Editor

Dear Dr Jaron,

We have received three thoughtful reviews of your manuscript entitled “Genomic evidence of paternal genome elimination in globular springtails”. The three referees are globally positive, as well as I was when I accepted to handle this preprint for a recommendation. As it is current practice, I’ll ask you to revise your ms according to referees’ concerns and provide a cover letter where you explain how you modified the ms. Referee 1 and 3 have only a few minor concerns to improve clarity that you should easily account for. Referee 2 who signed his review has a longest list of comments. This referee will likely see the revised version. I am looking forward to reading your revised ms and to work on a nice recommendation to advertise your interesting work.

Best regards,

Nicolas Bierne

Dear Nicolas,

Thank you for the evaluation of our manuscript. We have revised the manuscript according to the comments of the three reviewers.

The major change in the revised manuscript is that we provide a power analysis that shows biological and technical conditions for which the proposed two tissue model is applicable. The benchmarking made us re-evaluate how we construct the two tissue model and we ended up fitting the model to k-mer spectra instead of coverages of mapped reads. These adjustments allow us to analyse any whole genome sequencing dataset without need of a reference genome, making our technique more generally applicable.

Finally, we also worked hard to improve the logical flow of the manuscript. Now we more explicitly separate the role of the two tissue model for estimating the fraction of a tissue with a different karyotype and the role of the PGE model we use to test autosomal haplotype presence in sperm.

We will be looking forward to hearing back from you.

On behalf of all authors,
Kamil S. Jaron

Reviewer 1

This article present a new genomic approach for detecting an interesting type of reproduction: paternal genome elimination. In this reproductive mode, males inherit paternal and maternal genomes but only maternal genetic material is present in their spermatozoids. To detect such a reproductive mode, the authors suggest to check whether X-linked alleles show similar coverage than the major autosomal allele, which would be indicative of an excess of maternally inherited genetic material in sperm.

While I do not have the expertise to judge the mathematical aspects of the method, the approach is nicely presented with clear figures explaining the rationale, and the main result showing that Allacma fusca males practice paternal genome elimination is convincing. As authors aknowledge it, this approach is only possible in males with large quantity of sperm. While they discuss accordingly the fact that many invertebrates feature high proportion of male germ cells, I would like to know if they can estimate the fraction of sperm necessary to significantly detect paternal genome elimination depending on genome coverage. I think it would be a nice addition to help future scan studies using this kind of approach on several other species.

This is a great suggestion, we implemented a power analysis that simulates genomes with the same composition as Allacma fusca while varying the fraction of sperm, size of X chromosomes, heterozygosity and sequencing depth (L226 - 229, and SM Text 5). We showed that this approach has the potential to reveal abnormalities from about 10% of cell count with different ploidy ratios (in theory it does not have to be AAX0 and AX but any type of A to X variation).

The analysis also revealed that our estimator is conservative and underestimates the real fraction of sperm in the sample. We added a section in the discussion to reflect these new findings.

Also, I suggest to edit the title to be more spedific about the fact that the main result concerns only one species of globular springtails.

*The title is now "***Genomic evidence of paternal genome elimination in globular springtail *Allacma fusca****".*

Minor remarks:
L40: fix nested parentheses of citation
L50: I do not think that the verb "culture" can be used for insects
L53: citations should be placed before the comma
L98: add a comma after "In fungus gnats and gall midges"
L272: contains
L467: change "seems to be present already" on "seems to be already present"
L493: fix nested parentheses of citation
L513: fix nested parentheses of citation
Fig1: "heterochmatized" does not seem to be an existing word

We fixed all these issues.

Reviewer 2

PCIEvolBiol #525

Genomic evidence of paternal genome elimination in globular springtails
Kamil S. Jaron, Christina N. Hodson, Jacintha Ellers, Stuart JE Baird, Laura Ross
https://doi.org/10.1101/2021.11.12.468426

Overview
In this study, the authors analysed re-sequencing genomic data in order to investigate an established hypothesis that Paternal Genome Elimination occur in globular springtails. I believe their reasoning is sound, the results interesting, the study convincing and the title accurate. My criticism are only towards the clarity of the different steps taken by the authors to reach their conclusion and the length of the main text. This study is of good quality, however the formating of its manuscript should be improved.

The interest of some analyses is unclear to me, the limits of some interpretation are not explicitely stated and the potentiality of competing interpretation feel sometimes overlooked. I believe this study could be thoroughly stream-lined, by only keeping the necessary and sufficient results in order to build the argumentation towards the likely existence of PGE in globular springtails. This would i) ease the reading, ii) shorten the manuscript, and iii) and create space to discuss the limits and competing hypotheses possibly explaning the results presented. I do believe these limits do not hamper the authors to reach their conclusions, and that their final hypothesis is the most likely to date. But their conclusion will be even stronger after having taken the time to explain how potential issues are unlikely to impact the results or that they do not hamper to reach the author’s conlusions. Also it would be the opportunity to clarify precisely what conclusion can be drawn by what analyses. Taken alltogether, I agree with the interpretation of the results, but the process to get to this conclusion felt convoluted, sometimes spending time on trivial aspects and unecessary figures, sometimes not discussing or investigating potential issues.

Note that this reviewer is not an expert in these specific reproductive modes (PGE) and their evolutionary impact on organisms. These aspects of study were thus not « properly » reviewed here, and I would hold with my overall positive review as long as these aspects are considered correctly tackled by other reviewers (e.g. the evolution of PGE in arthropods being correctly depicted).

We thank the reviewer for an overall positive review. We improved the general flow of the manuscript by moving some of the details into supplementary materials (e.g. the Box 1) and separating more clearly the two models we apply to the data (two tissue and PGE model). We also addressed the issue of using an untested approach using simulation-based power analysis. See point to point answer below for details.

Mandatory revisions

line 109 : inappropriate figure reference : SM Figure 1 do not really show the elimination of one of the X1 and X2 chromosomes during early embryogenesis. Or it does but I didn't understand how. In both case, the authors need to modify this figure in order for the reader/reviewer to understand how it shows elimination of X1 and X2 chromosomes. Maybe I simply did not find the « Supplementary Figure 1 » ? In any case, this needs to be fixed/clarified.

We removed the reference (it was a reference we forgot to delete from the previous version of the manuscript). We unified all formats for supplementary figures (everything is now in "Figure SMX" format.)

Line 139-141 : This study does not « demonstrate » uniparental inheritance, it strongly suggests it. This is correctly stated in the abstract, but the sentence here is incorrect. Also, using the wording « co-segregate » is unclear (at least for me) regarding whether the segregation of X chromosomes and autosome co-occur (i.e. same timing) or is done in two different steps. This should be clarified here.

We changed the wording to "strongly suggests".

However, we kept the expression "co-segregation". Co-segregation of alleles is a common term to denote joint inheritance of alleles, usually due to physical linkage. Although the term is also used for chromosomes, for example in species where multiple chromosomes are involved in sex determination (10.1038/hdy.2016.22). To our knowledge, co-segregate is the only used term for this phenomenon.

We added "(i.e. are transmitted together)" to remind the reader of the meaning of this expression.

Line 146 : The figure 2 title should clearly establish that this is a hypothetical model, a working hypothesis. As of now, it might be confused by readers as a conclusion of this study (thus appearing way too soon in the manuscript), or worse, as a state-of-the-art introductory figure (which would render the entire study unnecessary, since PGE would already have been shown). Maybe a title like « Working hypothesis for PGE in globular springtails » ? Another solution (my favorite) would be to place this figure at the end of the manuscript, after PGE is strongly suggested by the results, as a scheme of the new, up-to-date, working hypothesis that the reader will remember.

We clarified it's a "model" in the title of the figure. Besides the figure itself we streamlined both method and result sections to make clear what are the individual questions we are testing at each step.

Line 194-198 : Instead of showing difficult-to-interpretate k-mer spectra due to unevenly spaced ploidy peaks, the authors shoud first (or only) show the distribution of mapping coverage on scaffolds (SM Fig 2 panels B, D and F). This would make understanding the results easier for the readers. Note that If they wish so, the authors they can later show the corresponding k-mer spectra. But these spectra do not add any value to this study, as what makes them more informative (i.e. estimating genome size and heterozygosity) is useless in the context of this manuscript. Consequently, this also diminishes the interest of the Box 1 in the main text. This box is not important to understand the study, and should be either removed or placed within supplementary text if the authors strongly want to keep it in the manuscript.

The major advantage of the fits of the k-mer spectra is that they are very generally applicable, whereas mapping coverage both requires a reference and suffers from all the issues of reference bias and is incomparably more computationally demanding. However, we agree the wording we have chosen was unnecessarily confusing and not very concise.

In the revisions we moved the mapping coverage part to supplements as well as Box 1. The main presented results of the two tissue models are now based purely on the k-mer spectra analysis, which allows us to drop the requirement of a reference genome to estimate the fraction of sperm in a body and streamlines reading of the manuscript.

The reference genomes are used later in the text to test PGE using coverage support of heterozygous alleles.

Line 279: Figure 4 is referred to before Figure 3 (line 342). These two figures should thus be switched in the manuscript. See also my remark on placing Figure 2 at the end of the Manuscript.

The two figures are now referred to in the order in the text. We worked hard to make a clear step-wise logical flow of Two tissue model (Figure 3) used without assemblies and allowing us to create expectations for the PGE model (Figure 4). Hopefully this will clarify why the figures simply can't be in the reverse order without breaking this flow.

The estimation of the fraction of sperm cells heavily rely on the accuracy of the peak of the allele frequency distribution shown in SM Figure 5. However, this distribution is very flat (thus, the exact position of the peak is uncertain).

*The fraction of sperm can be calculated either from the coverage peaks (using two tissue model) or allele coverages (using PGE model). In the current revisions we made clear that the most robust estimate of the fraction of sperm is from the positions of the two peaks on **Figure 3**. We also added a table with all the sperm estimates (**SM Table 3**) and discussed the robustness of these estimates.*

Finally, now we also conducted a power analysis that shows the cases for which we get robust estimates of the fraction of sperm (L226 - 229, SM Text 5).

While trying to estimate the fraction of sperm cells based on these data is very interesting, the authors should explicitely mention the inherent inaccuracy of these estimate every time they mention it in the manuscript. This does not seem to be a strong results of the study and should not get the quantity of attention it currently gets from the authors. Also, for clarity and for convincing the readers that their method is sound, they should also compute and present the corresponding fraction of sperm cells in Ocin2, as its peak is also not exactly positionned at 0.5. (of course, it is impossible to placed at exactly 0.5).

We appreciate this suggestion. The estimated Ocin2 fraction of sperm using the previously presented method (the mapping coverage based) is 10.3%. This is apparently wrong and rather disturbing. We investigated where it might come from and found that the reference assembly of O. cincta contains several long scaffolds that have coverage in between of 1n and 2n peaks - these are likely miss-assembled (chimeric) scaffolds made out of X-linked and autosomal contigs and as a consequence causing bias in estimates of 1n and 2n coverages.

To avoid this we implemented a k-mer spectra estimate of 1n and 2n peaks. With this model we found a negligible fraction of sperm in O. cincta (0.58%), which is within expected error margins (the positions of 1n and 2n peaks are not significantly different to 1:2 coverage ratio). All fraction of sperm estimates are now sorted in table SM Table 3.

Finally, they should also try their approach on an organism for which this fraction is known. As long as this is not done, I would definitely tone down theses results in the manuscript for two reasons : i) their accuracy is debatable and, more importantly, ii) they do not really help answering whether or not PGE occurs in globular springtails and are somewhat a distraction from the main argumentation line of this manuscript.

Our genomic test requires the average sperm karyotype to be different to soma. While there are many systems where this is true, none of them is studied well enough to have precise quantification of their sperm content in their body. Therefore such comparison is impossible at the moment. Instead, we conducted a power analysis that provides a clear picture about the robustness of the analyses.

Additionaly, authors should clearly indicate the respective sequencing depth of the data used for BH3-2 and Ocin2 in the main text and discuss it. Indeed, a much larger coverage in Ocin2 might be sufficient to produce a minor allele frequency distribution closer to 0.5, that is, with more accurate SNP-calling. Alternatively, the order of magnitude higher number of heterozygous variants in Ocin2 might be sufficient to produce a much clearer peak, potentially mechanisticaly closer to 0.5 (i.e. even if every single SNP is called with the same accuracy as for BH3-2).

We added two female samples with comparable coverage to the male sample to the ***SM Figure 9****. In both cases the coverage ratio is a lot closer to 0.5 than expected. We also added a more detailed explanation to the legend of the SM Figure 9.*

Overall, I believe the estimate of the fraction of sperm cells need to be taken with caution, and that the authors should discuss about the limitation of their otherwise interesting approach. Authors repeatedly uses PGE as the explanation of a coverage shifts. While this is a possible explanation, other types of GE, that are non-paternal genome elimination (or even simply different ploidy levels within an organism) could also produce such a shift.

We explored and presented other scenarios explaining the coverage shift (SM Figure 7). Our test subsequently narrowed down the search to a systematic genome elimination during spermatogenesis. On lines 224 - 225 we refer to SM Text 4, which discusses these scenarios in more detail.

However, the results presented in figure 4C clearly and convincingly suggest PGE in Allacma fusca (or at least in indvidual BH3-2). I thus suggest the authors to clarify their argumentation line, decomposing it in two steps : i) analysis of coverage data + interpretation of the peak being shifted allow to confirm various ploidy levels in the tissue (interpreted as due to genomic elimination, Figure 3) ; ii) coverage distribution of maternal allele fitting the distribution of « major » autosomal allele strongly suggests that the eliminated chromosome are, in fact, of paternal origin. I believe this would clarify the argumentation line of the study, explicitely stating the relative contribution of the different analyses to the main conclusion. This would also help to stream-line the writing.

In revisions of the manuscript we made a clearer separation of the two logical steps of our analysis - the two tissue model and the PGE model as the most likely explanation of the two tissues.

The authors should explicitely state whether their analyses bring light (or not, which would not be an issue) on the tempo of PGE : with their data, can they re-inforce or contradict the hypothesis that chromosome elimination happens in one step, or in two steps (elimination of X chromosome in early spermatogenesis, and elimination of chromosomes later during the process) ? I believe the answer is no : these data can not validate or invalidate this hypothesis, but this should be discussed in the manuscript, in order to clarify to the reader the state-of-the-art knowledge on reproductive mode in globular springtails.

As the scheme (Figure 2) indicates, the X chromosome elimination happens during embryogenesis and it was documented cytologically. The autosomes are not strictly speaking eliminated (in the sense of chromosomal elimination as known from germ-line restricted chromosomes). Instead, one of the two primary spermatocytes degenerate (the one that misses the two X chromosomes), which means the two X chromosomes are “eliminated” at the same time. The two step process is known from cytological studies and is described and referred throughout the manuscript.

Moreover, our data indeed is consistent with the previously observed cytological observations. For example we revealed no heterozygous sites on X chromosomes (consistent X chromosome elimination during early embryogenesis) and we detected only a single genotype in the sperm (consistent with aberrant spermatogenesis). All this is addressed in the manuscript.

The last paragraph of the discussion section should be removed, or placed elsewhere (maybe as supplementary text, or at the beginning of the discussion section). I personally do not think that this study « demonstrate the power of a careful bioinformatics analysis » : this is vague and emphatic. I do believe however that discussing the importance of quality checks and data exploration is interesting for the readers. This should however not be placed as the final paragraph of this study about reproductive mode of globular springtails.

We removed the paragraph and now discuss the generality of the approach together with the power analysis.

English spelling and typos need to thoroughly checked throughout the manuscript.

Suggestions, corrections and « typos »

While we appreciate the individual spotted typos, these generic comments are unnecessary.

line 35 : the current sentence finishing the first paragraph is not particularly convincing, it feels a bit « off » the rest of the paragraph. I suggesting re-phrasing it.
Table S1, row 11 : « confromation » => « confirmation »
line 113 : « The X chromosome lacking spermatocytes » => «The spermatocyte lacking the X chromosomes »
line 138 : « in-silico bioinformatics » is redundant, pick your favorite.
Line 138 : « to separate the effect of somatic and germline genomes ». What effect are your referring to ? Please re-phrase.
Line 139 : Please refrain to use wordings such as « innovative ». It is close to meaningless and feels emphatic. Readers will decide in the future wether this study is « innovative » or Not.
Line 165 : The authors need to explicitely specify whether DNA amplification step was used in this study during the production of cDNA libraries. Indeed, this would have an importance for later analyses of allelic frequency and the determination of maternal versus paternal chromosomes.
Line 171 : « the bistates » => « biallelic SNPs »

These are not "biallelic SNPs". We added an explanation, the section reads now "We used genomic positions with two nucleotides with coverage >1 mapped to it. This approach showed a higher abundance of these bistates around coverage ratios" (L: 410 - 411)

Line 187 : « monoploid ». Either the authors have a reason to not use the word « haploid » and should justify their choice within the text, or they should use « haploid ».

Haploid (although often misused) means "reduced". Meiotic product is haploid regardless of its ploidy. For regular diploid species, indeed haploid often means monoploid, but that's not the case for polyploid species, or species with distorted segregation patterns (like fungus gnat also discussed in the paper). It does not work the other way around either - monoploid captures also heterozygous loci, those are by no means "haploid".

Nevertheless, discussing this nomenclature is beyond scope of our manuscript. Especially given how prevalent are mistakes in modern genomic literature.

SM Figure 3 title : typo « allelie »
Line 214 : SM Figure 9 caption refers to SM Figure 9.
Line 472 : re-phrase the entire sentence, as the paragraph specifically explain that PGE is NOT the only explanation compatible with biology.
It is my opinion that the overall inclusion of misassigned variants (results and discussion) takes too much space in the current manuscript. While it is good the authors checked it, it did not occur to me when reading the study that it could have had a strong impact on the results. I suggest the authors mention this check in the manuscript, but reduces its Importance.

We believe the careful consideration of separating alleles by coverages strengthens our point. However, we agree some parts of the manuscript were overly convoluted for no reason and streamlining the manuscript using suggestions of all three reviewers helped us to make the story more to the point without the necessity of hiding the care we took while analysing the data.

Reviewer 3

Review by anonymous reviewer, 13 Dec 2021 14:59

In this paper, Jaron et al. test for systematic paternal genome elimination in a species of globular springtails (Allacma fusca) by developing computational approaches to disentangle germline and somatic genome contributions to a whole-body sample of gDNA. Specifically, the authors develop a two-tissue mixture model (soma and cells having undergone paternal genome elimination, here, secondary spermatocytes and mature sperms) to estimate the respective contribution of these tissues to the whole-body sample of gDNA and, hence, the sequencing library. This allows them to formally test for systematic paternal genome elimination based on maternal and paternal sequencing coverages. They also validate their approach with a control species (Orchesella cincta) that has XO sex determination. Overall, this manuscript is well written, the methods are clear, very well described and most likely reproducible. I only have a few minor comments that mostly involve clarifications.

- I found the wording “aberrant spermatogenesis” (L98) a bit confusing, as paternal genome elimination appears to be systematic. I understand this wording is a legacy from earlier publications, but I wondered if it could be reformulated at this point, given the evidence provided by the authors.

We removed the word “aberrant” when we describe the fly systems (previous L98), in the next paragraph we explain in greater detail what we mean by aberrant spermatogenesis, so the text is still compatible with the previous literature (L: 114 - 118).

- A minor caveat to the approach developed here is that it requires a reference genome assembly. This should be mentioned somewhere.

Indeed that was the case in the original preprint. However, in this revision we adjust the first step of the analysis from “mapping based coverages” to “kmer coverages”, therefore the technique is reference free now (which is the reason we made the change). We also streamline the text so the Two tissue model and PGE model are two clearly distinct steps, where only the second step needs a reference genome.

ABSTRACT

L21-23: “The genomic approach we developed allows for detection of genotypic differences between germline and soma in all species with sufficiently high fraction of germline in their bodies.”
- This statement is vague (what is a sufficiently high fraction?). Might it be possible to evaluation what is a sufficiently high fraction either by modelling or sequencing read resampling? If not (or beyond the scope of the present study), I suggest rephrasing.

As we got this request also from the Reviewer 1, we decided to perform a power analysis (L226 - 229, and SM Text 5), it has been very helpful to understand our approach better - we thank the reviewer for this suggestion.

- This statement is also potentially more generalizable ("high fraction of germline in their bodies") to include a tissue sample (e.g., gonads) containing a mixture of somatic and germline cell (and not only whole-body samples) in larger organisms (e.g., birds).

That is very true, and we are investigating other applications already. However, this manuscript presents only one model only and the model as we share it is specifically applicable to this particular karyotypic difference between two tissues. We added the possible more general applications of the model to the discussion (L: 439 - 444).

INTRODUCTION

L59: I would remove the comma.
Figure 1: Heterochmatized --> heterochromatinized? (or heterochromatinised)

We used (this time an existing) word "Heterochromatic".

DISCUSSION
L464: I wondered if "hybridogenesis" should not be mentioned as well for the Australian carp gudgeons.

We added a mention.

L505: How could PGE affect the evolution of reproductive isolation and increase diversification rate? Could it be related to the lack of recombination in males and strict maternal inheritance to reduced effective population size? I would clarify the rationale here.

We added a sentence briefly explaining the main difference to diplo-diploid hybridization: "caused by a generation lag of hybrid males that can be produced only if the mother is a hybrid already".

SUPPLEMENTARY MATERIAL
L129: I think this sentence is incomplete :)

We fixed the sentence (now L220 in supplements).

Github repository: some parts could be cleaned up a bit, but everything seems to be there.

We updated the GitHub repository to be more organised. Now it also contains quite a lot of data directly to the repository hence many of the figures and analyses can be reproduced.