Dear Recommender,

We have revised the preprint " A simple procedure to detect, test for the presence of stuttering, and cure stuttered data with spreadsheet programs: application to parasites and vectors". We have tried to take into account or discussed all referees' remarks. In the following, you will find below a detailed rebuttal letter were I discuss all amendments we made.

The new version of the preprint is in the file "DeMeeusStutteringTest-2ndRoundV2.docx", the same file with track-changes is in the file "DeMeeusStutteringTest-2ndRoundWithTrackChanges.docx".

These files, together with all supplementary files, are available in Zenodo at https://zenodo.org/record/6822660.

Rebuttal letter

Recommender's comments

we have now received two reviews on your manuscript. Both are positive, although one of the reviewers has some reservations. Please read the reviews carefully and refer to the points raised in your revision. Overall, both reviewers recommend shortening the text and bring out the added value.

Answer

We have tried to combine the results and discussion section into one paragraph and moved some tedious descriptions into the appendix.

We also have added a sentence or two to insist on the fact that the detection procedure proposed is much more powerful than the one available so far (in MicroChecker), and on the benefit of the different curing strategies. We still insist on the fact that, for non-model organisms, for which microsatellite markers still represent the best cost-benefit ratio.

Reviews

Reviewed by Thibaut Malausa, 05 Apr 2022 12:28

This article presents and tests a method for the detection of stutters in microsatellite genotyping projects. It also benchmarks this method and a data correction method proposed by de Meeûs et al. 2019 using several datasets from several case studies.
The proposed detection method is exposed in details and the technical choices at each step of the design are well justified. The added value of the proposed tool is documented and this method represents a decent alternative to the software Microchecker developed in 2003 and proving more and more incompatible with most systems. As such, this method deserves publication and can be of interest for teams using microsatellites.
However, as reader and possible user of such method, I must acknowledge that I am not convinced that I would use or recommend the proposed tools. I agree that microsatellite markers are still cost-efficient and informative for a wide range of projects (although the frequency of their use has decreased over years). However, next-generation sequencing has considerably facilitated the design of microsatellite markers, enabling to be very exigent in terms of markers' quality without spending much money. Hence, I am not sure that tools for stutter management are currently a top priority for most research teams.

Answer

Giving a straight answer to this argument is uneasy because Dr Malausa does not provide real examples (e.g. references) to support his assertions.

Firstly, and as far as we know, perfection and cost efficiency of NGS-based microsatellite design, as compared to traditional genotyping, has not been proven. This would indeed require a comparative approach on the same individuals and in different populations, which, to our knowledge again, has never been undertaken so far.

Secondly, perfection was the argument spotlighted by the first users of microsatellite markers, decades ago. Since then, several and frequent imperfections were found: null alleles, short allele dominance (SAD), allelic dropout and stuttering. Several efficient tools to detect and cure such problems were then developed. Saying that a technique provides perfect results is not proving that. We, in our Intertryp team, have very good reasons to seriously doubt that NGS-based techniques are flawless and as cost efficient as traditional microsatellite genotyping. We have indeed collaborated with Dr Olivier Lepais in Bordeaux to design and genotype trypanosomes and tsetse flies. The impressive number of polymorphic loci obtained was as impressive as the proportion found in significant linkage disequilibrium (11%) and the level of heterozygote deficits (F_{IS}=0.236) in one species of tsetse flies, due to various allele miscoring. All these data are still being analyzed and corrected for SAD, stuttering and null alleles, which takes a lot of time (for one tsetse fly species, so far). Even if the prices were reduced as compared to what NGS approaches used to cost, it still represented an important budget of 11200 € (for 60 loci and 96 individuals of three species), not ready to use, since days of data cleaning are still necessary before real population genetics data analyses can be undertaken.

Thirdly, we do not know what Dr Malausa means by "top priority". What we can say is that we will use this procedure in our future works, and will advise using it to all colleagues we work with. We cannot tell how many people will feel concerned, and we do not know what the threshold is before "priority" status is reached, but our feeling is that if it helps several researchers to improve their work, then it is worth being published. We understood that PCIs were not looking for ground breaking papers or for reaching the top IF of all journals, and this a reason why we support this initiative. We thus believe this preprint will be useful enough to the community.

Thibaut Malausa, 05 Apr 2022 12:28

The other argument limiting my enthusiasm is related to the apparent ergonomy and extent of added-value of the tool. The advantage of the method is its simplicity and portability over operating systems. However, it does not seem straightforward to automatise it or incorporate it in a high-throughput workflow. If I am mistaken on this point, I would recommend to be more explicit in the ms about the possibilities of use of the tool.

Answer

Automatization would indeed represent a problem for datasets with many loci. Nevertheless, with reasonably polymorphic microsatellite loci, 7-10 loci are quite enough. Researchers will easily reorganize allele frequencies output of Fstat and copy and paste the necessary commands to get expected frequencies of heterozygotes with one repeat difference, and compare it to observed ones. The procedure is really simple and does not require tremendous skills in calculus spreadsheet softwares. None of us are programmers, but I guess that many bioinformaticians can easily transpose the method into R. Once published, this may indeed encourage R-programmers to create a dedicated package.

Thibaut Malausa, 05 Apr 2022 12:28

In terms of added-value, I found that overall the ms did not underline a large added-value provided by its use.

Answer

We have added some sentences insisting on the performance of our new procedure as compared to the available one (Microcheker), its portability across any operating system, and on the efficiency of the cures proposed. We hope that these sentences will meet Dr Malausa's request.

Thibaut Malausa, 05 Apr 2022 12:28

In terms of form of the article, I also think that the text could be much shortened to convert the ms into a short and smooth methodological paper. In my opinion, some elements of contexts and some detailed results and interpretations provide limited added-value while considerably complexifying the reading. Also, the text would probably be easier to read with a section pooling “Results & Discussion”: adding a small interpretation and “take-home message” after each section of the results would likely lead to little or no loss when compared to the current discussion, and it would avoid many repetitions and statements that are too far from the result section to be easily understood.

Answer

We have now merged the results with the discussion.

Thibaut Malausa, 05 Apr 2022 12:28

Finally, I have a series of remarks, suggestions and comments, listed below:

· L20-21: This statement does not look self-explanatory. This may be inserted and explained in the introduction but I would suggest to remove it from the abstract.

Answer

We have deleted this sentence

Thibaut Malausa, 05 Apr 2022 12:28

· L27-32: I found this summary of the results hard to follow. The statements “work well”, “not perfectly”, “improve parameters” or “behaviour of their variations” sound fuzzy. Referring to more detailed results (comparisons of a set of criteria or indices used to benchmark the methods) may be more informative and clear. I would suggest to re-write the second part of the abstract.

Answer

We have rewritten this section of the abstract and hope it now meets referee's satisfaction.

Thibaut Malausa

· L41-42: I see reasons why microsatellites are useful for non-model species, but why are they particularly useful for small organisms and vectors? This may be explained here.

Answer

Small organisms are more difficult to study directly, by direct observation as for birds, or through mark-release-recapture methods. Most parasites and vectors are non-model organisms.

Thibaut Malausa

· L43 : represent

Answer

Done

Thibaut Malausa

· L48: kinds?

Answer

Done

Thibaut Malausa

· L50-60: I wonder to what extent it is useful to make a focus on SNPs here. It kind of breaks the train of ideas while providing little added-value (the article is not really about comparing SSR to SNP and sequencing).

· L56-60: The idea is clear but this sentence may be reorganized to be easier to read.

Answer

We deleted this paragraph.

Thibaut Malausa

· L60-65: I still do not understand why this is particularly true for parasitic organisms and their vectors. Many laboratories on many topics lack resources to carry out large genotyping projects. I would say that the constraint might be lower for laboratories working on species with high economic or health impact, so this would go against the statement in this sentence.

Answer

The idea that laboratories working on economically or medically relevant species get more funds seems intuitive but is not confirmed in real life. There is a big variance depending on what kind of species is studies and by what means. NGS studies on SRAS-cov-2 viruses will easily attract substantial amounts of big grants. Evolutionary ecology of trypanosomatid parasites and theirs vectors (tsetse flies, sand flies and triatomes), that affect millions of people in Southern countries, is a total different story. Nevetherless, we have deleted this paragraph. As underlined by Dr Malausa, it indeed did not bring much useful information.

Thibaut Malausa

· L66: On ? (is it correct to write « Polymerase Chain Reaction of the targeted DNA strand”?)

Answer

Yes, we agree that this sentence was not appropriately written and we changed it in a form that we hope will meet Dr Malausa's satisfaction.

Thibaut Malausa

· L83-88: I think it is not necessary to provide so many details. Stating that “Microchecker was developed in 2003 and displays incompatibility issues with most current systems” is sufficient to convince the readers.

Answer

We have deleted such detailed descriptions and added the sentence proposed by the Referee.

Thibaut Malausa

· L114 : 10,000
· L120 : simulation
· L120 : 10,000

Answer
Done.

Thibaut Malausa
· L121-122: Why 20? Whenever possible, people usually try to genotype 30-40 samples to get more reliable estimates of population genetics indices.

Answer
We have added a sentence explaining that, for subsamples, 20 is most of the time difficult to achieve in parasitic or vector populations, especially in the field of neglected tropical diseases, on which our laboratory work. We may add that if the procedure proves efficient enough with 20, there is no reason why it would not perform at least as efficiently with 30 or 40.

Thibaut Malausa
· L134 : Identity

Answer
Done.

Thibaut Malausa
· L136-137: What are the reasons behind this choice?

Answer
We added "arbitrarily". There is indeed no rational reason. 10% is quite small, so if the procedure had displayed no detection at that level, we would have tried 20%.

Thibaut Malausa
· L138-139: When less than 20 alleles were present, how many loci were recoded? 2 or 10%?

Answer
We have tried to make things clearer. All loci were recoded, except for loci that exhibited only alleles separated by more than one repeat difference.

Thibaut Malausa
· L139-140: This is probably not very important here, but in my experience, this is generally not the most realistic option. In general, stutters affect more than 10% of the alleles and sizes of stutters increase as the fragment size increases. Hence, most of the largest alleles are generally affected.

Answer
May be we will need to specifically go back to our genotypic profiles to check for this, but so far we never noticed obviously that more than 10% of heterozygotes were interpreted as homozygotes in our datasets, due to stuttering and that largest alleles were more affected, which should somehow mimic short allele dominance. Nonetheless, and as mentioned above, if detection works with 10%, then it should work even better with higher rates of stuttering.

Thibaut Malausa

· L148: I do not think it is indispensable to keep the formula in French. Most readers are not interested in the formula in French and French readers will easily translate the formula.

Answer

We have deleted French formulas all along the manuscript. This was for our colleagues from Francophone countries in Central and Western Africa. But Dr Malausa is right, finding the equivalent in French should be easy.

Thibaut Malausa

· L158 : 10,000

Answer

Done.

Thibaut Malausa

· L297: At this stage, the reader can hardly remember what is BH. I suggest to be explicit. It will not increase much the size of the text and will be easier to read.

Answer

Done.

Thibaut Malausa

· L327: What is SAD?

Answer

Short allele dominance. We know spell it in full in the amended manuscript.

Thibaut Malausa

· L331: Is it really obvious for the reader at this stage?

Answer

We are sorry, but we really do not understand what is not obvious exactly.

Thibaut Malausa

· L359: I suggest to remove “The performance of” as it implicitly sates that detecting stuttering is positive/successful while it is not the case under H0.

Answer

Done.

Thibaut Malausa

· L360: Placing comas around “respectively” may facilitate the reading

Answer

Done.

Thibaut Malausa

· L361: Reminding the H0 would help the reader here

Answer

Done.

Thibaut Malausa

· L363-364: See last comment: it would be better to be explicit earlier in the paragraph

Answer

Done.

Thibaut Malausa

· L370-372: This sounds already like a discussion. The previous paragraph was already clear and referring to Table 1 in this paragraph looks sufficient

Answer

Done.

Thibaut Malausa

· L382-383: Replace « the power to detect stuttering" by "stuttering detection"?

Answer

Done.

Thibaut Malausa

· L396: Again, the term « performance" seems to me misleading as it can be implicitly understood as a valuable property (while it is not under H0).

Answer

We deleted it. Nevertheless, by definition, with alpha=0.05, we expect 5% of significant tests under H0. This means that lower proportions may prejudge of low performance procedures. But such discussions obviously would bring no added value to the manuscript.

Thibaut Malausa

· L407: Replace « seemed" by "was"?

Answer

Done.

Thibaut Malausa

· L472-L486: I found this entire subsection "Clonal populations" is hard to follow. I think this comes from the first sentence that contains the statement "significant stuttering signature" that I do not find explicit; and at the end of the section I do not see as self-explanatory the fact that average FIS CI are consistent with the expectations

Answer

We have tried to make this more explicit. We can also add a figure if things still appear unclear.

Thibaut Malausa

· L474-476: This second statement is relative to the H1 I suppose?

Answer

The Referee is right and we have clarified this in the amended manuscript.

Thibaut Malausa

· L489-490 : This first sentence seems little useful

Answer

Deleted.

Thibaut Malausa

· L531-592: I find this section far too detailed. I would find it much more clear and convincing if it could be shorter with a focus on the qualitative/quantitative comparisons between the results obtained using the several methods.

Answers

We have simplified this section as much as we could and we hope that it now meets Dr Malausa's satisfaction.

Thibaut Malausa

· L540 : translate “Côte d’Ivoire”?

The political staff of Côte d'Ivoire has strictly specified that they refused that the name of their country be translated into any other language but French. We have no problem with "Ivory Coast", but in the (very unlikely) case where one of the officials of this country had a glance to this preprint, we found it wiser to keep it this way.

Thibaut Malausa

· L541: “authorS”?

Answer

Yes, but this sentence was deleted in the amended version.

Thibaut Malausa

· Discussion : in my opinion, this text would be much easier to read with a section pooling “Results & Discussion”. Overall, I think that adding a small interpretation and “take-home message” after each section of the results would result in little or no loss when compared to the current discussion, and it would avoid many repetitions and statements that are too far from the result section to be easily understood.

· L595: This statement is confusing (I also had the same feeling when reading the introduction): are we talking about the results under H0 (hence, this is a positive feature), under H1 (negative feature), or overall?

· L598: Reminding some results may be more convincing here

Answer

Discussion does not exist any longer.

Thibaut Malausa

· L685: Who is JBR ?

Answer

We apologize, this resulted from an uncontrolled copy and paste from another preprint. This acronym was deleted.

Reviewed by Thierry Rigaud, 06 May 2022 12:35

Despite the rise of NGS and genomics, microsatellite genetic markers remain useful in population genetics studies. Due to their long-standing use, we now have an excellent perspective on the advantages and disadvantages of their use. One caveat of their usage is stuttering during the amplification process, which produces artificially alleles with one repeat difference, generating artificial heterozygote deficits in population genetics studies. This paper proposes a new method to detect stuttering in microsatellite data. This paper is an important contribution to the field because, as well described in the text, this new method is more efficient in most cases than the only tool presently available (MicroChecker). This study is therefore useful and timely.

The paper present the method, compare its efficiency with MicroShecker, but also examine consequences on various F statistics. This is done by a combination of simulations and tests on real datasets. The paper provide results of simulations for exploring the detection of stuttering, and answer the question: is the new method improve detection? The answer is often “yes”, but high proportions of false stuttering detections were found in clonal organisms. Then the consequences of stuttering of F statistics in population genetics (FIS, FST) are explored. Interestingly, 10% stuttering induce significant deficit of heterozygotes, especially in populations with selfing, but stuttering do not influence occurrence of linkage disequilibrium between pairs of loci.

I found the paper well written (albeit sometimes a bit too detailed, see comments below) and methodology accurate. I only have few comments aiming to improve (I hope) the reading of the paper and the usage of the method. Please find below these comments:

1- My main comment is about the description of the alternative method, L. 190-260. Instead of the long and fastidious description of what and where paste the results from Fstat and formula, why not providing a template (or example) on an excel sheet as supplementary material? This would fit the paper tittle

Answer

We have introduced five supplementary files, corresponding to the first simulation, with parameters used, data files obtained for Fstat and Genepop, output file of Easypop providing different parameters' values for all generations, and two templates: one for the genesis of 10% stuttering for these data, and another describing stuttering detection procedures. These files are introduced in the preprint in the needed paragraphs of the Material and Methods section.

2- The end of the first paragraph of the introduction is a kind of mix between different ideas economic impact of the non-model diseases / Constraints for developing genetic markers. It is a bit hard to follow. I would suggest rewriting and separate these two things.

Answer

Following Referee 1's advice, most of these were deleted.

3- L. 81. Please explain why a “global test might be more …/… robust”

Answer

We expected that a global procedure would avoid extreme rare significant tests, with the average across all subsamples. However, from our results, it was clear that the problem never was about robustness, except for clonal populations (but for other reasons). We thus deleted "more robust" for the sake of clarity.

4- L. 134: “identitiy” should be “identity”

Answer

Done

5- L. 306-308, about curing data. I do not understand why this “otherwise” procedure is made here, but not in cases where alleles were less rare. In other words, why forcing the grouping with more than one repeat difference? Why not leave them as they are? (is it making a difference?)

Answer

Successive alleles with one repeat difference, if their allele frequencies sum to more than 0.05, can have a significant impact on perceived heterozygosity. We also wanted to avoid pooling rare alleles together. Pooling such alleles with the closest non-rare allele seemed the best compromise. Nevertheless, this particular situation did not arose very frequently. We have tried to add some more explanations in the new version of the preprint and hope these will meet Dr Rigaud's satisfaction.

6- L. 475-477. I think the sentence is not complete (the proportions increase, but why are they increasing?)

Answer

Referee 1 made the same remark and this was amended.

7- L. 532. I do not understand this beginning of sentence since the following sentence show that there are differences.

Answer

Referee 1 made the same remark and this was amended.

8- L. 557-561 (and in other examples). All these list of alleles are very specific. Can you please recall here the reference of where to find the data?

Answer

Following both referees' remarks, these lists were removed and placed in the Appendix.

9- L. 640. Change “tried out” by “considered” (?)

Answer

Done, in the new "Results and discussion" section.

10- L. 675. Remove “;”

Answer

Done.