuals of a population, but this is not what we are looking for here. With the F_{IS}/F_{ST} approach, we estimated Nm=1 and N_e=7, which, to our point of view, are enough to consider that this tick population is strongly subdivided. I (TdM) have recurrent experience with Bayesian clustering and I am left with the uncomfortable feeling that nobody controls what the software that handle those procedures really do, and to what biological entities the clusters obtained really correspond. I am thus really reluctant in using such algorithms unless I cannot avoid it, e.g. when the data are obviously affected by a Wahlund effect, in which case Bayesian clustering can prove useful (see for instance what we did in (Manangwa et al., 2019)). Nonetheless, if such analyses appeared mandatory, we would undertake them, but with no real hope that they will bring anything new. We would also take care to conduct analyses on contemporaneous subsamples only, with the cured data set and excluding loci with suspected null

alleles, because none of the available clustering algorithms control for temporal heterogeneities or amplification problems. In the following figure you can see a DAPC graphic, based on cured data, which provides quite a strong "structure" (mean assignment is 0.96), but, even if some geographic concordance can be noticed (Cluster8 is mainly Wisconsin), many individuals that belong to the same cluster originated from remote sites.



A new analysis provided a better partition (mean assignment is almost 1) but with a different partition and the same problems of gathering individuals from remote sites:



How are we supposed to interpret such outputs? What are these clusters about? Personally, I cannot think of a convincing biological interpretation. Note that in (Manangwa et al., 2019), we did not use the clustering of any partition, but we used the obvious two clades that happened in all Bayesian clustering analyses that clearly separated two clusters or two collections of clusters (depending on the analysis). We can add such figures as those above, together with the others that would be generated by Tess and STRUCTURE or even BAPS (why not?), which certainly would be different from one another for no obvious reason. Nevertheless, I seriously doubt it would help clarify the main results of our paper. In fact, I believe it would render it much more difficult to read.

Providing this additional result part and adequate reply/changes to the last minor comments by both reviewers, I believe that the article should qualify for acceptance in PCI Evol Biol in the very near future.

Best regards, and looking forward to the hopefully last version of the manuscript.

Aurelien Tellier

Additional requirements of the managing board:
As indicated in the 'How does it work?' section and in the code of conduct, please make sure that:
-Data are available to readers, either in the text or through an open data repository such as Zenodo (free), Dryad (to pay) or some other institutional repository. Data must be reusable, thus metadata or accompanying text must carefully describe the data.

Answer: We have amended the typo of the link to raw and cured datasets at the end of the manuscript. The correct link is http://www.t-de-meeus.fr/Data/DeMeeus-et-al-SAD&StutteringI-scapularisUSA-PCI-EvolBiol-TableS1.xlsx. Is it ok?

-Details on quantitative analyses (e.g., data treatment and statistical scripts in R, bioinformatic pipeline scripts, etc.) and details concerning simulations (scripts, codes) are available to readers in the text, as appendices, or through an open data repository, such as Zenodo, Dryad or some other institutional repository. The scripts or codes must be carefully described so that they can be reused.

Answer: All necessary information is presently in the manuscript

-Details on experimental procedures are available to readers in the text or as appendices.

Answer: All necessary information is presently in the manuscript

-Authors have no financial conflict of interest relating to the article. The article must contain a "Conflict of interest disclosure" paragraph before the reference section containing this sentence: "The authors of this preprint declare that they have no financial conflict of interest with the content of this article." If appropriate, this disclosure may be completed by a sentence indicating that some of the authors are PCI recommenders: “XXX is one of the PCI XXX recommenders.”

Answer: Done.

Reviews

Reviewed by Martin Husemann, 2019-08-03 09:50

Dear colleagues,

I have reviewed this paper before and find that the authors in their revisions have addressed most points satisfactory. I still think that the sample size for the subpopulations is rather on the low site, but the authors have ample experience and hence I believe their judgement. One thing I found a bit strange is that the authors consider the change from 22% of loci in LD (prohibitive) to 19% (reasonable) such a large difference. It seems rather minor, but certainly there is an improvement.

Answer: We agree with the referee that the term "prohibitive" was inaccurate and that 22% is not significantly different from 19%. We replaced "prohibitive" by "important". What is significant is that two tests remained significant after FDR correction in the raw data set, while none stay significant in the cured data set. We already insisted on that matter in the conclusion section. We have added some precisions on that point (FDR corrections) in the amended version of the Figure 6 (FlowChart).

At The last part of the methods still occur to me like a discussion (Lines 281-301). It is not clear to me which of these named potential remedies the authors performed.

Answer: We are sorry but we do not clearly understand what Dr Husemann means here. All these paragraphs describe techniques to correct for SAD, correct stuttering, compute bilateral *p*-values, and correct F_{ST} for excess of polymorphism. We used all these "remedies".

There is a small mistake I Line 265 with a duplication of the author names.

Answer: Done

In Line 278, the “instances” where the BH correction was applied should be named and explained.

Answer: We have added some precisions that we hope will meet Dr Husemann's satisfaction.

In Lines 248f the authors propose the presence of selection. It would be nice to know which minor evidence they found.

Answer: Dr Husemann probably meant "line 348". We have added some more precise information that we hope will meet Dr Husemann's satisfaction.

In figure 4, the authors should name the size of the micropeak in the header to make clearer what is meant.

Answer: Done

Besides I congratulate the authors to a nice and valuable contribution.

Answer: We thank Dr Husemann for his nice comment.

Kind regards
Martin

Reviewed by anonymous reviewer, 2019-08-09 18:17

De Meeûs et al. investigated the effects of common biases associated with SSR datasets, i.e. null allele, short allele dominance and stuttering, and proposed recommended steps for analyzing biased SSR dataset in general. The authors investigate this question on a total of 387 (? right) tick individuals sampled across the US and genotyped for nine SSR markers. I found the manuscript well written and the questions interesting but I still agree with the previous reviews. Indeed either the authors choose to:

-make a tick-centered analysis, i.e. analyze the dataset to infer population structure, diversity, demographic history, taking into account some biases associated with their SSR dataset, and if they wish, also explaining their framework to prune and analyze it;

-or either the authors choose to perform a methodological paper devoted to the analysis of 1) pseudo-observed data simulated under different assumptions (e.g. as suggested, testing different sampling size, but may be also reproductive systems, different number of biased and unbiased markers, different extent of null alleles, short allele dominance and stuttering…), and 2) of their dataset as done here, and if possible of previously published data

sets, to draw general conclusions on how to handle the biases and, if possible, providing new toolkits.

Therefore, at that stage the paper either lacks a theoretical analysis of pseudo-observed unbiased and biased SSR datasets to draw general conclusions, or do not provide deep population genetic analyses to understand the specific evolutionary history of the tick species/populations in the US.

If the theoretical analyses cannot be performed, population genetic structure analyses, and even demographic history of the species, should be provided to get a comprehensive view of the evolutionary history of the disease vector.

Major comments:
1) The authors repeat several times in the manuscript that their data suggest strong population subdivision. However, they do not provide analyses of population structure (e.g. with STRUCTURE and/or TESS softwares, DAPC analyses) with their biased and unbiased datasets. Such analyses should be added.

Answer: We are sorry but we strongly disagree with this opinion. F_{ST} and F_{IS} analyses, when handled on a correctly sampled dataset of dioecious species, can provide an accurate measure of the degree of subdivision in a biologically easy to interpret way, in contrast with Bayesian clustering methods, which are based on assumptions that nobody has ever clearly stated and provide debatable results (Latch et al., 2006; Kaeuffer et al., 2007; Frantz et al., 2009; Meirmans, 2015; Manangwa et al., 2019). The fact that we obtain, with the cured data set, substantially negative F_{IS} and substantially high F_{ST} estimates obviously argues in favor of a strong subdivision. We have added estimates of *Nm* in an Island model (here $Nm=1$) to illustrate this point.

Even inferences of the demographic history may be interesting to explore.

Answer: Sorry, but we do not understsand what Referee 2 means here.

2) The authors pooled alleles close in size to correct for stuttering. They chose their filtering threshold based on the assumptions that in small dioecious species population you expect heterozygous excess, so an extra care was performed to not remove rare alleles. Again, this assumption is a tick-centered, or a dioecious-centered, hypothesis.

Answer: This is inaccurate, as pooling rare alleles together will always generate a spurious excess of heterozygotes, even in large monoecious species. Therefore, this is not a hypothesis but a fact. We may also add that dioecious organisms, even if they are probably a minority in the biosphere, are however extremely common in population genetics studies.

If the authors would like to provide a framework for a large audience, they should provide wider assumptions, for different model systems for instance, and if possible, using already existing tools. For instance, in the same way the authors cite the FREENA software, there is the AUTOBIN macro (https://www6.bordeaux-aquitaine.inra.fr/biogeco_eng/Scientific-Production/Computer-software/Autobin). Perhaps the authors could use Autobin with different threshold to pool alleles and to provide a guideline for different model systems to correct for stuttering? This is just an idea, but at the moment I still found this pooling methodology a black box and very hard to apply in a general manner.

Answer: We do not really see what AUTOBIN would bring to the story, since this unpublished software apparently only produces alerts. We already detected problems with other means. The correction we proposed is anything but a black box. Stuttering makes it difficult to distinguish the two alleles in individuals that are heterozygous for alleles that are close in size. Then, pooling these alleles, pretending the genotyping technique cannot distinguish those, is a very easy to understand way to remove this problem. It is very easy to apply with, for instance, any spreadsheet software such as Excel, and we have precisely described in our manuscript what alleles we pooled and at what loci. Consequently, we find the terms "black box" and "very hard to apply" particularly unfair. This kind of correction can be undertaken for any kind of diploid organisms.

Minor comments:
-I am still confused with the number of samples, I think the author should summarize in Table 1 the total number of individuals per site (AL1, AL2, …), for instance by adding a line “TOTAL” for each site, and also adding a line TOTAL at the end of the table for the full dataset.

Answer: This information is already present in the last column "N", so we do not understand what Referee 2 means here.

-Figure 6: the authors should add the associated population genetic software to use for each step.

Answer: Done

-line 467, when the authors say that their “cures provided satisfactory results”. I am not fully convinced of that point as stated above. At the moment I am a bit frustrated with the results: I indeed miss either a “simulation” or an “evolutionary history” study. The authors should thus make a choice. And thus, I am also not convince about the following statements line 461 “this issue would require a full simulations study”

Answer: We do not understand Referee 2's frustration. For us, the fact that we detected null alleles SAD and stuttering is hardly disputable. The fact that the cured data set

offered much easier to interpret data is also hard to quarrel. Indeed, null allele detection and correction were extensively studied in other papers (Brookfield, 1996; Chapuis and Estoup, 2007; Séré et al., 2017; De Meeûs, 2018). The same can be said for SAD (Wattier et al., 1998; De Meeûs et al., 2004; Manangwa et al., 2019). For stuttering, the cure proposed can have no effect on heterozygote deficits if those do not come from stuttering. Numerous simulations would allow to study precisely the exact consequences of each of these phenomenon independently and in various combinations. Though we do not deny that such a study would bring interesting information, we do not see why our findings could not be successfully applied to any kind of diploid organism. Our tick population has no biological reason to display strong heterozygote deficits and highly significant LD. The cures we proposed are either extremely well documented (null alleles) or very logical and straightforward. The fact that these cures deliver a biologically interpretable dataset is in favor of the efficacy of such cures and their possible application to other biological models. We can add that to our knowledge; no simulation software can simulate null alleles, SAD and stuttering in various population scenarios, so what referee 2 is asking for really requires an additional work that we cannot ourselves afford.

and, line 480 “the correlation between mitochondrial clade and genetic structure is not the scope”.

Answer: This sentence has been changed.

References

Brookfield, J.F.Y., 1996. A simple new method for estimating null allele frequency from heterozygote deficiency. Mol. Ecol. 5, 453-455.

Chapuis, M.P., Estoup, A., 2007. Microsatellite null alleles and estimation of population differentiation. Mol. Biol. Evol. 24, 621-631.

De Meeûs, T., 2018. Revisiting F_{IS}, F_{ST}, Wahlund effects, and Null alleles. J. Hered. 109, 446-456.

De Meeûs, T., Humair, P.F., Grunau, C., Delaye, C., Renaud, F., 2004. Non-Mendelian transmission of alleles at microsatellite loci: an example in *Ixodes ricinus*, the vector of Lyme disease. Int. J. Parasitol. 34, 943-950.

Frantz, A.C., Cellina, S., Krier, A., Schley, L., Burke, T., 2009. Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: clusters or isolation by distance? J Appl Ecol 46, 493-505.

Kaeuffer, R., Reale, D., Coltman, D.W., Pontier, D., 2007. Detecting population structure using STRUCTURE software: effect of background linkage disequilibrium. Heredity (Edinb) 99, 374-380.

Latch, E.K., Dharmarajan, G., Glaubitz, J.C., Rhodes, O.E., 2006. Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. Conserv Genet 7, 295-302.

Manangwa, O., De Meeûs, T., Grébaut, P., Segard, A., Byamungu, M., Ravel, S., 2019. Detecting Wahlund effects together with amplification problems : cryptic species, null alleles and short allele dominance in Glossina pallidipes populations from Tanzania. Mol. Ecol. Res. 19, 757-772.

Meirmans, P.G., 2015. Seven common mistakes in population genetics and how to avoid them. Mol. Ecol. 24, 3223-3231.

Séré, M., Thévenon, S., Belem, A.M.G., De Meeûs, T., 2017. Comparison of different genetic distances to test isolation by distance between populations. Heredity 119, 55-63.

Wattier, R., Engel, C.R., Saumitou-Laprade, P., Valero, M., 1998. Short allele dominance as a source of heterozygote deficiency at microsatellite loci: experimental evidence at the dinucleotide locus Gv1CT in *Gracilaria gracilis* (Rhodophyta). Mol. Ecol. 7, 1569-1573.