THE ROYAL SOCIETY
PUBLISHING

PROCEEDINGS B

The genome of a daddy-long-legs (Opiliones) illuminates the evolution of arachnid appendages

Guilherme Gainett, Vanessa L. González, Jesús A. Ballesteros, Emily V. W. Setton, Caitlin M. Baker, Leonardo Barolo Gargiulo, Carlos E. Santibáñez-López, Jonathan A. Coddington and Prashant P. Sharma

Article citation details
Proc. R. Soc. B **288**: 20211168.
http://dx.doi.org/10.1098/rspb.2021.1168

Review timeline

Original submission:	15 February 2021
1st revised submission:	21 May 2021
2nd revised submission:	5 July 2021
Final acceptance:	14 July 2021

Note: Reports are unedited and appear as submitted by the referee. The review history appears in chronological order.

Review History

RSPB-2021-0372.R0 (Original submission)

Review form: Reviewer 1

Recommendation
Accept with minor revision (please list in comments)

Scientific importance: Is the manuscript an original and important contribution to its field?
Excellent

General interest: Is the paper of sufficient general interest?
Excellent

Quality of the paper: Is the overall quality of the paper suitable?
Excellent

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Reports © 2021 The Reviewers; Decision Letters © 2021 The Reviewers and Editors; Responses © 2021 The Reviewers, Editors and Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, provided the original author and source are credited

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
No

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
Yes

Is it clear?
Yes

Is it adequate?
Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
Gainett et al present the genome of Phalangium opilio alongside data on the genetic basis for the elongate legs seen in this species. This work is well-written and presented. In general, the authors are to be commended on the clarity and openness of the presentation of their methods and data.

There is one major issue with the manuscript as written that I would request is corrected before publication. I also have a number of minor comments and further suggestions, as noted below. Please also note that the UCSC genome browser link provided did not allow me access to the browser, and a more user-friendly URL should be generated and provided before publication - something like a tinyurl could work well.

Major comment:
The removal of genes from the annotation described on lines 146-148 of the supplement is not clearly described and could have a major impact on the analysis and interpretation of the results.
1) There needs to be clarity as to what "did not generate significant BLAST hits, functional annotation, or lacked transcript evidence" means. This could result in the exclusion of many excellent gene models. Please provide full details of criteria used for exclusion, and a summary of how many genes were thus removed.
2) "the gene set was then further refined with a 98% similarity threshold using CDHIT" - again, this could result in the trimming of many excellent gene models. This would particularly impact the Hox and Ks analyses - if close paralogs exist, they would be deleted by this step. Please
- note how many genes were thus removed
- provide clear evidence that no close matches to hox cluster genes were removed
- repeat the Ks analysis for Phalangium opilio on the untrimmed set if large numbers (perhaps >500?) genes were removed from the data shown, to ensure that paralogs (which could result from WGD) were not removed inadvertently by this step
3) please provide details as to how genes noted as duplication were removed from the gtf file. Was the shortest copy removed in all cases, for example? The CDHIT command given shows how the amino acid file was trimmed, but not how a consensus gtf (and thus the final set of gene models) was generated.

Minor comments:
Blobplot: please give blast/diamond settings used for the blast step in this identification. It is very surprising that "no hit" is by far the most common result, and no arthropod "hit" was found within the majority of sequences. This could be explained by overly stringent cutoffs, a poor

target library for blasting, or alternatively, if no gene model is present on the vast majority of these contigs. Please provide further information, as your gene annotation process suggests that most gene models should have a good blast hit to something. It is perhaps worth considering redoing the blobplot with more relaxed blast criteria.

-Table S5 or in text - please provide information as to what "Ns" in the genome represent. If these are consistently of a given size (e.g. 1000bp when inserted during scaffolding) please make this clear.

-Please provide a supplementary table summarising the results of your RepeatMasker/RepeatModeler analysis. This will provide basic information on repetitive elements for the community, as well as making it clear what %age of the genome has been soft masked.

Further suggestions:
- Fig 2C placement of node labels could be misconstrued, as the placement of these often is suggestive of wider groups than intended, as they are often slightly more towards the root than the clades they note. I suggest these are moved closer towards the node to which they refer, even if the figure needs to be increased in size.

- Line 265: maritma misspelled

- Supplementary: possible redundancy, consider rewriting, line 54: The Single Molecule Real-Time (SMRT) Cells were sequenced on (16) SMRT cells

Review form: Reviewer 2

Recommendation
Reject – article is not of sufficient interest (we will consider a transfer to another journal)

Scientific importance: Is the manuscript an original and important contribution to its field?
Excellent

General interest: Is the paper of sufficient general interest?
Marginal

Quality of the paper: Is the overall quality of the paper suitable?
Excellent

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
Yes

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
No

Is it clear?
N/A

Is it adequate?
No

Do you have any ethical concerns with this paper?
No

Comments to the Author
This paper brings an emerging model system in evo-devo, the daddy-long-legs Phalangium opilio into the genomic era. From a phylogenetic perspective, this is significant for two reasons. The first is developmental data from chelicerates are essential for robust inferences about the ancestral state in arthropods as a whole Second, developmental comparisons to other arthropods are simplified (both methodologically and conceptually) because the lineage is shown not to have undergone any whole genome duplications after diverging from the lineage that gave rise to the other major arthropod groups (whereas spiders and scorpions have). After presenting the genome, the paper presents RNA-based functional analyses of two conserved genes, the Hox gene Deformed and EGFR. The results are largely as one would predict, with some potentially interesting implications for the evolution of arthropod appendages.

The paper is very clearly written, the analyses appear well done, and each piece of the paper appears sound (with a few minor quibbles—see below). However, the genome and the two RNAi phenotypes did not come across as complementary—neither part strengthened the other part. Nor did there appear to be a broader conclusion that was enabled by combining these sets of data. As a consequence of doing so much, everything is short-changed (or in some cases not followed through in as great a depth as I would expect). In sum, this is high quality science, with multiple results that are individually interesting, but in my assessment it does not yield insights that I would expect to be compelling to a broad readership (beyond being about an interesting critter).

The single genes had clear phenotypes and there was some discussion about the ways in which these phenotypes were similar to and different from the phenotypes resulting from functional knockdown of homologous genes in other arthropods. However, the developmental work would be more compelling if it were more completely fleshed out. For Hox genes, this would mean generating RNAi phenotypes for all of the genes (and combinations of coexpressed genes), or at least the set that are expressed in legs. For example, is the absence of leg-pedipalp transformation in L3 and L4, which also express Dfd, due to Scr (expressed in L3 and L4, but not L1 or L2) also specifying leg identity? (Knockdowns of Scr on its own and the double Dfd/Scr knockdown should be done.) Also, as the authors acknowledge (l. 350-352), the homeotic effects can't be distinguished from a role regulating growth in L2, and thus the comparisons to Ubx in waterstriders are premature. For EGFR, the authors suggest that both early functions (distal tip) and late functions (related to expression in each leg segment) are conserved, based on their phenotype analyses. However, the expression data do not suggest the existence of a distal signaling center. Thus, expression/function of additional signaling pathway components are needed to make sense of this result.

Questions about results/interpretation:

Spider/scorpion Ks frequency distributions look very similar to the distribution in Phalangium. From this, it's not how these plots support/add evidence to the conclusion that Phalangium has not undergone a whole genome duplication while those other taxa have. It would also be helpful to give the number of gene families analyzed for each taxon, as large difference in this would

perhaps be meaningful. (Also, note that order of species in legend differs from order in figures.)

Are there features of embryogenesis that account for the strong asymmetries in RNAi phenotypes encountered?

Many of the support values shown in the EGFR tree are extremely low. In particular, the branches on which the conclusion of three independent duplication events within arthropods are based do not receive support (with the exception that the branch grouping the two Phalangium sequences appears robust). Thus, the scenario of there independent duplication events should not be presented as as well established as the text suggests, especially if there's a chance that subsequent gene conversion in any part(s) of the protein-coding region could have increased similarity between paralogues after duplication. (If there is other information that supports it, for example, gene structure features, or strong support in the UTRs than in the coding region, that would allay these concerns.)

Table S5: I'm puzzled that 15 BUSCO genes seem to have gone missing as a result of the purge haplotig command. Does this makes sense? If so, what happened to them? It would seem highly likely that they were real.

Data availability:
The EGFR alignment and tree should be made available in text/machine readable form.

Additional notes:
l. 142-144: Given the size of the dataset, there's no need for the computational shortcuts such as fast bootstrapping; a full analysis can be performed. Also, What model of amino acid substitution was used? (This may be implicitly stated through the commands, but please also state explicitly.)

l. 204-206: It looks like half the sentence went missing.

l. 270-271 and figure S8: truncation of EGFR-B. It would be helpful for readers to know you have confirmed that the truncation is genuine and not an artifact of incomplete genome or mRNA assembly.

Several supplemental figures appeared lower quality than necessary to be readable at the size presented. (This could be due to pdf processing through the manuscript handling system, but if not, then originals should be magnified/improved.) Affected figures included S3, S4, S7, S8

Figure S5: It took me a while to understand how some of the Dfd phenotypes could become wild type. Also, I recommend stacking the conditions in order of severity (so strong transformation at the top). And I wonder whether the figure would be more informative if it showed the bilateral combination frequencies (WT/weak, WT/Strong; weak/weak; weak/strong; strong/strong). Especially if the two sides are not independent, this seems like a more representative summary of the data.

Decision letter (RSPB-2021-0372.R0)

27-Mar-2021

Dear Mr Gainett:

I am writing to inform you that your manuscript RSPB-2021-0372 entitled "The genome of a daddy-long-legs (Opiliones) illuminates the evolution of arachnid appendages" has, in its current form, been rejected for publication in Proceedings B.

This action has been taken on the advice of referees, who have recommended that substantial revisions are necessary. With this in mind we would be happy to consider a resubmission, provided the comments of the referees are fully addressed. However please note that this is not a provisional acceptance.

The resubmission will be treated as a new manuscript. However, we will approach the same reviewers if they are available and it is deemed appropriate to do so by the Editor. Please note that resubmissions must be submitted within six months of the date of this email. In exceptional circumstances, extensions may be possible if agreed with the Editorial Office. Manuscripts submitted after this date will be automatically rejected.

Please find below the comments made by the referees, not including confidential reports to the Editor, which I hope you will find useful. If you do choose to resubmit your manuscript, please upload the following:

1) A 'response to referees' document including details of how you have responded to the comments, and the adjustments you have made.
2) A clean copy of the manuscript and one with 'tracked changes' indicating your 'response to referees' comments document.
3) Line numbers in your main document.
4) Data - please see our policies on data sharing to ensure that you are complying (https://royalsociety.org/journals/authors/author-guidelines/#data).

To upload a resubmitted manuscript, log into http://mc.manuscriptcentral.com/prsb and enter your Author Centre, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions," click on "Create a Resubmission." Please be sure to indicate in your cover letter that it is a resubmission, and supply the previous reference number.

Sincerely,
Professor Gary Carvalho
mailto: proceedingsb@royalsociety.org

Associate Editor
Board Member: 1
Comments to Author:
I enjoyed reading the paper by Gainett and collaborators, in which they present the genome sequence of an opinion and use it combined with other omics approaches and gene expression data to infer the genetic basis of the evolution of appendages in chelicerates. The manuscript is well-written, nicely presented, and the analyses are sound. I'd like to congratulate the authors on their work.

I agree with most of the concerns raised by the two expert reviewers, which for the most part are are clarifications and requests for more information. I'd like to encourage the authors to address these concerns in order to improve the quality of their manuscript, including the tone of the claims about the homeotic patterns of the manuscript, claims made relating to expression (early and late effects), and comparison with water striders.

Reviewer(s)' Comments to Author:
Referee: 1
Comments to the Author(s)
Gainett et al present the genome of Phalangium opilio alongside data on the genetic basis for the elongate legs seen in this species. This work is well-written and presented. In general, the authors are to be commended on the clarity and openness of the presentation of their methods and data.

There is one major issue with the manuscript as written that I would request is corrected before publication. I also have a number of minor comments and further suggestions, as noted below. Please also note that the UCSC genome browser link provided did not allow me access to the browser, and a more user-friendly URL should be generated and provided before publication - something like a tinyurl could work well.

Major comment:
The removal of genes from the annotation described on lines 146-148 of the supplement is not clearly described and could have a major impact on the analysis and interpretation of the results.
1) There needs to be clarity as to what "did not generate significant BLAST hits, functional annotation, or lacked transcript evidence" means. This could result in the exclusion of many excellent gene models. Please provide full details of criteria used for exclusion, and a summary of how many genes were thus removed.
2) "the gene set was then further refined with a 98% similarity threshold using CDHIT" - again, this could result in the trimming of many excellent gene models. This would particularly impact the Hox and Ks analyses - if close paralogs exist, they would be deleted by this step. Please
- note how many genes were thus removed
- provide clear evidence that no close matches to hox cluster genes were removed
- repeat the Ks analysis for Phalangium opilio on the untrimmed set if large numbers (perhaps >500?) genes were removed from the data shown, to ensure that paralogs (which could result from WGD) were not removed inadvertently by this step
3) please provide details as to how genes noted as duplication were removed from the gtf file. Was the shortest copy removed in all cases, for example? The CDHIT command given shows how the amino acid file was trimmed, but not how a consensus gtf (and thus the final set of gene models) was generated.

Minor comments:
Blobplot: please give blast/diamond settings used for the blast step in this identification. It is very surprising that "no hit" is by far the most common result, and no arthropod "hit" was found within the majority of sequences. This could be explained by overly stringent cutoffs, a poor target library for blasting, or alternatively, if no gene model is present on the vast majority of these contigs. Please provide further information, as your gene annotation process suggests that most gene models should have a good blast hit to something. It is perhaps worth considering redoing the blobplot with more relaxed blast criteria.

-Table S5 or in text - please provide information as to what "Ns" in the genome represent. If these are consistently of a given size (e.g. 1000bp when inserted during scaffolding) please make this clear.

-Please provide a supplementary table summarising the results of your RepeatMasker/RepeatModeler analysis. This will provide basic information on repetitive elements for the community, as well as making it clear what %age of the genome has been soft masked.

Further suggestions:
- Fig 2C placement of node labels could be misconstrued, as the placement of these often is suggestive of wider groups than intended, as they are often slightly more towards the root than the clades they note. I suggest these are moved closer towards the node to which they refer, even if the figure needs to be increased in size.

- Line 265: maritma misspelled

- Supplementary: possible redundancy, consider rewriting, line 54: The Single Molecule Real-Time (SMRT) Cells were sequenced on (16) SMRT cells

Referee: 2

Comments to the Author(s)

This paper brings an emerging model system in evo-devo, the daddy-long-legs Phalangium opilio into the genomic era. From a phylogenetic perspective, this is significant for two reasons. The first is developmental data from chelicerates are essential for robust inferences about the ancestral state in arthropods as a whole Second, developmental comparisons to other arthropods are simplified (both methodologically and conceptually) because the lineage is shown not to have undergone any whole genome duplications after diverging from the lineage that gave rise to the other major arthropod groups (whereas spiders and scorpions have). After presenting the genome, the paper presents RNA-based functional analyses of two conserved genes, the Hox gene Deformed and EGFR. The results are largely as one would predict, with some potentially interesting implications for the evolution of arthropod appendages.

The paper is very clearly written, the analyses appear well done, and each piece of the paper appears sound (with a few minor quibbles—see below). However, the genome and the two RNAi phenotypes did not come across as complementary—neither part strengthened the other part. Nor did there appear to be a broader conclusion that was enabled by combining these sets of data. As a consequence of doing so much, everything is short-changed (or in some cases not followed through in as great a depth as I would expect). In sum, this is high quality science, with multiple results that are individually interesting, but in my assessment it does not yield insights that I would expect to be compelling to a broad readership (beyond being about an interesting critter).

The single genes had clear phenotypes and there was some discussion about the ways in which these phenotypes were similar to and different from the phenotypes resulting from functional knockdown of homologous genes in other arthropods. However, the developmental work would be more compelling if it were more completely fleshed out. For Hox genes, this would mean generating RNAi phenotypes for all of the genes (and combinations of coexpressed genes), or at least the set that are expressed in legs. For example, is the absence of leg-pedipalp transformation in L3 and L4, which also express Dfd, due to Scr (expressed in L3 and L4, but not L1 or L2) also specifying leg identity? (Knockdowns of Scr on its own and the double Dfd/Scr knockdown should be done.) Also, as the authors acknowledge (l. 350-352), the homeotic effects can't be distinguished from a role regulating growth in L2, and thus the comparisons to Ubx in waterstriders are premature. For EGFR, the authors suggest that both early functions (distal tip) and late functions (related to expression in each leg segment) are conserved, based on their phenotype analyses. However, the expression data do not suggest the existence of a distal signaling center. Thus, expression/function of additional signaling pathway components are needed to make sense of this result.

Questions about results/interpretation:

Spider/scorpion Ks frequency distributions look very similar to the distribution in Phalangium. From this, it's not how these plots support/add evidence to the conclusion that Phalangium has not undergone a whole genome duplication while those other taxa have. It would also be helpful to give the number of gene families analyzed for each taxon, as large difference in this would perhaps be meaningful. (Also, note that order of species in legend differs from order in figures.)

Are there features of embryogenesis that account for the strong asymmetries in RNAi phenotypes encountered?

Many of the support values shown in the EGFR tree are extremely low. In particular, the branches on which the conclusion of three independent duplication events within arthropods are based do not receive support (with the exception that the branch grouping the two Phalangium sequences appears robust). Thus, the scenario of there independent duplication events should not be presented as as well established as the text suggests, especially if there's a chance that subsequent gene conversion in any part(s) of the protein-coding region could have increased

similarity between paralogues after duplication. (If there is other information that supports it, for example, gene structure features, or strong support in the UTRs than in the coding region, that would allay these concerns.)

Table S5: I'm puzzled that 15 BUSCO genes seem to have gone missing as a result of the purge haplotig command. Does this makes sense? If so, what happened to them? It would seem highly likely that they were real.

Data availability:
The EGFR alignment and tree should be made available in text/machine readable form.

Additional notes:
l. 142-144: Given the size of the dataset, there's no need for the computational shortcuts such as fast bootstrapping; a full analysis can be performed. Also, What model of amino acid substitution was used? (This may be implicitly stated through the commands, but please also state explicitly.)

l. 204-206: It looks like half the sentence went missing.

l. 270-271 and figure S8: truncation of EGFR-B. It would be helpful for readers to know you have confirmed that the truncation is genuine and not an artifact of incomplete genome or mRNA assembly.

Several supplemental figures appeared lower quality than necessary to be readable at the size presented. (This could be due to pdf processing through the manuscript handling system, but if not, then originals should be magnified/improved.) Affected figures included S3, S4, S7, S8

Figure S5: It took me a while to understand how some of the Dfd phenotypes could become wild type. Also, I recommend stacking the conditions in order of severity (so strong transformation at the top). And I wonder whether the figure would be more informative if it showed the bilateral combination frequencies (WT/weak, WT/Strong; weak/weak; weak/strong; strong/strong). Especially if the two sides are not independent, this seems like a more representative summary of the data.

Author's Response to Decision Letter for (RSPB-2021-0372.R0)

See Appendix A.

RSPB-2021-1168.R0

Review form: Reviewer 2

Recommendation
Accept with minor revision (please list in comments)

Scientific importance: Is the manuscript an original and important contribution to its field?
Good

General interest: Is the paper of sufficient general interest?
Excellent

Quality of the paper: Is the overall quality of the paper suitable?
Excellent

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
No

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

> **Is it accessible?**
> No
>
> **Is it clear?**
> N/A
>
> **Is it adequate?**
> Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
This remains a clear, thorough manuscript about an interesting organism. I believe that the extensive revisions have substantially clarified the main focus of the paper and I think it would be appropriate for the journal following minor revisions to clarify a few points.

Substantive questions about interpretation that should be clarified/fixed:

l. 203-205: As stated, this claim is incorrect because all 4 miRNA that appear duplicated in the harvestman also appear to be duplicated in Arachnopulmonata. Some of what precedes it is also vague (because of the 'most', when in fact it appears to be true of most). And key evidence is omitted (re. duplication outside of chelicerates).
Is the following rephrasing attempt accurate, starting with claims on l. 200? "Three of these four microRNAs appear [or are parsimoniously interpreted as having undergone' to have undergone an initial duplication in an arthropod ancestor and subsequent duplication in the Arachnopulmonata and horseshoe crabs. The other, miR-29, has two copies in harvestmen, horseshoe crabs and a subset of arachnopulmonates, which suggests multiple independent duplications rather than an ancestral duplication."
[I see that is getting complicated—perhaps the possible miR-29 history doesn't need to be explained, just noted that history does not suggest a shared ancestral duplication event.]

l. 346-349: I'm not fully following the argument for subfunctionalization. I think that's because the evidence isn't totally consistent with the conclusion as the spider data would have been predicted to show function of 1 copy in L1 and the other copy in L2 (as opposed to no evidence of Dfd function in L2).

l. 352-354: The way this conclusion is stated, it comes across as a novel conclusion made possible by the evidence in this study. But in fact expression data from tardigrades and onychophora, paleontological evidence supporting multiple origins of tagmata, and evidence that ancestral arthropods lacked tagmata, but would have inherited offset anterior Hox boundaries (which are found across bilateria). Even if the boundaries were different in the harvestman, the conclusion would still be supported. So the presentation should be altered to reflect the fact that this paper is adding minor confirmatory evidence rather than supporting a new interpretation.

l. 357-358: Clarify that the argument for functional redundancy only applies to L3 (whereas from context, one might expect the claim to apply to L1-L3)

Other recommendations:

l. 164 (and supplemental methods l. 162-165): In the ms. text, it is important to indicate that this is the number of predicted genes after considerable cleaning of the dataset. In the supplemental methods, please also indicate either the total numbers of genes initially output from BRAKER or the number removed at each cleaning step.

l. 254: clarify number; my first reaction was that 8/177 was surprisingly low, but from supplement, I see that's a result of mortality. Perhaps present as "Mortality was high (X of 177), but 8 of X surviving individuals showed partial p-to-l transformations"

l. 406-411: I found the concluding paragraph to be a letdown. I think a stronger version would point out that so far the genome has confirmed/extended evidence of conservation of patterning and then, if there are any clues, elaborate on how it may now be used to refine our understanding of new traits.

Fig. 2: consider deleting panel B (where new data is redundant with A, and a expected in the context of arthropod Hox clusters) and condensing A to the same width as C;

Fig. 3: I recommend replacing schematic in A with in situ images of the Hox gene expression, given that rings of expression occur along the legs in both genes, so the variation in expression levels shown diagrammatically is too oversimplified.
Also, bold font in C is hard to see; additional emphasis from colored font, would help these stand out.

Fig. 3 legend: please explain use of color in R; I also found "sum of halves" not entirely clear.
Supp EGFR tree: why were the 5′ and 3′ segments of EGFR from Conichochernes crassus analyzed separately? Their different placements on the tree call into question the robustness of these phylogenetic results.

Typos, etc.:

l. 78: change 'derived' to a term that does not reinforce the misconception that some extant groups are ancestral and others are derived.

l. 123: add space: "16 cells"

l. 166: add 'and' before 'the mite'

l. 181: rephrase 'non-collinearity': it sounds like you are suggesting that colinearity may be maintained (and that the apparent fragmentation indicates an incomplete assembly).

l. 219: delete 'similarly' or make it explicit what aspect of expression is similar between Dfd and Scr.

l. 329-330: perhaps rephrase as “Nevertheless, the assumed unduplicated condition.....has not been rigorously tested previously.”

l. 337 and l. 352: replace ‘notion’ with ‘inference’ or ‘conclusion’

l. 343: unclear whether you mean leg segments or body segments. Only evidence shown is for legs; what about changing ‘segments’ to ‘legs’?

l. 361: rephrase (abdominal segments acquire thoracic identity or transformation of abdominal to thoracic identity)

l. 365: Perhaps ‘informative’ would be a better descriptor than ‘powerful’?

l. 393: change ‘median’ to ‘medial’ or ‘intermediate’

l. 394-399: replace discussion of short germband taxa with an evolutionarily meaningful group.

Decision letter (RSPB-2021-1168.R0)

14-Jun-2021

Dear Mr Gainett:

Your manuscript has now been peer reviewed and the reviews have been assessed by an Associate Editor. The reviewers’ comments (not including confidential comments to the Editor) and the comments from the Associate Editor are included at the end of this email for your reference. As you will see, the reviewers and the Editors have raised some concerns with your manuscript and we would like to invite you to revise your manuscript to address them.

We do not allow multiple rounds of revision so we urge you to make every effort to fully address all of the comments at this stage. If deemed necessary by the Associate Editor, your manuscript will be sent back to one or more of the original reviewers for assessment. If the original reviewers are not available we may invite new reviewers. Please note that we cannot guarantee eventual acceptance of your manuscript at this stage.

To submit your revision please log into http://mc.manuscriptcentral.com/prsb and enter your Author Centre, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions”, click on "Create a Revision”. Your manuscript number has been appended to denote a revision.

When submitting your revision please upload a file under "Response to Referees" in the "File Upload" section. This should document, point by point, how you have responded to the reviewers’ and Editors’ comments, and the adjustments you have made to the manuscript. We require a copy of the manuscript with revisions made since the previous version marked as ‘tracked changes’ to be included in the ‘response to referees’ document.

Your main manuscript should be submitted as a text file (doc, txt, rtf or tex), not a PDF. Your figures should be submitted as separate files and not included within the main manuscript file.

When revising your manuscript you should also ensure that it adheres to our editorial policies (https://royalsociety.org/journals/ethics-policies/). You should pay particular attention to the following:

Research ethics:
If your study contains research on humans please ensure that you detail in the methods section whether you obtained ethical approval from your local research ethics committee and gained informed consent to participate from each of the participants.

Use of animals and field studies:
If your study uses animals please include details in the methods section of any approval and licences given to carry out the study and include full details of how animal welfare standards were ensured. Field studies should be conducted in accordance with local legislation; please include details of the appropriate permission and licences that you obtained to carry out the field work.

Data accessibility and data citation:
It is a condition of publication that you make available the data and research materials supporting the results in the article (https://royalsociety.org/journals/authors/author-guidelines/#data). Datasets should be deposited in an appropriate publicly available repository and details of the associated accession number, link or DOI to the datasets must be included in the Data Accessibility section of the article (https://royalsociety.org/journals/ethics-policies/data-sharing-mining/). Reference(s) to datasets should also be included in the reference list of the article with DOIs (where available).

In order to ensure effective and robust dissemination and appropriate credit to authors the dataset(s) used should also be fully cited and listed in the references.

If you wish to submit your data to Dryad (http://datadryad.org/) and have not already done so you can submit your data via this link
http://datadryad.org/submit?journalID=RSPB&manu=(Document not available), which will take you to your unique entry in the Dryad repository.

If you have already submitted your data to dryad you can make any necessary revisions to your dataset by following the above link.

For more information please see our open data policy http://royalsocietypublishing.org/data-sharing.

Electronic supplementary material:
All supplementary materials accompanying an accepted article will be treated as in their final form. They will be published alongside the paper on the journal website and posted on the online figshare repository. Files on figshare will be made available approximately one week before the accompanying article so that the supplementary material can be attributed a unique DOI. Please try to submit all supplementary material as a single file.

Online supplementary material will also carry the title and description provided during submission, so please ensure these are accurate and informative. Note that the Royal Society will not edit or typeset supplementary material and it will be hosted as provided. Please ensure that the supplementary material includes the paper details (authors, title, journal name, article DOI). Your article DOI will be 10.1098/rspb.[paper ID in form xxxx.xxxx e.g. 10.1098/rspb.2016.0049].

Please submit a copy of your revised paper within three weeks. If we do not hear from you within this time your manuscript will be rejected. If you are unable to meet this deadline please let us know as soon as possible, as we may be able to grant a short extension.

Thank you for submitting your manuscript to Proceedings B; we look forward to receiving your revision. If you have any questions at all, please do not hesitate to get in touch.

Best wishes,

Professor Gary Carvalho
mailto: proceedingsb@royalsociety.org

Associate Editor
Comments to Author:
I want to congratulate the authors for their efforts to improve the paper with more data and experiments, as well as by editing the manuscript itself.

I think the manuscript could be improved further by clarifying the situation with the miRNAs duplicated in both opiliones and arachnopulmonate, are these duplications considered not-homologous (they are in-paralogs of each lineage rather than out-paralogs?) due to evidence from gene trees, synteny, or something else?

I'd also like to encourage the authors to make the phylogeny in Figure 2C consistent with the other ones in the paper (i.e., collapsed), unless this is the result of clustering based on the number of miRNA copies, but I don't think that is the case .

Other than that, I'd like to kindly ask the authors to address the concerns of the referee.

Reviewer(s)' Comments to Author:
Referee: 2
Comments to the Author(s).
This remains a clear, thorough manuscript about an interesting organism. I believe that the extensive revisions have substantially clarified the main focus of the paper and I think it would be appropriate for the journal following minor revisions to clarify a few points.

Substantive questions about interpretation that should be clarified/fixed:

l. 203-205: As stated, this claim is incorrect because all 4 miRNA that appear duplicated in the harvestman also appear to be duplicated in Arachnopulmonata. Some of what precedes it is also vague (because of the 'most', when in fact it appears to be true of most). And key evidence is omitted (re. duplication outside of chelicerates).
Is the following rephrasing attempt accurate, starting with claims on l. 200? "Three of these four microRNAs appear [or are parsimoniously interpreted as having undergone' to have undergone an initial duplication in an arthropod ancestor and subsequent duplication in the Arachnopulmonata and horseshoe crabs. The other, miR-29, has two copies in harvestmen, horseshoe crabs and a subset of arachnopulmonates, which suggests multiple independent duplications rather than an ancestral duplication."
[I see that is getting complicated—perhaps the possible miR-29 history doesn't need to be explained, just noted that history does not suggest a shared ancestral duplication event.]

l. 346-349: I'm not fully following the argument for subfunctionalization. I think that's because the evidence isn't totally consistent with the conclusion as the spider data would have been predicted to show function of 1 copy in L1 and the other copy in L2 (as opposed to no evidence of Dfd function in L2).

l. 352-354: The way this conclusion is stated, it comes across as a novel conclusion made possible by the evidence in this study. But in fact expression data from tardigrades and onychophora, paleontological evidence supporting multiple origins of tagmata, and evidence that ancestral arthropods lacked tagmata, but would have inherited offset anterior Hox boundaries (which are found across bilateria). Even if the boundaries were different in the harvestman, the conclusion would still be supported. So the presentation should be altered to reflect the fact that this paper is adding minor confirmatory evidence rather than supporting a new interpretation.

l. 357-358: Clarify that the argument for functional redundancy only applies to L3 (whereas from context, one might expect the claim to apply to L1-L3)

Other recommendations:

l. 164 (and supplemental methods l. 162-165): In the ms. text, it is important to indicate that this is the number of predicted genes after considerable cleaning of the dataset. In the supplemental methods, please also indicate either the total numbers of genes initially output from BRAKER or the number removed at each cleaning step.

l. 254: clarify number; my first reaction was that 8/177 was surprisingly low, but from supplement, I see that's a result of mortality. Perhaps present as "Mortality was high (X of 177), but 8 of X surviving individuals showed partial p-to-l transformations"

l. 406-411: I found the concluding paragraph to be a letdown. I think a stronger version would point out that so far the genome has confirmed/extended evidence of conservation of patterning and then, if there are any clues, elaborate on how it may now be used to refine our understanding of new traits.

Fig. 2: consider deleting panel B (where new data is redundant with A, and a expected in the context of arthropod Hox clusters) and condensing A to the same width as C;

Fig. 3: I recommend replacing schematic in A with in situ images of the Hox gene expression, given that rings of expression occur along the legs in both genes, so the variation in expression levels shown diagrammatically is too oversimplified.
Also, bold font in C is hard to see; additional emphasis from colored font, would help these stand out.

Fig. 3 legend: please explain use of color in R; I also found "sum of halves" not entirely clear.
Supp EGFR tree: why were the 5′ and 3′ segments of EGFR from Conichochernes crassus analyzed separately? Their different placements on the tree call into question the robustness of these phylogenetic results.

Typos, etc.:

l. 78: change 'derived' to a term that does not reinforce the misconception that some extant groups are ancestral and others are derived.

l. 123: add space: "16 cells"

l. 166: add 'and' before 'the mite'

l. 181: rephrase 'non-collinearity': it sounds like you are suggesting that colinearity may be maintained (and that the apparent fragmentation indicates an incomplete assembly).

l. 219: delete 'similarly' or make it explicit what aspect of expression is similar between Dfd and Scr.

l. 329-330: perhaps rephrase as "Nevertheless, the assumed unduplicated condition.....has not been rigorously tested previously."

l. 337 and l. 352: replace 'notion' with 'inference' or 'conclusion'

l. 343: unclear whether you mean leg segments or body segments. Only evidence shown is for legs; what about changing 'segments' to 'legs'?

l. 361: rephrase (abdominal segments acquire thoracic identity or transformation of abdominal to thoracic identity)

l. 365: Perhaps 'informative' would be a better descriptor than 'powerful'?

l. 393: change 'median' to 'medial' or 'intermediate'

l. 394-399: replace discussion of short germband taxa with an evolutionarily meaningful group.

Author's Response to Decision Letter for (RSPB-2021-1168.R0)

See Appendix B.

Decision letter (RSPB-2021-1168.R1)

14-Jul-2021

Dear Mr Gainett

I am pleased to inform you that your manuscript entitled "The genome of a daddy-long-legs (Opiliones) illuminates the evolution of arachnid appendages" has been accepted for publication in Proceedings B.

You can expect to receive a proof of your article from our Production office in due course, please check your spam filter if you do not receive it. PLEASE NOTE: you will be given the exact page length of your paper which may be different from the estimation from Editorial and you may be asked to reduce your paper if it goes over the 10 page limit.

If you are likely to be away from e-mail contact please let us know. Due to rapid publication and an extremely tight schedule, if comments are not received, we may publish the paper as it stands.

If you have any queries regarding the production of your final article or the publication date please contact procb_proofs@royalsociety.org

Data Accessibility section
Please remember to make any data sets live prior to publication, and update any links as needed when you receive a proof to check. It is good practice to also add data sets to your reference list.

Open Access
You are invited to opt for Open Access, making your freely available to all as soon as it is ready for publication under a CCBY licence. Our article processing charge for Open Access is £1700. Corresponding authors from member institutions (http://royalsocietypublishing.org/site/librarians/allmembers.xhtml) receive a 25% discount to these charges. For more information please visit http://royalsocietypublishing.org/open-access.

Your article has been estimated as being 9 pages long. Our Production Office will be able to confirm the exact length at proof stage.

Paper charges
An e-mail request for payment of any related charges will be sent out after proof stage (within approximately 2-6 weeks). The preferred payment method is by credit card; however, other payment options are available

Electronic supplementary material:
All supplementary materials accompanying an accepted article will be treated as in their final form. They will be published alongside the paper on the journal website and posted on the online figshare repository. Files on figshare will be made available approximately one week before the accompanying article so that the supplementary material can be attributed a unique DOI.

Thank you for your fine contribution. On behalf of the Editors of the Proceedings B, we look forward to your continued contributions to the Journal.

Sincerely,
Professor Gary Carvalho
Editor, Proceedings B
mailto: proceedingsb@royalsociety.org

Associate Editor:
Board Member
Comments to Author:
(There are no comments.)

Appendix A

Associate Editor
Board Member: 1
Comments to Author:
I enjoyed reading the paper by Gainett and collaborators, in which they present the genome sequence of an opinion and use it combined with other omics approaches and gene expression data to infer the genetic basis of the evolution of appendages in chelicerates. The manuscript is well-written, nicely presented, and the analyses are sound. I'd like to congratulate the authors on their work.

I agree with most of the concerns raised by the two expert reviewers, which for the most part are clarifications and requests for more information. I'd like to encourage the authors to address these concerns in order to improve the quality of their manuscript, including the tone of the claims about the homeotic patterns of the manuscript, claims made relating to expression (early and late effects), and comparison with water striders.

Response: We thank the Associate Editor and the reviewers for this positive feedback. Following the reviewer's suggestions, in this revision, we have (1) clarified several aspects of the genome annotation, (2) performed a double RNAi experiment against *Dfd* and *Scr* (the first such experiment in *P. opilio*), and (3) added analyses of *pointed*, a member of the EGFR signaling pathway. We hope that this additions and revisions will satisfy the reviewers, and we very much thank them for their constructive suggestions, which have encouraged us to push the boundaries of what we have done in this system.

Reviewer(s)' Comments to Author:

Referee: 1

Comments to the Author(s)
Gainett et al present the genome of Phalangium opilio alongside data on the genetic basis for the elongate legs seen in this species. This work is well-written and presented. In general, the authors are to be commended on the clarity and openness of the presentation of their methods and data.

There is one major issue with the manuscript as written that I would request is corrected before publication. I also have a number of minor comments and further suggestions, as noted below. Please also note that the UCSC genome browser link provided did not allow me access to the browser, and a more user-friendly URL should be generated and provided before publication - something like a tinyurl could work well.

Response:
The link to the genome browser will be publicly hosted, however the link provided here was temporary for the reviewers. Upon acceptance of the publication, the final public link will be updated with a shortened URL. This temporary link has also been shortened for the reviewers here: https://bit.ly/3tYgzhU .

Major comment:
The removal of genes from the annotation described on lines 146-148 of the supplement is not clearly described and could have a major impact on the analysis and interpretation of the results.
1) There needs to be clarity as to what "did not generate significant BLAST hits, functional annotation, or lacked transcript evidence" means. This could result in the exclusion of many excellent gene models. Please provide full details of criteria used for exclusion, and a summary of how many genes were thus removed.

Response:

Additional information on refining the gene models from the automated annotation were added to the supplementary materials as well as specific commands for Blast and Blast2GO for clarity (Lines 160-192).

2) "the gene set was then further refined with a 98% similarity threshold using CDHIT" - again, this could result in the trimming of many excellent gene models. This would particularly impact the Hox and Ks analyses - if close paralogs exist, they would be deleted by this step. Please
- note how many genes were thus removed
- provide clear evidence that no close matches to hox cluster genes were removed

Response:

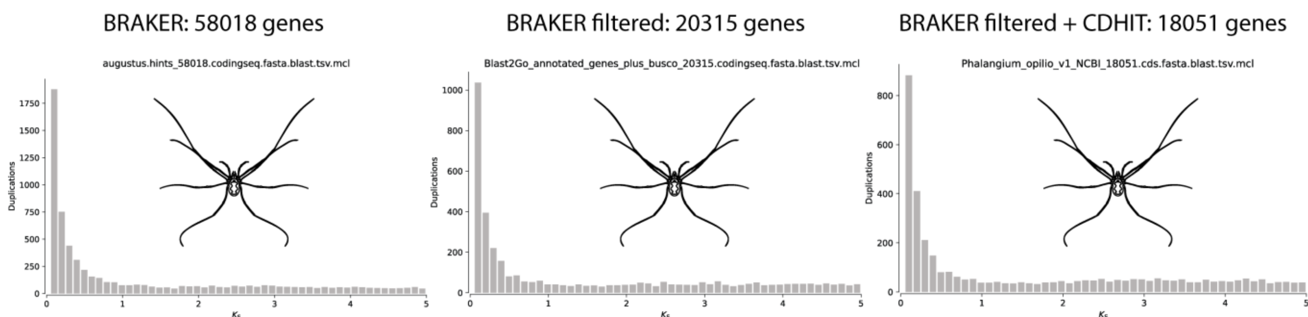
We take the reviewer's point that subset of the putative duplicates may potentially be real gene models. This was why we did not remove the associated scaffolds from the assembly. However, we did observe enough questionable duplicated genes in the gene set (e.g. 100% identical peptide sequences for well-characterized transcription factors like *Distal-less*; jg53.t1 and jg53015.t1) that we choose to establish a conservative gene model for this first draft genome.

In the case of Hox genes, three out of ten had models predicted in two separate scaffolds in the unfiltered BRAKER2 gene set (58,018 genes): *lab* (Contig7109 and Contig9232), *Hox3* (Contig4435 and Contig8730), and *Abd-B* (Contig8598 and Contig7106). The duplicates for *lab* and *Abd-A* were filtered after the pipeline resulting in the 20,315 gene set (see response above). The duplicated gene model for *Hox3* was filtered in the final CD-HIT 98% step (18,051 genes). Both predicted *Hox3* proteins have 477 amino acids, with a 98.32% similarity.
We infer that the paired scaffolds were duplicated artificially due to high allelic variation as evidenced by nearly identical genomic sequence present on separate scaffolds.

Thus, the 98% CD-HIT cutoff may be too stringent for some gene models but we anticipate further improvements to our gene models will require resequencing, including with new technologies. These efforts are underway, but do not impact the conclusions of this present work.

- repeat the Ks analysis for Phalangium opilio on the untrimmed set if large numbers (perhaps >500?) genes were removed from the data shown, to ensure that paralogs (which could result from WGD) were not removed inadvertently by this step

Response: In order to ensure that the lack of WGD signal in the Ks plots is not an artifact of filtering good gene models, we have conducted the Ks analysis in all three sets of gene models: (1) the unfiltered BRAKER output (58018 genes); (2) the BRAKER output filtered by the presence of any functional annotation (20315 genes); (3) the filtered BRAKER output further refined with a CDHIT similarity threshold of 98% (18051 genes). All three sets have almost identical Ks distributions, providing no evidence for a recent WGD, as shown is the following figure:



Action taken: Following a comment made by reviewer 2, we have decided to remove this analysis from the resubmission.

3) please provide details as to how genes noted as duplication were removed from the gtf file. Was the shortest copy removed in all cases, for example? The CDHIT command given shows how the amino acid file was trimmed, but not how a consensus gtf (and thus the final set of gene models) was generated.

Response:

CDHIT by default retains the longest representative of each identified cluster (at our set threshold of 98%), these identified redundant peptides were removed from the GTF file using their gene name identifier. A statement on how the genes were removed from the GTF has been added to the supplemental material for clarity (line 170-171).

Minor comments:

Blobplot: please give blast/diamond settings used for the blast step in this identification. It is very surprising that "no hit" is by far the most common result, and no arthropod "hit" was found within the majority of sequences. This could be explained by overly stringent cutoffs, a poor target library for blasting, or alternatively, if no gene model is present on the vast majority of these contigs. Please provide further information, as your gene annotation process suggests that most gene models should have a good blast hit to something. It is perhaps worth considering redoing the blobplot with more relaxed blast criteria.

Response:

Blast and Blobtools commands have been added to the supplementary materials (line 210-233). The Blast was run on the purged assembly and was done using the recommended stringency thresholds by the developers of Blobtools (-max_target_seqs 10 -max_hsps 10 -evalue 1e-25). Additionally, the spread of the sequences (based on coverage and GC content) seem to suggest that all sequences are likely of *Phalangium* origin and do not represent contaminations.

-Table S5 or in text - please provide information as to what "Ns" in the genome represent. If these are consistently of a given size (e.g. 1000bp when inserted during scaffolding) please make this clear.

Response:

During the final scaffolding phase of SSPACE-LongRead, the program calculates the gap-size between each contig. From this calculation, SSPACE-LongRead will merge contigs if this value is negative and there is an overlap identified; if the value is positive, a gap is inserted between the contigs depending on the gap size using N's. Therefore, N's here are estimated gap sizes and represent singular Ns. This has been added in the supplementary text (line 87-89).

-Please provide a supplementary table summarising the results of your RepeatMasker/RepeatModeler analysis. This will provide basic information on repetitive elements for the community, as well as making it clear what %age of the genome has been soft masked.

Response:

This table has been added to the supplementary materials as a table (Table S7).

Further suggestions:
- Fig 2C placement of node labels could be misconstrued, as the placement of these often is suggestive of wider groups than intended, as they are often slightly more towards the root than the clades they note. I suggest these are moved closer towards the node to which they refer, even if the figure needs to be increased in size.

Response: To improve clarity, we have replaced the name labels for numbers and added a description in the figure legend.

- Line 265: maritma misspelled

Response: The sentence was removed with the Ks plot analysis.

- Supplementary: possible redundancy, consider rewriting, line 54: The Single Molecule Real-Time (SMRT) Cells were sequenced on (16) SMRT cells

Response: We removed the second instance of "SMRT" as suggested.

Referee: 2

Comments to the Author(s)
This paper brings an emerging model system in evo-devo, the daddy-long-legs Phalangium opilio into the genomic era. From a phylogenetic perspective, this is significant for two reasons. The first is developmental data from chelicerates are essential for robust inferences about the ancestral state in arthropods as a whole. Second, developmental comparisons to other arthropods are simplified (both methodologically and conceptually) because the lineage is shown not to have undergone any whole genome duplications after diverging from the lineage that gave rise to the other major arthropod groups (whereas spiders and scorpions have). After presenting the genome, the paper presents RNA-based functional analyses of two conserved genes, the Hox gene Deformed and EGFR. The results are largely as one would predict, with some potentially interesting implications for the evolution of arthropod appendages.

The paper is very clearly written, the analyses appear well done, and each piece of the paper appears sound (with a few minor quibbles—see below). However, the genome and the two RNAi phenotypes did not come across as complementary—neither part strengthened the other part. Nor did there appear to be a broader conclusion that was enabled by combining these sets of data. As a consequence of doing so much, everything is short-changed (or in some cases not followed through in as great a depth as I would expect). In sum, this is high quality science, with multiple results that are individually interesting, but in my assessment it does not yield insights that I would expect to be compelling to a broad readership (beyond being about an interesting critter).

Response: One of the fundamental advantages of *P. opilio* as an arachnid model system is that it has not undergone WGD (as noted by this reviewer above). Every work that has come before on comparative arachnid genomics has accepted this unduplicated condition of the harvestman genome.

But there is one major problem with this claim—this inference has always been predicated on developmental transcriptome data, which may not always be effective assessments of WGD. We invite the reviewer to consider the case of pseudoscorpions. Just three years ago, we thought that pseudoscorpions were similarly free of WGD, on the basis of homeobox gene surveys in two transcriptomic libraries [1]. The publication of the first pseudoscorpion draft genome with intensive sequencing revealed that this order was in fact also part of the arachnopulmonate clade (common ancestor of spiders and scorpions) that is united by a WGD event [2].

<u>This manuscript is the first time we have been able to infer with confidence that certain genes (e.g., Hox) are indeed single copy in the harvestman.</u> If our interpretations are grounded in this inference, then we must include the genome as the prerequisite of our RNAi experiments, in order to substantiate this argument. We have rewritten several parts of the manuscript to make our logical bridge clear.

The single genes had clear phenotypes and there was some discussion about the ways in which these phenotypes were similar to and different from the phenotypes resulting from functional knockdown of homologous genes in other arthropods. However, the developmental work would be more compelling if it were more completely fleshed out. For Hox genes, this would mean generating RNAi phenotypes for all of the genes (and combinations of coexpressed genes), or at least the set that are expressed in legs. For example, is the absence of leg-pedipalp transformation in L3 and L4, which also express Dfd, due to Scr (expressed in L3 and L4, but not L1 or L2) also specifying leg identity? (Knockdowns of Scr on its own and the double Dfd/Scr knockdown should be done.)

Response: We agree with the reviewer that the study of *Sex combs reduced* could potentially reveal further aspects of leg specification. Throughout 2019 and 2020, we had conducted multiple RNAi experiments targeting *Popi-Scr*. We injected a total of five clutches, totaling 540 embryos. We did not detect phenotypic differences between hatchlings of treatment and control.

Following this review, we additionally investigated if the specification of L3 and L4 requires the combined input of *Dfd* and *Scr*, by conducting the first double-knockdown experiments in the harvestman, targeting both genes with RNAi. In this revised version of the manuscript, we show that double knockdown against *Dfd* and *Scr* results in additional leg-to-pedipalp transformation in the L3 segment.

Also, as the authors acknowledge (l. 350-352), the homeotic effects can't be distinguished from a role regulating growth in L2, and thus the comparisons to Ubx in waterstriders are premature.

Response: We take the reviewer's point that we cannot presently disentangle homeosis from a non-canonical role in growth. We have emended the manuscript extensively to revise this section and focus instead on the insights of *Dfd* and *Scr* activity in arachnid leg fate specification.

For EGFR, the authors suggest that both early functions (distal tip) and late functions (related to expression in each leg segment) are conserved, based on their phenotype analyses. However, the expression data do not suggest the existence of a distal signaling center. Thus, expression/function of additional signaling pathway components are needed to make sense of this result.

Response: Following the reviewer's suggestion, we investigated the expression pattern of an additional component of the EGFR signaling pathway: the gene *pointed* (*pnt*), which is an ETS transcription factor with an effector role in EGFR signaling [3]. We discovered that the *P. opilio* homolog, *Po-pnt*, has an early expression pattern at the tip of the appendages, and a later expression on the distal portion of the appendages, forming rings. This expression pattern is reminiscent of the expression of EGFR signaling components in two phases of leg patterning in *D. melanogaster* [4], and also the expression pattern in the embryonic legs of the beetle *T. castaneum* [5]. The similarity of *pnt* expression patterns between a daddy-long-legs and these insects strengthens our inference that EGFR signaling is involved in distal leg patterning in arachnids and was coopted for tarsomere patterning in *P. opilio*. Importantly, the early distal

expression of *Po-pnt* at the tip of the developing limb buds suggests the existence of a distal EGFR signaling center.

Questions about results/interpretation:

Spider/scorpion Ks frequency distributions look very similar to the distribution in Phalangium. From this, it's not how these plots support/add evidence to the conclusion that Phalangium has not undergone a whole genome duplication while those other taxa have. It would also be helpful to give the number of gene families analyzed for each taxon, as large difference in this would perhaps be meaningful. (Also, note that order of species in legend differs from order in figures.)

Response: We thank the reviewer for the suggestion. In accordance with a recent study [6], we corroborated a peak of Ks values in two horseshoe crab genomes, which suggests that at least one of the WGD duplications in Xiphosura is relatively recent (i.e., Cretaceous or younger). Ks plots did not show peaks in the arachnopulmonates, in accordance with the ancient timing inferred for this event (>430 Mya; the oldest crown-group scorpion fossils are Silurian in age).

Ks plots are limited to detecting only recent genome duplications (and have been critiqued for their high sensitivity in older splits), so our intent here was only to determine if there was evidence for a recent whole genome duplication specific to Opiliones. While we take the reviewer's point, this is a minor component of the work and we have therefore excluded these analyses from the paper.

Are there features of embryogenesis that account for the strong asymmetries in RNAi phenotypes encountered?

Response: To our knowledge, there are no specific attributes of *P. opilio* that would account for the strong asymmetries. We believe that the asymmetries are due to the time in ontogeny when the embryos are more tractable to be injected, which is immediately before the germ band forms, or at the germband stage (~5 dAEL). Considering that the cells at this stage are not syncytial, as presumably is the case during the first cell divisions at the center of the egg, the assimilation of the dsRNA may be limited to the cells most close to the site of injection.

Many of the support values shown in the EGFR tree are extremely low. In particular, the branches on which the conclusion of three independent duplication events within arthropods are based do not receive support (with the exception that the branch grouping the two Phalangium sequences appears robust). Thus, the scenario of there independent duplication events should not be presented as as well established as the text suggests, especially if there's a chance that subsequent gene conversion in any part(s) of the protein-coding region could have increased similarity between paralogues after duplication. (If there is other information that supports it, for example, gene structure features, or strong support in the UTRs than in the coding region, that would allay these concerns.)

Response: We agree with the reviewer's point. Since the main purpose of the phylogenetic analysis was to confirm the identity of the *Egfr* candidates in *P. opilio*, we have removed from the text the sentences about the conclusion of the independent duplication events.

Table S5: I'm puzzled that 15 BUSCO genes seem to have gone missing as a result of the purge haplotig command. Does this makes sense? If so, what happened to them? It would seem highly likely that they were real.

Response:

A known trade off when purging haplotypic duplications is the decrease in overall BUSCO scores, though algorithmic developments have been made to have minimal impacts on these metrics (Guan et al. 2020). In decreasing haplotypic duplications, as evidenced by the removal of duplications from known single copy genes in our assembly from 52.4% (531 genes) to 7.1% (72 genes), there was a tradeoff of 7 genes that are now fragmented, and 8 genes are missing for a total of 15 BUSCO gene difference. The 8 genes that are missing likely sit on duplicated regions that are now identified as alternative contigs are no longer present in our higher quality curated, haplotype-purged assembly.

Data availability:
The EGFR alignment and tree should be made available in text/machine readable form.

Response: The EGFR alignment and tree are now deposited in the Dryad repository, which is described in the Data Availability Statement.

Additional notes:
l. 142-144: Given the size of the dataset, there's no need for the computational shortcuts such as fast bootstrapping; a full analysis can be performed. Also, What model of amino acid substitution was used? (This may be implicitly stated through the commands, but please also state explicitly.)

Response: We reran 1000 non-parametric standard bootstrap analyses and updated the figure. The model selected by ModelFinder was JTT+G4, which is now explicitly mentioned in the supplementary material.

l. 204-206: It looks like half the sentence went missing.

Response: We amended the sentence with the missing part.

l. 270-271 and figure S8: truncation of EGFR-B. It would be helpful for readers to know you have confirmed that the truncation is genuine and not an artifact of incomplete genome or mRNA assembly.

Response: We have clarified this point by including the following sentence on the main text: "A 3' UTR for *Po-EgfrB* was assembled in both embryonic transcriptomes and corroborated by the

genome assembly, disfavoring fragmentary assembly as a possible explanation for missing domains".

We additionally designed primers and conducted in situ hybridization using the same protocols and stages probed for *Po-EgfrA*. However, we did not detect any signal in embryos assayed for *Po-EgfrB* anti-sense probes. Even though we cannot rule out that *Po-EgfrB* is expressed below the levels detectable by in situ hybridization in these stages, we at present conclude that that the truncation evidenced by the genome and transcriptomes is correct and that this gene has diverged in expression and function.

Primers used (appended with T7 ends):
>Popi_Egfr_B_911bp_F
ggccgcggATGCTCAAAGTGCGACGATC
>Popi_Egfr_B_911bp_R
cccggggcGACCTTGAACCTGTTGCTCG

Several supplemental figures appeared lower quality than necessary to be readable at the size presented. (This could be due to pdf processing through the manuscript handling system, but if not, then originals should be magnified/improved.) Affected figures included S3, S4, S7, S8

Response: We have now replaced the mentioned figures with higher resolution files. We believe this was just an issue with the manuscript handling system, which automatically reduces file sizes to 72 dpi.

Figure S5: It took me a while to understand how some of the Dfd phenotypes could become wild type. Also, I recommend stacking the conditions in order of severity (so strong transformation at the top). And I wonder whether the figure would be more informative if it showed the bilateral combination frequencies (WT/weak, WT/Strong; weak/weak; weak/strong; strong/strong). Especially if the two sides are not independent, this seems like a more representative summary of the data.

Response: To address this issue, we have recoded all the *Dfd* experiment. Instead of creating classes by making assumptions about the penetrance, we focused on the presence/absence of homeosis in each leg as a meaningful and assumption-free description of the *Po-Dfd* RNAi, that would also be applicable for the results obtained in the double knockdown of *Po-Dfd+Scr*. To describe the presence of mosaicism, we followed the reviewer's suggestions when coding individual mosaics in categories that combined all the possible coding schemes. For example, for the *Po-Dfd* KD, the four observed conditions were wild type (WT; no homeosis), homeosis in L1 and L2 (both); homeosis in L1 only (L1); and homeosis in L2 only (L2). Therefore, mosaics were classified into nine possible combinations: WT/both; WT/L1; WT/L2; both/both; both/L1; both/L2; L1/L1; L1/L2; L2/L2. This is now depicted with schematics in the new electronic supplementary material, figure S4.

References:

1. Leite, D. J. et al. 2018 Homeobox gene duplication and divergence in arachnids. *Mol. Biol. Evol.* **35**, 2240–2253. (doi:10.1093/molbev/msy125)

2. Ontano, A. Z. et al. 2021 Taxonomic sampling and rare genomic changes overcome long-branch attraction in the phylogenetic placement of pseudoscorpions. *Mol. Biol. Evol.* (doi:10.1093/molbev/msab038)

3. Brunner, D., Dücker, K., Oellers, N., Hafen, E., Scholzi, H. & Klämbt, C. 1994 The ETS domain protein Pointed-P2 is a target of MAP kinase in the Sevenless signal transduction pathway. *Nature* **370**, 386–389. (doi:10.1038/370386a0)

4. Galindo, M. I., Bishop, S. A. & Couso, J. P. 2005 Dynamic EGFR-Ras signalling in *Drosophila* leg development. *Dev Dyn.* **233**, 1496–1508. (doi:10.1002/dvdy.20452)

5. Grossmann, D. & Prpic, N.-M. 2012 Egfr signaling regulates distal as well as medial fate in the embryonic leg of *Tribolium castaneum*. *Dev Biol* **370**, 264–272. (doi:10.1016/j.ydbio.2012.08.005)

6. Roelofs, D., Zwaenepoel, A., Sistermans, T., Nap, J., Kampfraath, A. A., Van de Peer, Y., Ellers, J. & Kraaijeveld, K. 2020 Multi-faceted analysis provides little evidence for recurrent whole-genome duplications during hexapod evolution. *BMC Biol.* **18**, 57–13. (doi:10.1186/s12915-020-00789-1)

Appendix B

Associate Editor
Comments to Author:
I want to congratulate the authors for their efforts to improve the paper with more data and experiments, as well as by editing the manuscript itself.

Answer: We thank the Associate Editor for the positive feedback.

I think the manuscript could be improved further by clarifying the situation with the miRNAs duplicated in both opiliones and arachno pulmonate, are these duplications considered not-homologous (they are in-paralogs of each lineage rather than out-paralogs?) due to evidence from gene trees, synteny, or something else?

Answer: We have amended to text to clarify that we consider those duplications to be lineage specific (mir-29), or the result of an ancient duplication stemming from the most recent common ancestor of Arthropoda (mir-2, mir 87, and mir-263). We additionally implemented suggestions in this section following reviewer 2 (see below).

I'd also like to encourage the authors to make the phylogeny in Figure 2C consistent with the other ones in the paper (i.e., collapsed), unless this is the result of clustering based on the number of miRNA copies, but I don't think that is the case .

Answer: We have now collapsed the nodes as suggested.

Other than that, I'd like to kindly ask the authors to address the concerns of the referee.

Reviewer(s)' Comments to Author:

Referee: 2

Comments to the Author(s).
This remains a clear, thorough manuscript about an interesting organism. I believe that the extensive revisions have substantially clarified the main focus of the paper and I think it would be appropriate for the journal following minor revisions to clarify a few points.

Substantive questions about interpretation that should be clarified/fixed:

l. 203-205: As stated, this claim is incorrect because all 4 miRNA that appear duplicated in the harvestman also appear to be duplicated in Arachnopulmonata. Some of what precedes it is also vague (because of the 'most', when in fact it appears to be true of most). And key evidence is omitted (re. duplication outside of chelicerates).
Is the following rephrasing attempt accurate, starting with claims on l. 200? "Three of these four microRNAs appear [or are parsimoniously interpreted as having undergone' to

have undergone an initial duplication in an arthropod ancestor and subsequent duplication in the Arachnopulmonata and horseshoe crabs. The other, miR-29, has two copies in harvestmen, horseshoe crabs and a subset of arachnopulmonates, which suggests multiple independent duplications rather than an ancestral duplication."
[I see that is getting complicated—perhaps the possible miR-29 history doesn't need to be explained, just noted that history does not suggest a shared ancestral duplication event.]

Answer: We have clarified this passage as follows: "These microRNAs, with the exception of mir-29, are also duplicated in most other chelicerates and outgroup arthropods, (electronic supplementary material, table S2), suggesting the origin of paralogs at the arthropod common ancestor (figure 2c). The presence of duplicated mir-29 in harvestmen, horseshoe crabs and a subset of Arachnopulmonata suggests separate independent duplication events in these lineages, although this parsimonious inference is contingent upon the resolution of the position of these groups in arachnid phylogeny."

l. 346-349: I'm not fully following the argument for subfunctionalization. I think that's because the evidence isn't totally consistent with the conclusion as the spider data would have been predicted to show function of 1 copy in L1 and the other copy in L2 (as opposed to no evidence of Dfd function in L2).

Answer: We agree with the reviewer that the current data does not outright support a function of *Ptep-DfdB* in L2, as there are no reported functional studies for this copy. Nonetheless, the data in *Phalangium opilio* demonstrating a two-segment function for the single copy *Dfd* (plesiomorphic condition), together with the evidence that *Ptep-DfdA* patterns only one segment (L1), are suggestive that there was a subdivision of function in arachnopulmonates. This is further substantiated by the spatial and temporal restriction of spider *DfdB* to the ventral ectoderm of the L1 and L2 segments [1], in contrast to *Ptep-DfdA* and the known single-copy *Dfd* expression domains of harvestman, mite (Telford and Thomas 1998), and sea spider [2]. We agree that demonstrating this hypothesis will require RNAi against *Ptep-DfdB* and double-knockdown of both spider *Dfd* paralogs, but these experiments are outside the scope of our investigation into harvestman leg patterning.

Action taken: We have amended the text to clarify the differences in expression between the spider *Dfd* paralogs, and emphasized the conjectural character of this inference based on available evidence.

l. 352-354: The way this conclusion is stated, it comes across as a novel conclusion made possible by the evidence in this study. But in fact expression data from tardigrades and onychophora, paleontological evidence supporting multiple origins of tagmata, and evidence that ancestral arthropods lacked tagmata, but would have inherited offset

anterior Hox boundaries (which are found across bilateria). Even if the boundaries were different in the harvestman, the conclusion would still be supported. So the presentation should be altered to reflect the fact that this paper is adding minor confirmatory evidence rather than supporting a new interpretation.

Answer: We agree with the reviewer. We have modified the text as follows, to denote the confirmatory role of this claim: "These results bring further support for the notion that the establishment of some Hox anterior boundaries predates the evolution of tagmata, with further substantiation from Hox anterior boundaries in Onychophora [3]."

l. 357-358: Clarify that the argument for functional redundancy only applies to L3 (whereas from context, one might expect the claim to apply to L1-L3)

Answer: As suggested, we now specified that the functional redundancy applies to L3 only.

Other recommendations:

l. 164 (and supplemental methods l. 162-165): In the ms. text, it is important to indicate that this is the number of predicted genes after considerable cleaning of the dataset. In the supplemental methods, please also indicate either the total numbers of genes initially output from BRAKER or the number removed at each cleaning step.

Answer: We have now clarified that the numbers refer to "after filtering steps". The total number of the initial BRAKER output is currently available in supplementary methods, in the section referring to BRAKER (line 164). The number of genes in each filtering step is also mentioned there.

l. 254: clarify number; my first reaction was that 8/177 was surprisingly low, but from supplement, I see that's a result of mortality. Perhaps present as "Mortality was high (X of 177), but 8 of X surviving individuals showed partial p-to-l transformations"

Answer: clarified as suggested.

l. 406-411: I found the concluding paragraph to be a letdown. I think a stronger version would point out that so far the genome has confirmed/extended evidence of conservation of patterning and then, if there are any clues, elaborate on how it may now be used to refine our understanding of new traits.

Answer: Here we will respectfully request some latitude from the reviewer and the Associate Editor. Every research group has their own priorities for future efforts, as well

as their own subjective decisions as to the writing of a paper's conclusion. This manuscript reflects ours.

Action taken: None.

Fig. 2: consider deleting panel B (where new data is redundant with A, and a expected in the context of arthropod Hox clusters) and condensing A to the same width as C;

Answer: We thank the reviewer for the suggestion, but we subscribe that the panel B is essential to for readers not familiar with arthropod Hox evolution and to assist with reaching a broader audience that may be interested in *Phalangium* genome.

Action taken: None.

Fig. 3: I recommend replacing schematic in A with in situ images of the Hox gene expression, given that rings of expression occur along the legs in both genes, so the variation in expression levels shown diagrammatically is too oversimplified.
Also, bold font in C is hard to see; additional emphasis from colored font, would help these stand out.

Answer: In order to follow the reviewer suggestion and push further the limits of the *Phalangium opilio* system, we performed for the first time in arachnids a Hybridization Chain Reaction v.3 (HCR) [4] in situ hybridization to assess the expression of both *Dfd* and *Scr* in the same embryo. This technique provided the unprecedented resolution of transcripts at the cellular level and new data on the precise colocalization of *Dfd* and *Scr* transcripts. The merged figures are provided in figure 3*a*, and the individual expression of each gene is included as supplementary material. We also included a description of the HCR methods in the supplementary material.

Fig. 3 legend: please explain use of color in R; I also found "sum of halves" not entirely clear.

Supp EGFR tree: why were the 5' and 3' segments of EGFR from Conichochernes crassus analyzed separately? Their different placements on the tree call into question the robustness of these phylogenetic results.

Answer:

(1) We included the following explanation: "Lighter color indicates weaker penetrance.".
(2) We replaced "sum of halves" for "# of affected halves".
(3) The two fragments are inferred to be the result of a fragmentary assembly, not two paralogs. We do not possess evidence to allow joining these two sequences, so we opted to analyze them as separate terminals, as we cannot rule out the

possibility that they could be separate genes. While the reviewer is right to point out that the phylogenetic analysis is affected by the introduction of missing data in those sequences, the main purpose of the analysis was to confirm the orthology of the *Phalangium* sequences. We therefore contend that no further action is required to substantiate our present claim of homology (the identity of the *Phalangium* sequences as *Egfr* homologs is not in question). Future studies interested in understanding the evolution of *Egfr* paralogs in Chelicerate should certainly consider a broader sample and complete gene sequences.

Typos, etc.:

l. 78: change 'derived' to a term that does not reinforce the misconception that some extant groups are ancestral and others are derived.

Answer: Replaced with "across insects".

l. 123: add space: "16 cells"

Answer: Corrected.

l. 166: add 'and' before 'the mite'

Answer: Added.

l. 181: rephrase 'non-collinearity': it sounds like you are suggesting that colinearity may be maintained (and that the apparent fragmentation indicates an incomplete assembly).

Answer: We clarified this sentence as follows: "In addition to the small size, these scaffolds contained very few or no adjacent genes, suggesting that position of these four Hox genes outside of the main cluster is an artifact of fragmentary assembly."

l. 219: delete 'similarly' or make it explicit what aspect of expression is similar between Dfd and Scr.

Answer: Deleted.

l. 329-330: perhaps rephrase as "Nevertheless, the assumed unduplicated condition.....has not been rigorously tested previously."

Answer: rephrased as suggested.

l. 337 and l. 352: replace 'notion' with 'inference' or 'conclusion'

Answer: Replaced as suggested.

l*. 343: unclear whether you mean leg segments or body segments. Only evidence shown is for legs; what about changing 'segments' to 'legs'?*

Answer: Rephrased: "...affects the identity of legs 1 and 2. "

l. 361: rephrase (abdominal segments acquire thoracic identity or transformation of abdominal to thoracic identity)

Answer: Rephrased for clarification.

l. 365: Perhaps 'informative' would be a better descriptor than 'powerful'?

Answer: Modified as suggested.

l. 393: change 'median' to 'medial' or 'intermediate'

Answer: Modified as suggested.

l. 394-399: replace discussion of short germband taxa with an evolutionarily meaningful group.

Answer: We emended this text to provide parenthetical examples of specific insect models that exhibit these developmental dynamics.

1. Schwager, E. E. et al. 2017 The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. *BMC Biol.* **15**, 62. (doi:10.1186/s12915-017-0399-x)

2. Jager, M., Murienne, J., Clabaut, C., Deutsch, J., Guyader, H. L. & Manuel, M. 2006 Homology of arthropod anterior appendages revealed by Hox gene expression in a sea spider. *Nature* **441**, 506–508. (doi:10.1038/nature04591)

3. Janssen, R., Eriksson, B. J., Tait, N. N. & Budd, G. E. 2014 Onychophoran Hox genes and the evolution of arthropod Hox gene expression. *Front Zool* **11**, 22. (doi:10.1186/1742-9994-11-22)

4. Choi, H. M. T., Schwarzkopf, M., Fornace, M. E., Acharya, A., Artavanis, G., Stegmaier, J., Cunha, A. & Pierce, N. A. 2018 Third-generation in situ hybridization

chain reaction: multiplexed, quantitative, sensitive, versatile, robust. *Development* **145**. (doi:10.1242/dev.165753)