Dear Editor,

Dear Reviewers,

We are grateful for this comprehensive and constructive review of our research. The four independent reviewers provided in-depth revisions that helped us improving our manuscript tremendously.

Particularly, similar comments, consistent across the different reviewers, triggered us to increase the number of samples that were verified through cloning. Therefore, we had to re-do all the analyses, including those new cloning data. While the main message of the manuscript did not change, we are confident that all alleles were validated through different approaches as outlined in Supplemental information 2. This is the reason why our revision took more time. We also removed from the manuscript the notion of "singleton", which seemed confusing to several reviewers, as the proportion of unshared alleles dropped sharply through allele validation, and the remaining unshared alleles do not hide the signal anymore.

We also decided to present the data separately for the filtered dataset (*i.e.* including only the individuals for which all marker sequences were available) and the complete dataset, highlighting the extensive genetic mixing between species in the genus *Corbicula*. The title of the paper therefore changed from 'Distinct biogeographic origins of androgenetic *Corbicula* lineages followed by genetic captures' to 'Extensive genetic mixing between worldwide-collected freshwater *Corbicula* lineages'. Our data also still suggest distinct biogeographic origins of androgenetic lineages, which is presented in the result and discussion sections, but the extensive allele sharing found between invasive and native *Corbicula* lineages is worth highlighting. This also prevents taxonomic species identification in this genus.

The authors' contributions have changed, with Martin Vastrade being first author because he ended up adding more data, validating the alleles through different approaches and performing all the analyses. He wrote the manuscript with Karine Van Doninck.

Please find below our answer to all comments made by each reviewer. For clarity, their comments have been put in bold, while our answer are not.

Decision

by Chris Jiggins, 2019-05-08 20:34
Manuscript: https://doi.org/10.1101/590836

Revision of 'Distinct biogeographic origins of androgenetic Corbicula lineages followed by genetic captures'

The reviews generally accept the broad interest of the paper and the advances, and I agree with this.

However, there are some important comments on the paper that need to be addressed. The key comments include:

1) the framing of the paper and making it more appealing to a general audience - it is currently written very much about the Corbicula clams, but it is unclear how it relates to

evolutionary processes on other organisms. This needs work both in the introduction and perhaps also in the conclusions - how does the unique reproductive mode affect the genetic variation in these organisms compared to other systems?

We agree with this comment. The evolutionary context of androgenesis was developed in the present manuscript, both in the introduction (lines 45 - 74) and discussion (lines 685 - 689).

2) Clear statement of the hypotheses being tested. This is particularly important - the final paragraph of the introduction focusses on methods used (i.e. Haplowebs) but does not really outline what the paper is trying to test. Could there be a list of specific hypotheses laid out in this paragraph?

Our final paragraph in the introduction (lines 178 - 192) was indeed re-written to highlight the aim of this manuscript before presenting the novelty of the method.

3) Clarify nomenclature such as 'egg parasitism' - I agree with the reviewer that this seems a misleading term and request that a better term is devised.

We have explained in the introduction (lines 64 - 66) why egg parasitism occurs in an androgenetic system. We hope it is now clear why this term is used (also in the literature on androgenesis). We added the reference Lethonen *et al.* 2013 which clearly explains this term.

4) Better clarify what is novel in this paper relative to previous work on the same system. There are a number of references to previous studies but it should be clarified which data comes from previous papers (and if published data are not included why was this the case) and which questions can be addressed here that were not addressed before. More specifically, what exactly is novel here apart from the use of a web-based method of analysis.

We decided to change the title of the paper for this reason, highlighting the novel results of this research. In the introduction (lines 185 - 186) we emphasize the novelty of our analysis which allows us to show for the first time the **extensive** mixing between species in this genus, the sharing of alleles between native and invasive *Corbicula* lineages and the distinct biogeographic origins of the invasive lineages. The abstract, the results and the discussion were also adapted to highlight the novelty of our findings compared to previous work. The inclusion of data from previous researches was only made when the raw data (chromatograms) were available, allowing us to check the exact sequence and to analyse a strong, complete dataset.

I am also concerned about some of the methodological questions raised by reviewers. How many clones were sequenced per individual and why were triploid allele patterns not detected - is this just because too few clones were sequenced. Also clarify the 'remove singletons' option in the Haploweb program. Why were singletons removed?

Details on the cloning methods were added in Supplemental Information 2: between 20 and 30 clones were sequenced for each individual. Triploid patterns were detected and are presented in Table S2 and S3, the three alleles were considered in the allele sharing analysis. This point is also developed in Supp. Info. 2. The singletons were not removed from the new analyses. Previously singletons (unshared alleles) occurred in high proportion hiding the informative alleles. However, the number of singleton alleles have dropped dramatically through allele validation and the remaining unshared alleles do not hide the signal anymore. The haploweb

analyses was performed twice, once on the entire dataset (n=359 in total but with sequences lacking for many individuals for at least one marker) and once on the filtered dataset, including only individuals for which all four markers were available (n=141). Haplowebs on filtered data being easier to interpret, they are included in the main text (Fig. 2) while haplowebs done on the complete dataset are presented in supp. Data (Fig. S2).

There are a large number of additional minor suggestions that should be addressed in the reviews.

Subject to addressing these points I think the manuscript is appropriate for a recommendation.

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

Reviews

Reviewed by Arnaud Estoup, 2019-04-18 09:54

The questions addressed in the paper by Etoudy et al. are of high interest and refers to an evolutionary topic of general interest. The dataset is impressing and the statistical treatments (based on previously published methodological frameworks) are innovative and insightful in the complex context of worldwide Corbiculata genetic variation. The paper brings new insights into the biogeographic origins of androgenetic Corbicula lineages, both in the native and invasive ranges of the species. I believe that this work deserves publication and recommendation in PCI EvolBiol.
I feel however that the ms could (should) be improved on different aspects to make it (even) more interesting and readable (especially for non-expert readers). I propose that the authors consider the following comments (four major points followed by a number of minor points) to improve their ms.
1/ The study context has been nicely formalized in the Introduction section which I globally like very much. I feel however that the addition of a (very first) general section

about variation in sexual systems with a particular emphasis on the rare and intriguing androgenetic reproduction system would give a more generic perspective to this work; see for instance Rey O, Facon B, Foucaud J, Loiseau A, Estoup A. 2013 Androgenesis is a maternal trait in the invasive ant Wasmannia auropunctata. Proc R Soc B 280: 20131181. http://dx.doi.org/10.1098/rspb.2013.1181.
A more general section was indeed lacking in our introduction and has been added (lines 45 - 74). We thank Arnaud Estoup for this suggestion, providing now the general context of our system, which is important for non-expert readers.

2/ The Results section is difficult to read especially for non-expert readers. The authors should change its structure to make it easier to read. At the moment results include too many details that are presented somewhat as a list which "submerges" the reader. Try to follow a kind of two-step presentation. For each result section, start first by describing "the big picture" of the results so that readers get the main message(s). Then you can go into details that might be of importance for specialist readers.
We re-wrote the result section entirely and followed the recommendations of this reviewer, distinguishing the details from the main message. We hope it is easier to read in the present version.

The illustrations (i.e. figures) that should be kept in the main text should preferentially correspond to those that integrate results on multiple-loci. Concretely, keep Fig 1 (cited from the introduction which is fine to me), then keep Fig 3 and add Fig S3 in the main text, as the two later figures provide some nice synthetic summaries/illustrations of results. You might also include in the main text one figure of the locus-haploweb treatments (choose the "best" locus), but other locus-haploweb figures are redundant in what they show and should be hence included in the Supplementary Materials.
(see below for presentation details of the figures that should be improved to make them easier to interpret)
We agree with the redundancy of the different haplowebs in the previous version. However, the latest analyses have revealed some interesting differences between the markers. Moreover, the haplowebs are key to understand the Field for Recombination (FFR) delimitation and the introgressions or allele sharing between them. We therefore decided to keep them in the main text of this revised manuscript that highlights this extensive genetic mixing. We however did the haploweb analyses twice, on the complete dataset (n=359 in total but with many individuals lacking at least one marker) and on the filtered dataset, including only individuals for which all four markers are available (n=141). Both haplowebs gave similar results but the filtered dataset is easier to interpret and included in the main text (Fig. 2) while the other haplowebs are presented in supp. data (Fig. S2).

3/The discussion section is nice. My only point is that the authors should avoid pointing recurrently to all their tables and figures (see a symptomatic example with lines 403 – 405). This is expected in the Results section but not in the discussion section.
We agree and we have reduced this recurrent refencing.

4/ As a last main point I have to say that I am (very) surprized by the absence of triploid molecular signatures in the analysed dataset, a feature which would be expected as indicated by Fig 1b. How do the authors explain that? This point might be evoked somewhere in the Discussion section.
This is one of the points that made us verify and increase the number of samples that were checked through cloning. Actually, we found triploid signatures, these are given in Table S2-3.

Details on the phasing of triploids and how they were included in the analysis were added in Supplemental Information 2.

Other following points correspond to minor points that might also help improving the ms:
- Abstract; L23 "egg parasitism between distinct lineages": please explain what this notion means through a few simple words.
Generally speaking, I am not fan of the "egg parasitism" terminology used all over the ms. It is not clear to me why the term "parasistism" is used. The "parasitism" terminology" involves precise biological meanings which do not seem to apply here. I would prefer that you just refer to the genetic/evolutionary mechanism itself (mt or nuclear DNA capture) without referring to "parasitism".
While egg parasitism is broadly used in the scientific literature on androgenesis, we agree this should not be mentioned in the abstract as it requires additional explanations. However, we included this terminology in the introduction (line 64 - 66) as androgenesis is indeed reported as an egg-based sexual parasitism (see Lethonen *et al.* 2013). This reference was not explicitly quoted in the previous version, we added it.

- L61: the reference Schwander & Oldroy 2016 is missing in the reference list.
Indeed, very important reference. It was added.

- Figure S1 is great and very useful: refers to it more thoroughly in the material and methods as well as later in the ms (for example in the legends of haploweb figures). Regarding Fig S1 itself: marker of interest => say clearly that it actually means sequence haplotypes. "median-joining network": please explain/specify. "Sharing mutually exclusive pools of alleles" is unclear => please reword and specify. Fields of recombination are clearly delineated by hand which is great: pleased do the same in your illustrations of haploweb-based figures.
The legend was clarified, and a new reference to this figure was made in the result section (line 351), before the description of the haplowebs, and in the legend of the figures (line 1031 and 1070), to help the reader understand the method and the haplowebs (Figures 2 and S2).
However, the addition of hand-made delimitations of the FFR on the main figure would make it less clear, the colour codes are referring to the delimited clusters (through CoMa or KoT analysis).

- L281: it is not clear from Fig 3 where the sexual lineage C. sandai is located on the figure
Lineages information was added to the figure (currently Figure 4) in the legend of each idiotype.
- L320: "reveal"…really ??? Please reword to tone down this term here.
This section was completely modified and we did not use "reveal".
- Check all over the ms if the word "indeed" is really appropriate cf. a very "French – English" wording. Most of the time the sentence is fine without "indeed".
We have addressed this.
- L366: Two patterns => evolutionary scenarios might be better.
We re-phrased this in term of "hypotheses" (line 641 – 644).
- L425: "Genetic or Evolutionary" origin would be more appropriate
There must be a mistake, I cannot see where your comment applies here.
- Fig 2a,b c: avoid repeating legends – try to delineate fields of recombination by hand to make them more visible.
The addition of hand-made delimitations of the FFR on the main figure would not make it clearer, we used colour codes instead.

- Fig 2c and S2: FFR2 and 4 seem to have the same colour. Use clearly different colour codes for the different FFRs.
This original figure S2 does not exist anymore.
- Fig 2d: what means FW ??? Please specify.
These are the original barcode names of the COI haplotypes of the invasive *Corbicula* lineages. We made it clear in the introduction section (lines 129 - 137).
- Fig 3: great figure: please add a few important things in the figure. Help readers by specifying in some ways the androgenetic and sexual haplotypes in the figures. Specify the native and invasive haplotypes too. Finally please specify where C. sandai, C. moltkiana, C. fluminalis, and C. African are located in the figure to help readers to following what you mention (recurrently) in the results (and discussion) section.
Information on the occurrence of sexual and androgenetic lineages was added to each group (or idiotype). The information on the different species was also added, the coloured alleles have been specified (invasive alleles in colour, native alleles in grey). Note this figure was renamed Figure 4.

Reviewed by Simon Henry Martin, 2019-04-29 23:03

The authors investigated the origins of androgenetic lineages of Corbicula clams. The origin of a distinct reproductive mode represents a key evolutionary transition, and is of particular relevance in Corbicula because it appears to be associated with increased invasiveness. However, past studies based on one or two markers have failed to pin down the origin of androgenesis. In the present study, the authors analysed sequences of four nuclear markers and COI from several hundred individuals including both invasive androgenetic and native sexual populations. They made use of an analysis of allele pairs in diploids to delineate distinct gene pools and conclude that there are at least three distinct geographic origins of androgenetic lineages. While I think that this study has certainly extended the knowledge of the history of this genus, I have several concerns about core aspects of this study, including the hypotheses tested, conclusions drawn and methodological choices.

1. **Hypotheses and main conclusions My main issue is that I found the overall message of this paper somewhat ambiguous. Firstly, it should be made clearer in the introduction what question is being addressed, or what hypotheses are being tested. As far as I can tell, the main question is: Does androgenetic reproduction have a single origin or multiple origins in Corbicula? This was not clear to me upon first reading. The first direct description of the two hypotheses, and the expected evidence that would support one or the other, appears deep in the Discussion (line 366-372).**

Our final paragraph in the introduction was indeed written to highlight the novelty of the method, maybe occulting the aims of this manuscript. We have re-written this last paragraph (lines 178 – 192) outlining the aims of this work and why this specific method of allele sharing was used in the context of the genus *Corbicula*. The main results are the extensive genetic mixing between *Corbicula* species (from both the native and invasive regions) and the distinct biogeographic origins of invasive *Corbicula* lineages, the title has been changed to highlight this extensive allele sharing.

Immediately after the above statement of hypotheses, the authors seem to claim that the data support the alternative hypothesis of distinct origins of androgenetic lineages.

However, they then immediately state (line 374) "While our results do not disentangle the origin of the peculiar reproductive mode of androgenesis in Corbicula, it shows for the first time the distinct biogeographic origins of androgenetic Corbicula lineages [...]".

How are the two parts of this statement reconciled? I agree with the first part. Given the extent of hybridisation and 'nuclear capture' in androgenetic lineages (as documented in previous studies and this one). It seems impossible to rule out that there was a single origin of androgenesis followed by nuclear captures that could have wiped out traces of the original ancestry of form C/S for example (at least at the four nuclear markers were used in this study). I am therefore confused by the second part of the above quote - the idea that this study supports "distinct biogeographic origins" of androgenetic lineages. If we cannot rule out the hypothesis of a single origin followed by nuclear capture in certain lineages, how can we accept that there are distinct origins?

Perhaps the authors are making a subtle distinction between the (more recent) biogeographic source of a particular invasion and the (possibly older) origin of androgenesis as a reproductive mode. If so, this must be stated more clearly to remove ambiguity. The title might need to change too, as this appears to claim that nuclear captures occurred AFTER distinct origins of androgenetic lineages, rather than nuclear captures simply creating the appearance of distinct origins.

We agree with this reviewer, there is a distinction between the more recent biogeographic origin of the invasive lineages we found in Europe and America and the origin of androgenesis as reproductive mode in *Corbicula*. This distinction is discussed in the new version of the manuscript (lines 532 – 535 and 577 – 580). Due to nuclear captures, it is very difficult to trace the origin of the reproductive mode. In our manuscript, we highlight two main results: 1) the extensive genetic mixing between freshwater *Corbicula* "species" (from both the native and invasive regions) which prevents the clear identification of taxonomic clusters, 2) the distinct biogeographic origins of invasive *Corbicula* lineages since we were able to identify three different genetic clusters containing invasive lineages (with limited allele sharing between these 3 FFRs). Androgenesis in *Corbicula* results in genetic mixing between distinct lineages, therefore blurring their delimitations, but since the origins of the invasive lineages appear recent, we still observed a signal of distinct origins (as discussed lines 532 – 535). We have highlighted in the title of the revised manuscript this extensive genetic mixing rather than the distinct biogeographic origins, but both results are important in this study.

1. **Exclusion of singletons Singleton alleles were excluded as they "are not informative about allele-sharing". This lead to a huge amount of data being disregarded: nearly 90% of sequences in the case of the atps gene, including all Rlc and Indonesian individuals. Singletons might not be indicative of sharing of identical alleles, but their genetic distances from other alleles are still informative about ancestry. A more quantitative approach that considers all haplotypes and the genetic distances among them would surely add power, and potentially reveal additional sharing of highly-similar (but not identical) haplotypes between sexual and androgenetic lineages. I think it is important to at least show that inclusion of these singleton alleles would support and not contradict the main patterns found using the limited number of shared identical haplotypes.**

We agree with that comment. Singletons were not removed anymore in the new analyses. Previously singletons (unshared alleles) occurred in high proportion hiding the informative

alleles. However, the proportion of unshared alleles dropped sharply through allele validation, and the remaining unshared alleles do not hide the signal anymore.

2. **The results section describing Figure 2 only mentions sharing between C/S and A/R forms at 28S in 3 putatively hybrid individuals (line 251-255). However, Figure S3 (Zone A) suggests allele sharing between a larger number of "FFR3" and "FFR4" individuals. Is it right that there are more than 3 putative hybrids? Perhaps some of these are the C/S individuals mentioned later that share alleles with Lake Biwa at the amy gene (line 279)? Why is this not mentioned in the description of Figure 2? Similarly, the paragraph from line 289-293, describing allele sharing between different invasive lineages does not mention the sharing between C/S and A/R at all. I think omitting these findings from the description of results is problematic as it could create the impression the C/S is more distinct than it really is.**

These sections have been completely reworked and the sharing of alleles between FFRs is now consistent and clearly highlighted in Figure 4 and Tables S2, S3. Please find a thorough description of the allele sharing between forms A/R and C/S in the Result section (lines 416 – 426).

3. **Line 325-326. "In this study, the haploweb approach and the conspecificity matrix gave consistent results for all markers tested." This statement is misleading. Firstly, the conspecificity matrix is simply a summary of the Haploweb results and does not represent an independent source of information. Secondly, due to the exclusion of singletons, the placement of many individuals in a particular FFR is supported by only one of the four markers, so these individuals cannot be used to support the argument of consistency across markers. Finally, there are several individuals for which the placement is ambiguous (i.e. zones A and B of Figure S3). I therefore think that this part of the paper needs to be reworded to be more true to the findings.**

This section was reworked entirely, and we have now a very robust allele dataset. Any sentence that would describe the conspecificity matrix as a new source of information has been modified, because indeed it summarizes the Haploweb results (see lines 351 - 426). The singletons were not excluded (see explanation above) and the filtered dataset contains all individuals for which all four markers could be analysed.

4. **Markers Is any information about the independence of the four nuclear markers available? For example, are they on different linkage groups? This would help with interpretation of cases where the results are consistent among markers. Moreover, I think the paper needs to point out that a larger number of markers, such as GBS or WGS data might be able to resolve the questions further. I appreciate the huge effort that must have gone into generating the sequence data for the present study. However, with hindsight I think it is fairly clear that information from just four markers would be unlikely to resolve the origins of androgenesis given the extent of hybridisation in this genus. It would be a shame for future researchers to invest time in generating similar large data sets from a small number of markers just to discover that they lack sufficient power to infer genomic ancestry. Moreover, assuming that there are loci associated with androgenesis, genome-wide sequence**

data would potentially enable future researchers to map these loci through association analysis, which could also shed light on its origins.

No information is available on the independence of the markers used (there is not yet a genome assembly of *Corbicula*) and the results are not 100% consistent among markers which is briefly discussed in the present manuscript (lines 675 - 685). We also discussed the need to increase the power of further analyses (lines 577 – 580 and 597 - 599) and the next study indeed will include genome wide analyses of hybrids. We thank this reviewer for this interesting comment.

5. **Samples from previous studies Why were previously studied samples, such as those from Iberia (Peñarrubia et al. 2017) and Russia (Bespalayaet al. 2018) not included in this study? Is it simply because they only used 28S and COI markers? Since so many singletons were excluded from the present study, this doesn't seem to be a very strong argument for excluding these potentially interesting populations.**

Given the difficulty to obtain reliable alleles when phasing polyploid individuals – which justified the use of cloning, see the new Supp. Info. 2 – we decided to only include in our work sequences for which the raw data (chromatograms) were available, because we need the chromatograms to correctly phase the alleles, validated through cloning for the ambiguous cases.

6. **Line 342. Since the reproductive mode of the South African individuals was not identified, is it not possible that this represents an invasive population of form C/S rather than a geographic origin?**

Africa belongs to the native range of *Corbicula*, populations there might be (very) old (Pleistocene) and include sexual species but unfortunately this continent is poorly studied for *Corbicula*. The presence of form C/S in Europe is recent (dating from 1980s). The clustering of form C/S with South African individuals rather than with Asian populations (as observed for Form A/R, Rlc and B), suggests a distinct biogeographic origin for this form rather than the presence of the same form on both continents. We cannot however exclude that the South African population is also an invasive lineage and therefore added this in the discussion (lines 551- 558).

7. **I personally find Figure S3 to be a more easily interpretable summary of the Haploweb results than the multiple panels of Figure 2. However it would be even more useful if information about the geographic origin or morphological form, or both of each individual can be included along the axes, rather than simply summarising which groups of individuals fall in each clade.**

Labelling each individual in this figure would not be readable. This figure is ordered based on the conspecificity scores only, which means the individuals are not ordered by geographic origin or morphology. The results in the revised manuscript include both the haplowebs (on the filtered dataset), the conspecificity matrix and the circos, providing complementary information: the haplowebs (Fig. 2 and S2) show the details on allele sharing per marker, the conspecificity matrix highlights the FFR clusterization defined by the CoMa method, and the circos shows the allele sharing between all lineages.

8. **The lines connecting different populations in Figure 3 sometimes differ in width in the two populations. Why is this? I would expect that if 10 alleles are shared between two populations, the line should have width 10 at both ends?**

The width is proportional to the number of occurrences of this allele in each population. The same allele can be found twice in a population but 10 times more in another one. This information was not mentioned explicitly in the legend, it has been added now to be more comprehensive.

9. **Nuclear capture Fertilization by diploid sperm with retention of maternal chromosomes is called nuclear capture in line 59 and simply hybridization in Figure 1b. Is there a clear distinction, and if not, can a single term be used throughout the paper?**

These different terms indeed refer to the same process and are used as synonyms from their first use (line 101) in the new manuscript.

10. **Line 111 cites Doyle 1995 for FFRs, whereas Figure S2 cites Doyle 1992. Is this intentional?**

No, this was a mistake and has been changed. Thanks.

Reviewed by anonymous reviewer, 2019-05-03 13:48

This study investigates the phylogeography of Corbicula clams, a genus that includes both sexual species and asexual relatives. The asexuals have a special reproduction system ("androgenesis") that results in clonal transmission of the paternal nuclear genome, while the formation of both eggs and sperm are still required. Moreover, due to maternal inheritance of mitochondria, phylogenies may show "cytonuclear mismatch" due to "capture" of mitochondria from divergent lineages. Finally, nuclear maternal chromosomes may sometimes be retained, which potentially leads to formation of hybrid nuclear genomes with increased ploidy and to diversity among asexual lineages.

The study is carried out carefully and the manuscript is well-written. However, the importance of this work beyond the particular study system is not sufficiently well explained. I can see several potential approaches, which may be taken to place the study in a more general framework. For instance, phylogeographic studies in asexuals with complicated reproduction systems may be addressed in a more general way, and the advantages of the main method for data analysis employed here ("allele sharing") to analyze such systems could be explained more explicitly. The current manuscript, however, is mostly set up from a system-specific perspective. Moreover, several similar studies have been carried out on this system before, and some of the results of the current paper are of confirmatory nature (though being based on a larger sample and more suitable genetic markers than previous studies).

The manuscript has been rewritten to include the more general framework (see lines 45 - 74 in the introduction and lines 685 - 689 in the discussion). We also highlighted the novelty of our findings compared to previous studies (*e.g.*, lines 510 – 515, and throughout the whole discussion).

Besides these general remarks on the interest of the manuscript to a general evolutionary readership, I have a few specific comments and suggestions for further improving/clarifying the paper:

- **L. 132: If mitochondria are inherited maternally but the nuclear genome paternally, assigning "lineage identity" based on mitochondrial data only appears to be questionable.**

We agree with that comment. Mitochondrial barcode COI was formerly used to describe species within the genus *Corbicula*, but in an androgenetic system such as *Corbicula*, using only mitochondrial sequences to delimit these species is not valid. We decided to add these names in lineages descriptions (Table S2, between brackets) to allow a comparison with previous studies but as outlined in the manuscript (lines 630 - 632) we strongly encourage the use of both nuclear and mitochondrial markers in *Corbicula* to identify lineages. One of our main messages is to challenge this lineage identification method because of the extensive genetic mixing between distinct genetic lineages.

- **L. 173: Triploidy is not necessarily expected to lead to triple peaks on chromatograms because the SNPs between the three haplotypes are not necessarily at the same site. More generally, some clarification on the ploidy of the investigated asexuals would be welcome.**

We agree with this reviewer. This section on ploidy results was completely reworked. The ploidy was not studied by itself, but can be partially estimated from the number of alleles retrieved in each individual. This information is accessible in Tables S2-3 and we outline in Supp info 2 how triploid alleles were included in the allele sharing analysis.

- **L. 177: Were all cloned individuals asexuals? Were they triploid but only two haplotypes were found?**

The number of samples that were verified through cloning was increased, with 20 to 30 clones per individual. We cloned individuals for which the haplotypes were doubtful based on direct sequencing. This information has been added to Table S2. Some cloned individuals are asexual (androgens based on sperm morphology) but for some, the reproductive mode is unknown. We added a supp. Info. 2 to clarify the phasing on triploids. As no direct measure of ploidy was made in this study, triploids were detected based on the occurrence of more than two alleles. The occurrence of triploid individuals with only two haplotypes is likely but unknown.

- **L. 199: It would appear to be better to exclude COI from these analyses (or analyze COI separately) because it is expected to follow a different inheritance pattern.**

We agree with this comment. The conspecificity matrix (Fig. 3) was redone, and appears to be clearer and easier to interpret with only the 3 nuclear markers. We decided to analyse COI separately, using a new analysis (KoT analysis) which is described in the methods. These two results are represented separately on the haplowebs, allowing an easy comparison of nuclear and mitochondrial markers because indeed their inheritance history is different. Thanks for this interesting comment which improved the clarity of our results.

- **L. 216: Please state whether these analyses were carried out on all data or on the data where singleton SNPs or singleton haplotypes were removed (which in both**

cases would further underestimate diversity). In addition, some information on how triploids were analyzed would be welcome.

Singletons were not removed anymore in the latest analyses. Previously singletons (unshared alleles) occurred in high proportion hiding the informative alleles. However, the number of singleton alleles have dropped dramatically through allele validation and the remaining unshared alleles do not hide the signal anymore. Genetic diversity was calculated on the entire populations, including all alleles independently of the occurrence of triploids. More information on the analysis of polyploids is available in Supp. Info 2.

- **L. 230: May the higher homozygosity of locus atps be explained by null alleles? Presumably "homozygous" in triploids indeed means that all three alleles are identical. Hence one may expect lower homozygosity rates in triploids. It may thus be good to split sexual and asexual lineages for these analyses.**

This point is discussed on the lines 645 - 649. The reproductive mode is unknown for several lineages, we therefore decided not to split asexuals and sexuals in this analysis. Notice these values have been edited based on the new analyses (line 347-349).

- **L. 234: Same remark as above: combining all four markers, including the mitochondrial one, does not seem safe for asexuals.**

The analyses of nuclear and mitochondrial data have been separated. Thanks for this comment.

- **L. 264 and throughout the manuscript: Clarify what "singletons" refer to, haplotypes or SNPs.**

As this notion seemed misleading and we removed singletons from our new data analysis (see explanation above). The new dataset is robust and has clarified this singleton aspect.

- **L. 422: Again, some more concrete information on ploidy seems necessary to understand what is meant here (currently the paper says that no individuals with three alleles were identified, but this seems to be contradicted here).**

This section has been modified. Occurrence of more than two alleles in a given individual can be visualized in Tables S2-3.

Reviewed by anonymous reviewer, 2019-05-05 20:57

This study seeks to better understand the origin of the genetic diversity observed in different asexual lineages of Corbicula clams. The authors investigated the pattern of allele sharing between different asexual lineages and related sexual species using a collection of individuals sampled worldwide. They identify three distinct genetic clusters containing asexual lineages with different biogeographic origins. Generally, I found the approach of using haplowebs to clarify the relationships among the different lineages very interesting. However, I found that some parts of the manuscript were hard to follow, especially for readers that are not familiar with this system. I provide below major and minor comments that I hope will help improve the manuscript and making it more accessible to a general audience.

Major comments:

1/ I found that some basic information regarding androgenesis in this species was missing. This type of information is important as the authors rely on particular assumptions for their haploweb analysis and as this could help interpreting their data. For example, it is unclear whether self-fertilization in androgenetic individuals involves haploid or diploid sperm and whether recombination can occur between homologues (L 42). The authors could also better explain the mode of reproduction of polyploid individuals. For example, in Fig. 1b would the resulting triploid individual produce only haploid gametes with a blue chromosome and diploid gametes with red chromosomes? The authors do not discuss the fact that they do not detect more than two alleles in samples that could potentially be polyploid (L 173). Does that mean that triploids only arise through self-fertilization of diploid or triploid individuals and not through outcrossing with other lineages?

The introduction has been modified including a more general context (lines 45 - 74), we hope it is more accessible to a general public. All the information known on *Corbicula* androgenesis has been included in the introduction. Androgenetic lineages in *Corbicula* appear to produce only unreduced spermatozoa (no haploid sperm), that seem to be used for both cross-fertilization or self-fertilization. Crossing-over can probably still occur during gamete production, but this was never studied or shown; this point is discussed (lines 682 - 683). In our new analyses, we detected more than 2 alleles in some individuals, this is outlined in tables S2-3 and in the supp info 2, we clarify how these triploid alleles were included in our analysis.

2/ I found that the authors could better explain how recombination between distant haplotypes in sexual or asexual individuals could affect their results. I am not familiar with the haploweb approach, but this seems to be an issue to me. The authors did check for the absence of chimeric sequences and for the validity of their method by performing both cloning and direct sequencing on 5 individuals. However, no information is provided on the way these individuals were chosen. To make a stronger point, I guess I would have chosen individuals from a diverse sexual population such as Lake Biwa, rather than from asexual populations from Europe or America (L177). Also, the authors did not provide the number of clones sequenced per individual. I thought that "(done twice)" meant that they only sequenced two clones per individuals, which seems an unreasonably low number.

We agree with this comment and we have improved our entire cloning analysis, having added additional cloning data in the revised manuscript and having verified through different approaches all ambiguous alleles (see Supp. info 2). This information is presented in Table S2. Between 20 and 30 clones were sequenced for each individual of the present work. The selection of individuals for cloning in each population was based on phasing difficulties with direct sequencing chromatograms only (see Supp. Info 2), to confirm each allele for each nuclear marker.

3/ I guess the invasive asexual lineages have a very different demography compared to the non-invasive ones. Would it be possible to compare the structure of haplotype networks for these two types of lineages?

The invasive and native lineages share alleles, it would therefore seems inappropriate to separate those lineages in our analyses. In our manuscript, we highlight two main results: 1) the extensive genetic mixing between freshwater *Corbicula* "species" (from both the native and

invasive regions) which prevents the clear identification of taxonomic clusters, 2) the distinct biogeographic origins of invasive *Corbicula* lineages since we were able to identify three different genetic clusters containing invasive lineages (with limited allele sharing between these 3 FFRs).

L 21: Maybe define "all-male asexuality" as this term is does not really say more than "androgenesis".

We agree this is not self-explanatory. This term is not used anymore in the new manuscript.

L25-27: These two sentences are not very clear.

This section was completely modified.

L62-119: I find this section hard to follow. Adding a figure with a map would help simplify the text, using a color code similar to the one in Fig. 2 would be useful. I would suggest streamlining this section by listing the different androgenetic lineages and where they are found so that the reader can understand that some forms are invasive (A/R, B, Rlc and C/S) and some others are not (e.g. Vietnam).

This section was modified in the revised manuscript (lines 129 - 150) to make this point clearer.

L 97: "Nuclear capture between androgenetic and divergent sympatric sexual or androgenetic lineages"?

This sentence was rewritten.

L126: "form" is a bit obscure, maybe it would possible to be more specific and to mention if these taxa are defined based solely on morphological data or on both morphological and genetic data.

We agree these notions of lineages and forms are confusing. We tried to clarify this in the manuscript (lines 127 - 135) while still retaining the nomenclature used in the Pigneur papers.

L 163: How would the strong departure from Hardy-Weinberg (high number of heterozygous loci in asexual lineages) affect the phasing?

With a statistical-based phasing (such as the use of PHASE software only, see Supp. Info 2), this could indeed be an issue. This is also for this reason that we increased the number of samples that were controlled via cloning (see Table S2).

L 256 "and C. fluminalis africana away from FFR4 individuals?

This sentence was modified.

Fig. 2: I found the captions not very informative. Maybe you could explicitly mention the androgenetic lineages and the sexual species in the caption?

The captions have been modified, we hope they are more informative now. However, it is not easy to represent the reproductive mode of each lineage in this figure. We rather choose to add it on figure 4.