THE ROYAL SOCIETY
PUBLISHING

PROCEEDINGS B

Target-capture phylogenomics provide insights on gene and species tree discordances in Old World Treefrogs (Anura: Rhacophoridae)

Kin Onn chan, Carl R. Hutter, Perry L. Wood, L. Lee Grismer and Rafe M. Brown

Article citation details
Proc. R. Soc. B **287**: 20202102.
http://dx.doi.org/10.1098/rspb.2020.2102

Review timeline

Original submission:	17 June 2020
1st revised submission:	27 August 2020
2nd revised submission:	9 November 2020
Final acceptance:	13 November 2020

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

Review History

RSPB-2020-1438.R0 (Original submission)

Review form: Reviewer 1

Recommendation
Major revision is needed (please make suggestions in comments)

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

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

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

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?**
> Yes
>
> **Is it adequate?**
> No

Do you have any ethical concerns with this paper?
Yes

Comments to the Author
The authors sought to investigate the sources of discordance in phylogenomics using a combination of different types of molecular data, from high-throughput to the inclusion of a small number of traditional markers. I liked the intention, especially because it tried to address a group of anurans with historical taxonomic problems. However, there are serious issues about the analysis and the assumptions made in this paper, which need to be addressed in order for their claims to be supported.

Major observations:

1) The authors dismissed hidden paralogy, which is a fundamental issue and a potential source of discordance in their data. Hence one of the major claims in the paper: that the discordance observed is solely due to ILS is not entirely correct unless they prove that their dataset does not have hidden paralogy. I encourage the authors to consult some papers which have investigated this issue:

- Siu-Ting et al. 2019. Inadvertent Paralog Inclusion Drives Artifactual Topologies and Timetree Estimates in Phylogenomics. Molecular Biology and Evolution. https://doi.org/10.1093/molbev/msz067

- Struck, T. 2013. The Impact of Paralogy on Phylogenomic Studies – A Case Study on Annelid Relationships. PLOS ONE. https://doi.org/10.1371/journal.pone.0062892

In order to make the paper more robust, the authors should show how they identified orthologs and paralogs and check for these in their data. Siu-Ting et al. 2019 presents a method to investigate this issue in gene trees, which could be useful and a rapid way to evaluate paralogous gene trees.

2) Carry out an estimation of saturation for all gene trees - this is key in particular when dealing with nucleotides rather than amino acids. I would like to see if this is or not an issue for their datasets.

Following from this, it would be interesting if the authors had explored if carrying out a phylogenetic inference on the amino acid alignment for the exon dataset would retrieve similar results.

3) Perform a partitioned analysis for concatenated datasets using the models inferred for those gene trees obtained. IQ-Tree allows for specification of multiple partitions with very large datasets when run using multiple processors (such as those in High Performance Clusters).

4) The authors make statements such as "correct topology" and "wrong topology". What was the criteria used to determine which is the right and the wrong topology?
My suggestion is to carry out AU tests for all topologies and all datasets to discard that the other topologies retrieved are not equally good descriptions of the data. I think this is particularly important to support such statements.

5) The chi-square test used by the authors "We then used a chi-square test to determine whether the frequency of gene trees (gCF) and sites (sCF) supporting the two alternate topologies was significantly different. Insignificant p-values ($p > 0.05$) indicate a failure to reject the hypothesis of equal frequencies, indicating that discordance among gene trees and/or sites is likely due to ILS." It seems to me a very large leap from topologies having all equal frequencies inferring that this is due to ILS. It's an assumption that is not clear.

6) I feel that the authors have missed on the opportunity to discuss further on types of data and how these can improve phylogenomic resolution. I think this would be one of the most informative and helpful things for researchers doing these type of analyses to know. But instead, only a couple of short sentences in the discussion were given for this. I would like to see the authors explore the impact of the types of data, link this to phylogenetic informativeness, to topology support and if resolution is actually improved or not and what could be happening. In my opinion devoting more space to discussing this would be more appropriate for a general audience, and will distinguish it from a paper that seems to be more focused for amphibian taxonomists.

Minor observations:
- Line 64-67: "Discordance between gene trees and the species tree can also result from biological processes such as introgression, horizontal gene transfer, and incomplete lineage sorting [35,38–44]". The authors have forgotten to list here one of the major sources of discordance is gene duplication and gene loss. In particular, failing to identify orthologous from paralogous genes will affect phylogenetic estimation greatly. See point 1) of major observations.
- Line 178-179: Authors assume that the only source of incongruence that they have is ILS without discarding other sources of incongruence. See point 1) of major observations.
- Lines 209-210: specify units, did you mean base pairs for numbers supplied?
- Line 221: congruent: do you mean that all topologies were identical here? Following from this, how do you reconcile here which is your optimal tree?
- Line 272: not clear what is the test carried out: what is the null hypothesis and what is meant by equal gene tree frequencies. Needs to be more clear in the main text.
line 285: "anomalous gene trees" - can you be more specific, what do you mean by anomalous in this case?
- Lines 296-299: "Through systematic analysis of different classes of data, we were able to demonstrate that ILS caused by rapid diversification events was responsible for most of the deep-level discordances in Old World treefrogs." I don't think this is true since authors have not discarded other sources of discordance. See major observation point 1).
- Missing more details in the methods: especially in the design of the bates used. At least provide a short summary.
- Other details missing in the Supplementary Materials Methods: Needs to provide more details about parameters and thresholds used for it to be fully reproducible. In particular in the matching against probe reference, since this point is key for the identification of orthologs and paralogs.
- What constituted the "Legacy" dataset? from what publications? include references.
- In Methods: specify software or code used to calculate gCF and sCF.
- Provide more details about software used to determine phylogenetic informative sites.

- Bibliography has several spelling mistakes which need to be double-checked and addressed.
- Figure 2 B seem a bit pointless. In Fig. 2B you are comparing values bootstrap values that are all 100 against branchlengths, it is obvious that you will not be able to see a correlation. I don't see the point of this graph.
- Figure 3 caption needs to be more informative. What filtering was used for each type of dataset?
- Supplementary Materials font seems to be different on the second
 half of the document. was that supposed to be like that?
- All alignments and gene trees inferred should be provided in a repository to ensure results will be replicable.

Review form: Reviewer 2 (Jeffrey Streicher)

Recommendation
Major revision is needed (please make suggestions 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?
Good

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?
Yes

Is it clear?
Yes

Is it adequate?
Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
RSPB 2020-1438

This was a co-review and overall we both enjoyed reading this manuscript. We thought that it used a large, novel dataset to demonstrate some important qualities of how 'short and deep' branches relate to gene tree discordance and bootstrap proportions. We both feel that it is a manuscript with potential broad appeal to the readership of Proceedings B. However, we thought several key analyses should be reconsidered (particularly the direct comparison of gene trees estimated using different model selection criteria) and in several instances that the methods lacked explanatory detail. We have listed below some major and minor comments that we hope will be helpful in constructing a revised version of the manuscript prior to publication.

Major comments:

1. Different introns can be under different selection pressures (e.g. intron with regulatory regions/intron without, selfsplicing intron/non-selfsplicing intron, etc.), might this affect the results? Would it be worth trying to ascertain whether the introns used can be binned into neutral/positive/negative/stabilising selection? This is not a request for the authors to do this, but it is worth considering the potential implications to the results.

2. Were non-Sanger and non-UCE datasets checked for paralogs? If these are present they can result in ILS-like behaviour.

3. The authors' worry about computational intractability is understandable, but means that the results of the concatenated analysis are not comparable to the gene tree analyses under the estimated models: The concatenated analysis was run under the GTR+G model with an unpartitioned alignment (why not GTR+I+G?); gene trees are necessarily "partitioned", and they were subjected to model testing. We suggest running all gene trees under the GTR+G model, and compare the results to the gene trees obtained under the selected models. Alternatively, rate heterogenous models can be used in lieu of the more commonly used models and would preclude the need for model testing. The concatenated tree should also be run under gene partitions. If there are no significant differences, then the comparison between the concatenated and gene tree analyses stands. If not, the conclusions will need to be re-interpreted and revised.

4. Were quartet scores checked in the Astral trees, or just local PP?

5. Which IQTree output was used as input for ASTRAL? Binary best ML tree, greedy consensus or 50% majority-rule consensus? The level of input tree resolution affects the resolution of the tree summarised by ASTRAL. All binary trees can lead to erroneously well resolved trees, while the 50%MRC can lead to more conservative estimates of topology with the output tree being less resolved than the "true" tree. Check "5.1.1 Gene tree uncertainty" in Mirarab 2019 arXiv:1904.03826v2

6. While an explanation about the anomaly zone is given in the supplementary materials a bit more detail would be helpful to understand the manuscript. At the very least, the assumptions of the approach should be stated explicitly, and how they may affect the results discussed.

7. Given that the paper describing concordance factors has only recently been published, it would be helpful to have more a more detailed description of the method in the manuscript. It is quite likely that people that do not regularly use IQTree are not aware of concordance factors, even if the preprint has been around for at least two years.

8. Parsimony Uninformative and Constant Sites contribute to branch lengths, but not topology. Thus, it makes sense that their removal would not affect site concordance factors. Might be worth saying explicitly how the BLs changed between the PIS filtered and unfiltered analyses. Would make for easier reading.

Minor comments:
Page 2, line 49: "...high bootstrap support values..." would be better as "... high bootstrap

proportions...", previous clause already mentions branch support

Page 3, Line 67: "...ancestral polymorphisms fail..." would be clearer as "...ancestral polymorphisms don't..."

Page 3, Line 68: "...failure of lineages to coalesce..." is confusing. Coalescence and lineage sorting might be considered different temporal concepts; one backwards in time (coalescence) the other forwards in time (lineage sorting). Please clarify.

Page 4, line 116: "i.e." should be substituted for ":", "i.e." is used to explain the same information with different wording, not to signify enumeration

Page 4, line 118: "...holdings of University..." reads better as "...holdings of the University..."

Page 5, line 138/9: "Capella-gutiérrez" should be "Capella-Gutiérrez". Also should it be referenced as a number in Proc B style?

Page 5, line 139/140: consider using "...at leat 50% of..." instead of "...at least 50 percent of..."

Page 5, line 145: "...missing taxon representation and..." reads better as "...missing taxa and..."

Page 5, line 160: ASTRAL-III was not cited in the manuscript

Page 6, line 185: Does "polytomy" refer to a hard or soft polytomy? Is the polytomy due to uncertainty in the data (soft), or is the "true" tree non-binary (hard)?

Page 6, line 188: "...and will generate..." would be better as "...and can generate/yield...", even though the branches are short there is still p=1/3 that the correct topology is recovered

Page 6, line 190/191: "...ancestor-and-descendent..." should be "...ancestor-and-descendant...", this is also present in the supplementary materials

Page 9, line 295: "...soft polytomy...", How was hard vs soft ascertained? Links to Page 6, line 185

Page 9, line 296-299: "Through systematic ... treefrogs." should be re-worded to show that it is a hypothesis. Until time machines are available the causes of ILS in frogs will probably remain hypothetical. Also the datasets were not tested for the potential causes of ILS, only whether there was ILS.

Page 10, line 326-329: "This shows ... before." To our current understanding of the manuscript, the possibility of hard polytomies was not discarded, so the lack of resolution at short internodes may not be artefactual. Also, we are aware of several other studies that have demonstrated short internodes remain challenging with large 'phylogenomic' datasets (e.g. Streicher et al. 2016. Syst. Biol. 65: 128-145; Burbrink et al. 2020 Syst. Biol. 69: 502-520; Roycroft et al. Syst. Biol. 69: 431-444; Singhal et al. Syst. Biol. In Press), so it would be appropriate to indicate via citations that this is a wider discussion in the literature at the moment.

Page 10, line 330: "total evidence" While multiple datasets/types were used, we do not feel that the term total evidence is warranted. There were no morphological data used, and there was only a passing reference to life-history traits.

Page 11, line 357: "... Pyron and Wiens [2001]) and ..." should be "... Pyron and Wiens [2001] and ...", there is an extra ")"

Page 12, line 376: "... funded NSF..." reads better as "...funded by NSF..."

Figure 1: Adding a cladogram for topology T1 would make comparison to the other cladograms easier. Suggest keeping phylogram as is and rearrange cladograms to show all 5 topologies.

Figure 4: Might be easier to interpret if each topology has a different colour globally, rather than just each topology with each node.
Also changing the y-axis label to "Cumulative percentage" might make the graph easier to interpret.

Citations: In-text citation style is inconsistent, some have Author (date), Author [date], Author date or [#]. Under the journal's style they should be: Author [#] at the start of a sentence, or [#] in all other instances. Also, update Minh BQ, Hahn MW, Lanfear R. 2018 New methods to calculate concordance factors for phylogenomic datasets. bioRxiv , doi: http://dx.doi.org/10.1101/487801. It has now been published.

Thank you very much for the opportunity to review this manuscript. We are happy for the authors to contact us if they need any clarification on our comments. Ana Serra Silva (a.da-silva@nhm.ac.uk) and Jeff Streicher (j.streicher@nhm.ac.uk).

Decision letter (RSPB-2020-1438.R0)

07-Aug-2020

Dear Dr Chan:

I am writing to inform you that your manuscript RSPB-2020-1438 entitled "Target-capture phylogenomics provide insights on gene and species tree discordances in Old World Treefrogs (Anura: Rhacophoridae)" 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,
Dr Sasha Dall
mailto: proceedingsb@royalsociety.org

Associate Editor
Board Member: 1
Comments to Author:
Two experts in the field have reviewed your work, and both have identified methodological issues with the phylogenetic analysis (e.g. the possible misleading effect of hidden paralogy). Furthermore, one of the reviewers pointed out the lack of any collecting permit authorisations for collecting the samples. Finally, all alignments, trees and scripts need to be made accessible in a public repository upon the paper acceptance. Considering the nature of the reviewers' comments, I cannot recommend the MS for publication on Proc of the Royal Society B.

Reviewer(s)' Comments to Author:
Referee: 1

Comments to the Author(s)
The authors sought to investigate the sources of discordance in phylogenomics using a combination of different types of molecular data, from high-throughput to the inclusion of a small number of traditional markers. I liked the intention, especially because it tried to address a group of anurans with historical taxonomic problems. However, there are serious issues about the analysis and the assumptions made in this paper, which need to be addressed in order for their claims to be supported.

Major observations:

1) The authors dismissed hidden paralogy, which is a fundamental issue and a potential source of discordance in their data. Hence one of the major claims in the paper: that the discordance observed is solely due to ILS is not entirely correct unless they prove that their dataset does not have hidden paralogy. I encourage the authors to consult some papers which have investigated this issue:
- Siu-Ting et al. 2019. Inadvertent Paralog Inclusion Drives Artifactual Topologies and Timetree Estimates in Phylogenomics. Molecular Biology and Evolution. https://doi.org/10.1093/molbev/msz067

- Struck, T. 2013. The Impact of Paralogy on Phylogenomic Studies – A Case Study on Annelid Relationships. PLOS ONE. https://doi.org/10.1371/journal.pone.0062892

In order to make the paper more robust, the authors should show how they identified orthologs and paralogs and check for these in their data. Siu-Ting et al. 2019 presents a method to investigate this issue in gene trees, which could be useful and a rapid way to evaluate paralogous gene trees.

2) Carry out an estimation of saturation for all gene trees - this is key in particular when dealing with nucleotides rather than amino acids. I would like to see if this is or not an issue for their datasets.
Following from this, it would be interesting if the authors had explored if carrying out a phylogenetic inference on the amino acid alignment for the exon dataset would retrieve similar results.

3) Perform a partitioned analysis for concatenated datasets using the models inferred for those gene trees obtained. IQ-Tree allows for specification of multiple partitions with very large datasets when run using multiple processors (such as those in High Performance Clusters).

4) The authors make statements such as "correct topology" and "wrong topology". What was the criteria used to determine which is the right and the wrong topology?
My suggestion is to carry out AU tests for all topologies and all datasets to discard that the other topologies retrieved are not equally good descriptions of the data. I think this is particularly important to support such statements.

5) The chi-square test used by the authors "We then used a chi-square test to determine whether the frequency of gene trees (gCF) and sites (sCF) supporting the two alternate topologies was significantly different. Insignificant p-values ($p > 0.05$) indicate a failure to reject the hypothesis of equal frequencies, indicating that discordance among gene trees and/or sites is likely due to ILS." It seems to me a very large leap from topologies having all equal frequencies inferring that this is due to ILS. It's an assumption that is not clear.

6) I feel that the authors have missed on the opportunity to discuss further on types of data and how these can improve phylogenomic resolution. I think this would be one of the most informative and helpful things for researchers doing these type of analyses to know. But instead, only a couple of short sentences in the discussion were given for this. I would like to see the authors explore the impact of the types of data, link this to phylogenetic informativeness, to topology support and if resolution is actually improved or not and what could be happening. In my opinion devoting more space to discussing this would be more appropriate for a general audience, and will distinguish it from a paper that seems to be more focused for amphibian taxonomists.

Minor observations:
- Line 64-67: "Discordance between gene trees and the species tree can also result from biological processes such as introgression, horizontal gene transfer, and incomplete lineage sorting [35,38–44]". The authors have forgotten to list here one of the major sources of discordance is gene duplication and gene loss. In particular, failing to identify orthologous from paralogous genes will affect phylogenetic estimation greatly. See point 1) of major observations.
- Line 178-179: Authors assume that the only source of incongruence that they have is ILS without discarding other sources of incongruence. See point 1) of major observations.
- Lines 209-210: specify units, did you mean base pairs for numbers supplied?
- Line 221: congruent: do you mean that all topologies were identical here? Following from this, how do you reconcile here which is your optimal tree?
- Line 272: not clear what is the test carried out: what is the null hypothesis and what is meant by equal gene tree frequencies. Needs to be more clear in the main text.
line 285: "anomalous gene trees" - can you be more specific, what do you mean by anomalous in this case?
- Lines 296-299: "Through systematic analysis of different classes of data, we were able to demonstrate that ILS caused by rapid diversification events was responsible for most of the deep-level discordances in Old World treefrogs." I don't think this is true since authors have not discarded other sources of discordance. See major observation point 1).
- Missing more details in the methods: especially in the design of the bates used. At least provide a short summary.
- Other details missing in the Supplementary Materials Methods: Needs to provide more details about parameters and thresholds used for it to be fully reproducible. In particular in the matching against probe reference, since this point is key for the identification of orthologs and paralogs.
- What constituted the "Legacy" dataset? from what publications? include references.
- In Methods: specify software or code used to calculate gCF and sCF.
- Provide more details about software used to determine phylogenetic informative sites.
- Bibliography has several spelling mistakes which need to be double-checked and addressed.

- Figure 2 B seem a bit pointless. In Fig. 2B you are comparing values bootstrap values that are all 100 against branchlengths, it is obvious that you will not be able to see a correlation. I don't see the point of this graph.
- Figure 3 caption needs to be more informative. What filtering was used for each type of dataset?
- Supplementary Materials font seems to be different on the second
half of the document. was that supposed to be like that?
- All alignments and gene trees inferred should be provided in a repository to ensure results will be replicable.

Referee: 2
Comments to the Author(s)
RSPB 2020-1438

This was a co-review and overall we both enjoyed reading this manuscript. We thought that it used a large, novel dataset to demonstrate some important qualities of how 'short and deep' branches relate to gene tree discordance and bootstrap proportions. We both feel that it is a manuscript with potential broad appeal to the readership of Proceedings B. However, we thought several key analyses should be reconsidered (particularly the direct comparison of gene trees estimated using different model selection criteria) and in several instances that the methods lacked explanatory detail. We have listed below some major and minor comments that we hope will be helpful in constructing a revised version of the manuscript prior to publication.

Major comments:

1. Different introns can be under different selection pressures (e.g. intron with regulatory regions/intron without, selfsplicing intron/non-selfsplicing intron, etc.), might this affect the results? Would it be worth trying to ascertain whether the introns used can be binned into neutral/positive/negative/stabilising selection? This is not a request for the authors to do this, but it is worth considering the potential implications to the results.

2. Were non-Sanger and non-UCE datasets checked for paralogs? If these are present they can result in ILS-like behaviour.

3. The authors' worry about computational intractability is understandable, but means that the results of the concatenated analysis are not comparable to the gene tree analyses under the estimated models: The concatenated analysis was run under the GTR+G model with an unpartitioned alignment (why not GTR+I+G?); gene trees are necessarily "partitioned", and they were subjected to model testing. We suggest running all gene trees under the GTR+G model, and compare the results to the gene trees obtained under the selected models. Alternatively, rate heterogenous models can be used in lieu of the more commonly used models and would preclude the need for model testing. The concatenated tree should also be run under gene partitions. If there are no significant differences, then the comparison between the concatenated and gene tree analyses stands. If not, the conclusions will need to be re-interpreted and revised.

4. Were quartet scores checked in the Astral trees, or just local PP?

5. Which IQTree output was used as input for ASTRAL? Binary best ML tree, greedy consensus or 50% majority-rule consensus? The level of input tree resolution affects the resolution of the tree summarised by ASTRAL. All binary trees can lead to erroneously well resolved trees, while the 50%MRC can lead to more conservative estimates of topology with the output tree being less resolved than the "true" tree. Check "5.1.1 Gene tree uncertainty" in Mirarab 2019 arXiv:1904.03826v2

6. While an explanation about the anomaly zone is given in the supplementary materials a bit more detail would be helpful to understand the manuscript. At the very least, the assumptions of the approach should be stated explicitly, and how they may affect the results discussed.

7. Given that the paper describing concordance factors has only recently been published, it would be helpful to have more a more detailed description of the method in the manuscript. It is quite likely that people that do not regularly use IQTree are not aware of concordance factors, even if the preprint has been around for at least two years.

8. Parsimony Uninformative and Constant Sites contribute to branch lengths, but not topology. Thus, it makes sense that their removal would not affect site concordance factors. Might be worth saying explicitly how the BLs changed between the PIS filtered and unfiltered analyses. Would make for easier reading.

Minor comments:

Page 2, line 49: "...high bootstrap support values..." would be better as "... high bootstrap proportions...", previous clause already mentions branch support

Page 3, Line 67: "...ancestral polymorphisms fail..." would be clearer as "...ancestral polymorphisms don't..."

Page 3, Line 68: "...failure of lineages to coalesce..." is confusing. Coalescence and lineage sorting might be considered different temporal concepts; one backwards in time (coalescence) the other forwards in time (lineage sorting). Please clarify.

Page 4, line 116: "i.e." should be substituted for ":", "i.e." is used to explain the same information with different wording, not to signify enumeration

Page 4, line 118: "...holdings of University..." reads better as "...holdings of the University..."

Page 5, line 138/9: "Capella-gutiérrez" should be "Capella-Gutiérrez". Also should it be referenced as a number in Proc B style?

Page 5, line 139/140: consider using "...at leat 50% of..." instead of "...at least 50 percent of..."

Page 5, line 145: "...missing taxon representation and..." reads better as "...missing taxa and..."

Page 5, line 160: ASTRAL-III was not cited in the manuscript

Page 6, line 185: Does "polytomy" refer to a hard or soft polytomy? Is the polytomy due to uncertainty in the data (soft), or is the "true" tree non-binary (hard)?

Page 6, line 188: "...and will generate..." would be better as "...and can generate/yield...", even though the branches are short there is still p=1/3 that the correct topology is recovered

Page 6, line 190/191: "...ancestor-and-descendent..." should be "...ancestor-and-descendant...", this is also present in the supplementary materials

Page 9, line 295: "...soft polytomy...", How was hard vs soft ascertained? Links to Page 6, line 185

Page 9, line 296-299: "Through systematic ... treefrogs." should be re-worded to show that it is a hypothesis. Until time machines are available the causes of ILS in frogs will probably remain hypothetical. Also the datasets were not tested for the potential causes of ILS, only whether there was ILS.

Page 10, line 326-329: "This shows ... before." To our current understanding of the manuscript, the possibility of hard polytomies was not discarded, so the lack of resolution at short internodes may not be artefactual. Also, we are aware of several other studies that have demonstrated short internodes remain challenging with large 'phylogenomic' datasets (e.g. Streicher et al. 2016. Syst. Biol. 65: 128-145; Burbrink et al. 2020 Syst. Biol. 69: 502-520; Roycroft et al. Syst. Biol. 69: 431-444; Singhal et al. Syst. Biol. In Press), so it would be appropriate to indicate via citations that this is a wider discussion in the literature at the moment.

Page 10, line 330: "total evidence" While multiple datasets/types were used, we do not feel that the term total evidence is warranted. There were no morphological data used, and there was only a passing reference to life-history traits.

Page 11, line 357: "... Pyron and Wiens [2001]) and ..." should be "... Pyron and Wiens [2001] and ...", there is an extra ")"

Page 12, line 376: "... funded NSF..." reads better as "...funded by NSF..."

Figure 1: Adding a cladogram for topology T1 would make comparison to the other cladograms easier. Suggest keeping phylogram as is and rearrange cladograms to show all 5 topologies.

Figure 4: Might be easier to interpret if each topology has a different colour globally, rather than just each topology with each node.
Also changing the y-axis label to "Cumulative percentage" might make the graph easier to interpret.

Citations: In-text citation style is inconsistent, some have Author (date), Author [date], Author date or [#]. Under the journal's style they should be: Author [#] at the start of a sentence, or [#] in all other instances. Also, update Minh BQ, Hahn MW, Lanfear R. 2018 New methods to calculate concordance factors for phylogenomic datasets. bioRxiv , doi: http://dx.doi.org/10.1101/487801. It has now been published.

Thank you very much for the opportunity to review this manuscript. We are happy for the authors to contact us if they need any clarification on our comments. Ana Serra Silva (a.da-silva@nhm.ac.uk) and Jeff Streicher (j.streicher@nhm.ac.uk).

Author's Response to Decision Letter for (RSPB-2020-1438.R0)

See Appendix A.

RSPB-2020-2102.R0

Review form: Reviewer 1

Recommendation
Major revision is needed (please make suggestions in comments)

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

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

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

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?
Yes

Is it clear?
Yes

Is it adequate?
Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
Overall the manuscript has improved and it's more readable. The authors have made an effort to make it of more interest to a general audience. The explanation of the equal frequencies test in the context of the Multispecies Coalescent is more clear and now the results from these make sense.
But, following from the issues raised before, even though they have tackled several of these, there are still 2 major ones that I do not feel the authors have satisfactorily answered or tackled in their paper (see below Hidden Paralogy and AU Test). I have also added 2 more (Data and Different Data and topologies) that won't take too much for the authors to address and I think it would enhance the paper and the discussion greatly.
As for the numerous minor observations, I also hope that the authors consider making these changes, which will help improve the quality and detail of their paper. I apologise in advance for the lengthy text.

Major observations:

- Hidden Paralogy:
In their response the authors claim that "Namely, the target markers were matched to the Nanorana parkeri and Xenopus genomes, where markers that matched to more than one location in each genome were initially removed altogether from the probe set. It is also possible that lineages may have lineage-specific paralogs not originally found, so during the bioinformatic processing of each sample, we removed any assembled contigs that matched substantially to more than one of the target markers".

Having only 1 match to a probe alignment hardly constitutes a convincing way to determine orthology (and I am not referring to in paralogs either). I will explain why:

As all Amphibians (indeed vertebrates) have undergone 2 whole rounds of genome duplication and multiple gene losses, for 3 species A, B and C, it would be expected that they will have 4 copies of a same locus/gene: A1, B1, C1, A2, B2, C2, A3, B3, C3, A4, B4, C4. Please see attached file with a figure explaining this. In this example, true orthologs will depict the following species tree: ((A+B)+C);
Now imagine the case where after gene losses, in particular cases of late gene loss, you only end up with copies of loci A1, B3, C3, which will appear in single copy in your sequencing (and you will think it is ortholog), but these will yield the topology: (A(B,C)); which is not the species tree topology. This is a case of hidden paralogy, and this is something that you cannot discard in your method, even using probes. This pattern, it would be very difficult to discern in your data, as it will look like ILS, when it is not.
Because of the taxonomic level at which you are working (inter Genera relationships within a Family) it is likely that most of your discordance issues are due to ILS and that hidden paralogy may be a small proportion of your data, but still you cannot discard hidden paralogy. This is an issue that is not unique to this study and it will take a long while to resolve.
Following from this explanation, I would feel more comfortable if the authors added an explicit caveat in their discussion that they cannot discard that there is hidden paralogy in their data in those sentences where they claim that the source of discordance is ILS. You can use as citations the references I gave before to support this (https://doi.org/10.1093/molbev/msz067 and https://doi.org/10.1371/journal.pone.0062892)

In addition to this, I invite the authors again to read those papers because some of these issues are explained there more clearly and it will also come in useful as they state somewhere in their FrogCap pipeline that they intend to include a code/software to identify orthologs in this type of data in the future.

- AU-Test and "optimal tree": Upon the suggestion that the authors perform an AU test, the response was: "We consider the correct topology to be the one derived from a total evidence approach; i.e., the largest and most comprehensive dataset (all-combined). In our opinion, a topology test would not help in this case, because the test is dependent on which alignment was used as a reference (intron, exon, UCEs etc.). This inherently introduces user selection bias. To avoid confusion, we have changed the terminology from “correct topology” to “optimal topology” whenever referring to our results"

I disagree with this explanation, in fact, choosing a "total evidence" as the "optimal" (which is the same as saying "best") topology is also subjective and biased. Why is T4 worse than T1? or T2, T3, T5? only because it doesn't match what they have chosen as their preferred topology? I still am of the opinion that the authors should carry out the AU test, or any hypothesis testing to discard that the alternative topologies are not equal descriptions of the data. Ideally, they should carry out an AU test for each alignment they tested against all topologies. This is something that can be carried out very simply in IQtree. At the minimum, they should carry out this test with their "optimal alignment" and testing all different retrieved topologies against this. This analysis would also add support to the authors' taxonomy suggestions at the end of the Discussion (section 4.4).

Bottomline is, unless the authors can prove that one topology is better than another through a test as explained before, using the terms "best", optimal", "worse", "suboptimal", "right", "wrong" is not correct and not supported.

- Data displayed and available: Although the authors have now included the supporting alignments and most of the resulting trees, they still need to include some that I was not able to find (namely the results of the SVDQuartet) and also it is my opinion that the paper would benefit enormously if the authors provided each of the resulting topologies in the Supplementary information (see comments for "Additional results below").

- Different data (exon, intron, UCE, Legacy) yield different topologies because these data are likely undergoing different selection pressures and it is something that has been already observed in phylogenomic studies. I think it warrants a mention in the discussion (I'd say it would go well in section 4.3 Causes of discordance). You should have a look at Shen et al. 2018 https://www.nature.com/articles/s41559-017-0126 and Bravo et al. 2019 https://peerj.com/articles/6399/ .

Minor observations:

- Trees deposited in the Data Dryad: I only found the trees for ASTRAL and IQtree, but couldn't find the SVDQuartet trees. This is either missing or the labels need to be made more clear.

- Methods and Results: 4 or 5 outgroups? In the main text for the methods it says 5 outgroups (which should be all fully listed in the main text). But then table S1 only has 4 outgroups listed. Also there is a major inconsistency with the outgroup species listed: Table S1: missing Scaphiophryne marmorata and Arthroleptis_variabilis. Also said table has Abavorana nazgul which doesn't appear anywhere in the alignments. Please correct this.
Also following from this point, you should add a sentence in the Methods main text of why you chose those 5 outgroups, and how distantly related are they from the ingroup.

- Materials and Methods: Bioinformatics: Line 146. Alignment information should be mentioned here in the main text, also add the software used.

- Lines 190-205: it's missing information on which datasets were Bootstrap carried out and needs to specify the bootstrap calculated (parametric Bootstrap or the Fast bootstrap which are different and both available in IQTree).

- Line 282-286: Where are the analyses that support the statements made? can you supply the distance between the gene trees and the "large" trees or how did you check for this. These statements need reference to the results from your analyses to back it. These should be supplied for both topology and branch lengths.

- Outgroups missing. I was surprised that I couldn't see any of the outgroups in any of the supplied figures. Figure 1a is the only tree you show and it does not depict the full tree of your results. In the spirit of being fully transparent with the information, it would be better to show the outgroups somewhere, such as resulting trees in the Supplementary information (see point "Additional results")

- Figure 1: Also the way the groups are summarised in T2, T3, T4 and T5 are very confusing and one has to spend ages to make sense of the comparisons and the positions of Nodes N1, N2, and N3 across all 5 topologies. I would suggest that the authors label N1-N3 in each of the 5 topologies.

- Table 2: Needs to add more relevant data to support why certain nodes from the different analyses fall in the anomaly zone. You state in the methods and in the Supplementary Methods that you performed "Anomaly Zone calculations". The results of these calculations and the branchlengths should be provided in this table.

- Fig S2: missing labels on the left hand for all vignettes. Also it would be good if you could supply a sentence or 2 in the caption for this to state what is the point you want to highlight, the major result from this.

- Line 380-381: "Quartet scores for UCE datasets were also higher compared to introns and exons, indicating that UCE markers may be less affected by ILS" add reference to results for this: Table 2.

- Additional results: Also I would like to ask the authors to provide a visualisation of the resulting trees for each of the 11 datasets and all the different methods in the supplementary information, as this will help any reader to compare the underlying results.I personally had to copy and paste each of them in a tree viewer, why make life harder to your reviewers when it is better to show all your results in an explicit manner? Please add in each tree a label of Nodes N1, N2 and N3 and which topology they result in (T1, T2, T3, T4 or T5). And also add the support values calculated for each branch based on whatever method used.

- Figures 1 and 3: Node 1: very confusing this "Stelladerma" appearing out of nowhere (not shown in Fig 1, not appearing in any alignments) until you reach the end of the paper and even then, it's not clear the clade of Theladerma that the authors refer to. Anybody who is not familiarised with Rhacophorid or even frog taxonomy are going to find this very confusing. Names for clades and species should be consistent in all figures.

- Figure 3: I noticed that you labeled some of the topologies T1, T2, T3 in some cases next to some of the alternative topologies. But nowhere in these appear T4 ot T5. Is this because of an error? I actually think it would help to illustrate some of your points in the discussion if you labeled in this figure which topology matches which of these hypotheses (for example take Line 425-426 you mention at Node 1 you retrieve T4. Label T4 in the hypothesis that corresponds here for this node in this figure). I think you should do these for all Topologies (T1-T5).

- Phylogenomic relationships (4.4): I find most of this whole section confusing and difficult to follow because the authors do not point to which parts of the trees they refer to (Node 1, Node 2, Node 3).
Also, can you name/find any other supporting evidence for your proposed hypotheses? like morphology, ecology, etc in disfavour of the "Stelladerma hypothesis". It would also help that you define in a very brief way what is this hypothesis.

- Lines 391-392: "Bootstrap values were not correlated with topological concordance and routinely produced strong support for highly discordant nodes (Fig. 2). " You should add this is in agreement with what has already been explain in Minh et al 2020 (https://doi.org/10.1093/molbev/msaa106).

- Line 407: "However, higher concordance values can also be an artefact of small datasets" same as previous observation, this is exactly what was explained in Minh et al 2020 (https://doi.org/10.1093/molbev/msaa106), so it would be good to add this as reference here.

Review form: Reviewer 2 (Jeffrey Streicher)

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?**
> Yes
>
> **Is it clear?**
> Yes
>
> **Is it adequate?**
> Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
This was a co-review of the revised manuscript. We were listed as Referee 2 of the original submission. Overall, we thought the authors did a great job of addressing the reviewer comments. Below we have provided a few minor suggestions for improving the main text prior to publication.

Multiple places in manuscript: When referring to statistical significance please use "non-significant" instead of "insignificant".

Page 2, line 36-37: "We showed ... internal branches." Please clarify, not clear whether causal relationship is between ILS and short branches, or discordance and short branches. Also, short branches result from ILS, not the other way around.

Page 2, line 53: Remove the word "support"

Page 3, line 100: Substitute "UCE's" by "UCEs".

Page 4, line 132: "each unfiltered dataset" maybe remove "unfiltered"

Page 4, line 133: "filtered at 75% sampling completeness", worth clarifying whether it is taxon sampling or length of sequence, wording isn't clear

Page 5, line 165: Consider substituting "(respectively)" for ", respectively,", based on the nearness to the use of parentheses introducing abbreviations and linking support value and software

Page 6, line 173-174: "indicate" and "indicating" within the same sentence, consider swapping one of them for a synonym

Page 6, line 179: Consider "we tested ... for the presence of polytomies" or just "we tested ... for polytomies"

Page 6, line 184: Switch "ancestor-and-descendent" for "ancestor-and-descendant". This persists in the Supplementary materials in section 1.3.

Page 7, line 235: "larger proportion" instead of "larger portion"?

Page 8, line 249 and 251: Add comma before "but"

Page 8, line 266: "efficacious" Was the desired use related to efficacy or efficiency?

Page 9, line 277: "fewer markers" not "less markers" Markers can be counted

Page 9, line 291: Place "such as ... bootstrapping" between commas

Page 9, line 294 and 297: Substitute ";" for ","

Page 10, line 311: Maybe add some citations after "other phylogenomic studies"

Page 10, line 321: Use "distinct" instead of "differential"

Page 11, line 341-348: Rearrange/revise paragraph, the opening sentence is a bit jarring/confusing as it sounds like the authors established the subgenus Stelladerma etc. in the present study.

Page 11, Line 359: Maybe replace 'further corroborated by' with 'consistent with'

Decision letter (RSPB-2020-2102.R0)

09-Oct-2020

Dear Dr Chan:

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,
Dr Sasha Dall
mailto: proceedingsb@royalsociety.org

Associate Editor
Comments to Author:
Two experts in the field have reviewed your revised manuscript, and both agree that it is improved with respect to the previous version. However, reviewer #2 has identified several weakness in the methodological part, e.g. hidden paralogy and the AU-test that should be addressed. Regarding the AU-test, following the reviewer's comment, it is sufficient to carry out this test with the optimal alignment and testing all different retrieved topologies against this.
In conclusion, considering these comments and the methodological issues, I cannot recommend the MS for publication in its current status.

Reviewer(s)' Comments to Author:

Referee: 2

Comments to the Author(s).
This was a co-review of the revised manuscript. We were listed as Referee 2 of the original submission. Overall, we thought the authors did a great job of addressing the reviewer comments. Below we have provided a few minor suggestions for improving the main text prior to publication.

Multiple places in manuscript: When referring to statistical significance please use "non-significant" instead of "insignificant".

Page 2, line 36-37: "We showed ... internal branches." Please clarify, not clear whether causal relationship is between ILS and short branches, or discordance and short branches. Also, short branches result from ILS, not the other way around.

Page 2, line 53: Remove the word "support"

Page 3, line 100: Substitute "UCE's" by "UCEs".

Page 4, line 132: "each unfiltered dataset" maybe remove "unfiltered"

Page 4, line 133: "filtered at 75% sampling completeness", worth clarifying whether it is taxon sampling or length of sequence, wording isn't clear

Page 5, line 165: Consider substituting "(respectively)" for ", respectively,", based on the nearness to the use of parentheses introducing abbreviations and linking support value and software

Page 6, line 173-174: "indicate" and "indicating" within the same sentence, consider swapping one of them for a synonym

Page 6, line 179: Consider "we tested ... for the presence of polytomies" or just "we tested ... for polytomies"

Page 6, line 184: Switch "ancestor-and-descendent" for "ancestor-and-descendant". This persists in the Supplementary materials in section 1.3.

Page 7, line 235: "larger proportion" instead of "larger portion"?

Page 8, line 249 and 251: Add comma before "but"

Page 8, line 266: "efficacious" Was the desired use related to efficacy or efficiency?

Page 9, line 277: "fewer markers" not "less markers" Markers can be counted

Page 9, line 291: Place "such as ... bootstrapping" between commas

Page 9, line 294 and 297: Substitute ";" for ","

Page 10, line 311: Maybe add some citations after "other phylogenomic studies"

Page 10, line 321: Use "distinct" instead of "differential"

Page 11, line 341-348: Rearrange/revise paragraph, the opening sentence is a bit jarring/confusing as it sounds like the authors established the subgenus Stelladerma etc. in the present study.

Page 11, Line 359: Maybe replace 'further corroborated by' with 'consistent with'

Referee: 1

Comments to the Author(s).
Overall the manuscript has improved and it's more readable. The authors have made an effort to make it of more interest to a general audience. The explanation of the equal frequencies test in the context of the Multispecies Coalescent is more clear and now the results from these make sense. But, following from the issues raised before, even though they have tackled several of these, there are still 2 major ones that I do not feel the authors have satisfactorily answered or tackled in their paper (see below Hidden Paralogy and AU Test). I have also added 2 more (Data and Different Data and topologies) that won't take too much for the authors to address and I think it would enhance the paper and the discussion greatly.
As for the numerous minor observations, I also hope that the authors consider making these changes, which will help improve the quality and detail of their paper. I apologise in advance for the lengthy text.

Major observations:

- Hidden Paralogy:
In their response the authors claim that "Namely, the target markers were matched to the Nanorana parkeri and Xenopus genomes, where markers that matched to more than one location in each genome were initially removed altogether from the probe set. It is also possible that lineages may have lineage-specific paralogs not originally found, so during the bioinformatic processing of each sample, we removed any assembled contigs that matched substantially to more than one of the target markers".

Having only 1 match to a probe alignment hardly constitutes a convincing way to determine orthology (and I am not referring to in paralogs either). I will explain why:

As all Amphibians (indeed vertebrates) have undergone 2 whole rounds of genome duplication and multiple gene losses, for 3 species A, B and C, it would be expected that they will have 4 copies of a same locus/gene: A1, B1, C1, A2, B2, C2, A3, B3, C3, A4, B4, C4. Please see attached file with a figure explaining this. In this example, true orthologs will depict the following species tree: ((A+B)+C);
Now imagine the case where after gene losses, in particular cases of late gene loss, you only end up with copies of loci A1, B3, C3, which will appear in single copy in your sequencing (and you will think it is ortholog), but these will yield the topology: (A(B,C)); which is not the species tree topology. This is a case of hidden paralogy, and this is something that you cannot discard in your method, even using probes. This pattern, it would be very difficult to discern in your data, as it will look like ILS, when it is not.
Because of the taxonomic level at which you are working (inter Genera relationships within a Family) it is likely that most of your discordance issues are due to ILS and that hidden paralogy may be a small proportion of your data, but still you cannot discard hidden paralogy. This is an issue that is not unique to this study and it will take a long while to resolve.
Following from this explanation, I would feel more comfortable if the authors added an explicit caveat in their discussion that they cannot discard that there is hidden paralogy in their data in those sentences where they claim that the source of discordance is ILS. You can use as citations the references I gave before to support this (https://doi.org/10.1093/molbev/msz067 and https://doi.org/10.1371/journal.pone.0062892)

In addition to this, I invite the authors again to read those papers because some of these issues are explained there more clearly and it will also come in useful as they state somewhere in their FrogCap pipeline that they intend to include a code/software to identify orthologs in this type of data in the future.

- AU-Test and "optimal tree": Upon the suggestion that the authors perform an AU test, the response was: "We consider the correct topology to be the one derived from a total evidence approach; i.e., the largest and most comprehensive dataset (all-combined). In our opinion, a topology test would not help in this case, because the test is dependent on which alignment was used as a reference (intron, exon, UCEs etc.). This inherently introduces user selection bias. To avoid confusion, we have changed the terminology from “correct topology” to “optimal topology” whenever referring to our results"

I disagree with this explanation, in fact, choosing a "total evidence" as the "optimal" (which is the same as saying "best") topology is also subjective and biased. Why is T4 worse than T1? or T2, T3, T5? only because it doesn't match what they have chosen as their preferred topology? I still am of the opinion that the authors should carry out the AU test, or any hypothesis testing to discard that the alternative topologies are not equal descriptions of the data. Ideally, they should carry out an AU test for each alignment they tested against all topologies. This is something that can be carried out very simply in IQtree. At the minimum, they should carry out this test with their "optimal alignment" and testing all different retrieved topologies against this. This analysis would also add support to the authors' taxonomy suggestions at the end of the Discussion (section 4.4).

Bottomline is, unless the authors can prove that one topology is better than another through a test as explained before, using the terms "best", optimal", "worse", "suboptimal", "right", "wrong" is not correct and not supported.

- Data displayed and available: Although the authors have now included the supporting alignments and most of the resulting trees, they still need to include some that I was not able to find (namely the results of the SVDQuartet) and also it is my opinion that the paper would benefit enormously if the authors provided each of the resulting topologies in the Supplementary information (see comments for "Additional results below").

- Different data (exon, intron, UCE, Legacy) yield different topologies because these data are likely undergoing different selection pressures and it is something that has been already observed in phylogenomic studies. I think it warrants a mention in the discussion (I'd say it would go well in section 4.3 Causes of discordance). You should have a look at Shen et al. 2018 https://www.nature.com/articles/s41559-017-0126 and Bravo et al. 2019 https://peerj.com/articles/6399/ .

Minor observations:

- Trees deposited in the Data Dryad: I only found the trees for ASTRAL and IQtree, but couldn't find the SVDQuartet trees. This is either missing or the labels need to be made more clear.

- Methods and Results: 4 or 5 outgroups? In the main text for the methods it says 5 outgroups (which should be all fully listed in the main text). But then table S1 only has 4 outgroups listed. Also there is a major inconsistency with the outgroup species listed: Table S1: missing Scaphiophryne marmorata and Arthroleptis_variabilis. Also said table has Abavorana nazgul which doesn't appear anywhere in the alignments. Please correct this.
Also following from this point, you should add a sentence in the Methods main text of why you chose those 5 outgroups, and how distantly related are they from the ingroup.

- Materials and Methods: Bioinformatics: Line 146. Alignment information should be mentioned here in the main text, also add the software used.

- Lines 190-205: it's missing information on which datasets were Bootstrap carried out and needs to specify the bootstrap calculated (parametric Bootstrap or the Fast bootstrap which are different and both available in IQTree).

- Line 282-286: Where are the analyses that support the statements made? can you supply the distance between the gene trees and the "large" trees or how did you check for this. These statements need reference to the results from your analyses to back it. These should be supplied for both topology and branch lengths.

- Outgroups missing. I was surprised that I couldn't see any of the outgroups in any of the supplied figures. Figure 1a is the only tree you show and it does not depict the full tree of your results. In the spirit of being fully transparent with the information, it would be better to show the outgroups somewhere, such as resulting trees in the Supplementary information (see point "Additional results")

- Figure 1: Also the way the groups are summarised in T2, T3, T4 and T5 are very confusing and one has to spend ages to make sense of the comparisons and the positions of Nodes N1, N2, and N3 across all 5 topologies. I would suggest that the authors label N1-N3 in each of the 5 topologies.

- Table 2: Needs to add more relevant data to support why certain nodes from the different analyses fall in the anomaly zone. You state in the methods and in the Supplementary Methods that you performed "Anomaly Zone calculations". The results of these calculations and the branchlengths should be provided in this table.

- Fig S2: missing labels on the left hand for all vignettes. Also it would be good if you could supply a sentence or 2 in the caption for this to state what is the point you want to highlight, the major result from this.

- Line 380-381: "Quartet scores for UCE datasets were also higher compared to introns and exons, indicating that UCE markers may be less affected by ILS" add reference to results for this: Table 2.

- Additional results: Also I would like to ask the authors to provide a visualisation of the resulting trees for each of the 11 datasets and all the different methods in the supplementary information, as this will help any reader to compare the underlying results.I personally had to copy and paste each of them in a tree viewer, why make life harder to your reviewers when it is better to show all your results in an explicit manner? Please add in each tree a label of Nodes N1, N2 and N3 and which topology they result in (T1, T2, T3, T4 or T5). And also add the support values calculated for each branch based on whatever method used.

- Figures 1 and 3: Node 1: very confusing this "Stelladerma" appearing out of nowhere (not shown in Fig 1, not appearing in any alignments) until you reach the end of the paper and even then, it's not clear the clade of Theladerma that the authors refer to. Anybody who is not familiarised with Rhacophorid or even frog taxonomy are going to find this very confusing. Names for clades and species should be consistent in all figures.

- Figure 3: I noticed that you labeled some of the topologies T1, T2, T3 in some cases next to some of the alternative topologies. But nowhere in these appear T4 ot T5. Is this because of an error? I actually think it would help to illustrate some of your points in the discussion if you labeled in this figure which topology matches which of these hypotheses (for example take Line 425-426 you mention at Node 1 you retrieve T4. Label T4 in the hypothesis that corresponds here for this node in this figure). I think you should do these for all Topologies (T1-T5).

- Phylogenomic relationships (4.4): I find most of this whole section confusing and difficult to follow because the authors do not point to which parts of the trees they refer to (Node 1, Node 2, Node 3).
Also, can you name/find any other supporting evidence for your proposed hypotheses? like morphology, ecology, etc in disfavour of the "Stelladerma hypothesis". It would also help that you define in a very brief way what is this hypothesis.

- Lines 391-392: "Bootstrap values were not correlated with topological concordance and routinely produced strong support for highly discordant nodes (Fig. 2). " You should add this is in agreement with what has already been explain in Minh et al 2020 (https://doi.org/10.1093/molbev/msaa106).

- Line 407: "However, higher concordance values can also be an artefact of small datasets" same as previous observation, this is exactly what was explained in Minh et al 2020 (https://doi.org/10.1093/molbev/msaa106), so it would be good to add this as reference here.

Author's Response to Decision Letter for (RSPB-2020-2102.R0)

See Appendix B.

Decision letter (RSPB-2020-2102.R1)

13-Nov-2020

Dear Dr Chan

I am pleased to inform you that your manuscript entitled "Target-capture phylogenomics provide insights on gene and species tree discordances in Old World Treefrogs (Anura: Rhacophoridae)" 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

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 10 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,
Dr Sasha Dall
Editor, Proceedings B
mailto: proceedingsb@royalsociety.org

Associate Editor:
Comments to Author:
Dear Dr chan:

Following your response to the reviewers, I am glad to recommend your manuscript for publication on Proceeding B.
Best wishes,
Roberto Feuda

Appendix A

Response to reviewers (Manuscript ID RSPB-2020-1438)

Dear Editor,

We thank you and the reviewers for providing valuable insights and comments that have significantly improved our manuscript. We have carefully considered and addressed every comment, which are documented in detail below (our response are in blue font beginning with >). All relevant files generated from this study (alignments, gene trees, species trees, partition files) have been uploaded to Dryad (link provided) and raw sequences will be uploaded to GenBank SRA should this manuscript be accepted (a BioProject has already been created). We have included permitting information in the Acknowledgements corresponding to KU specimens collected by us. Some information on permits is not available because some samples were included in our study via tissue samples borrowed from other museums. Please do not hesitate to contact me (corresponding author's email: chankinonn@gmail.com) if you have any questions.

Reviewer(s)' Comments to Author:

Referee: 1

Comments to the Author(s)
The authors sought to investigate the sources of discordance in phylogenomics using a combination of different types of molecular data, from high-throughput to the inclusion of a small number of traditional markers. I liked the intention, especially because it tried to address a group of anurans with historical taxonomic problems. However, there are serious issues about the analysis and the assumptions made in this paper, which need to be addressed in order for their claims to be supported.

Major observations:

1) The authors dismissed hidden paralogy, which is a fundamental issue and a potential source of discordance in their data. Hence one of the major claims in the paper: that the discordance observed is solely due to ILS is not entirely correct unless they prove that their dataset does not have hidden paralogy. I encourage the authors to consult some papers which have investigated this issue:
- Siu-Ting et al. 2019. Inadvertent Paralog Inclusion Drives Artifactual Topologies and Timetree Estimates in Phylogenomics. Molecular Biology and Evolution. https://doi.org/10.1093/molbev/msz067

- Struck, T. 2013. The Impact of Paralogy on Phylogenomic Studies – A Case Study on Annelid Relationships. PLOS ONE. https://doi.org/10.1371/journal.pone.0062892

In order to make the paper more robust, the authors should show how they identified orthologs and paralogs and check for these in their data. Siu-Ting et al. 2019 presents a method to investigate this issue in gene trees, which could be useful and a rapid way to evaluate paralogous gene trees.

> Thank you for bringing this to our attention. Paralogy detection and removal in our study is robust, and was stated in the supplementary material and references therein. Paralogy detection is accomplished during the probe design and bioinformatic stages. Namely, the target markers were matched to the *Nanorana parkeri* and *Xenopus* genomes, where markers that matched to more than one location in each genome were initially removed altogether from the probe set. It is also possible that lineages may have lineage-specific paralogs not originally found, so during the bioinformatic processing of each sample, we removed any assembled contigs that matched substantially to more than one of the target markers. More details are in the original publication of the FrogCap method (Hutter et al., 2019), but we have also included a brief description in the revision for added clarification and to emphasize the added-value component of this novel method for this work (and demonstrate how we alleviated paralogy concerns here).

2) Carry out an estimation of saturation for all gene trees - this is key in particular when dealing with nucleotides rather than amino acids. I would like to see if this is or not an issue for their datasets. Following from this, it would be interesting if the authors had explored if carrying out a phylogenetic inference on the amino acid alignment for the exon dataset would retrieve similar results.

>As suggested, we performed a saturation analysis using the program DAMBE. Because analysing more than 12k loci individually was not feasible, we followed the approach by Breinholt & Kawahara (2013) by conducting the saturation test on the concatenated exon dataset. We assessed substitution saturation for the 1st, 2nd and 3rd codon positions separately. Test results showed that saturation was insignificant at all codon positions and this was confirmed by saturation plots (all results are provided in supplementary material). We also performed phylogenetic inference on the amino acid alignment for the exon dataset as suggested. The resulting topology was similar to the topology from other IQ-TREE exon datasets (topology T3), indicating that saturation did not have a significant impact on phylogenetic inference.

3) Perform a partitioned analysis for concatenated datasets using the models inferred for those gene trees obtained. IQ-Tree allows for specification of multiple partitions with very large datasets when run using multiple processors (such as those in High Performance Clusters).

>We tried running a partitioned + model testing IQ-TREE analysis, but this was not computationally tractable for the largest, unfiltered datasets. The problem was not so much computational power but available memory, which is the limiting factor in IQ-TREE model testing. Our analysis on the unfiltered datasets consistently ran out of memory even when 100 GB of RAM was allocated. As a compromise, we conducted a partitioned analysis using the most parameter rich GTR+G model for all partitions. For large phylogenomic datasets, this strategy has recently been shown to be as, if not more, effective than conducting model testing for individual loci (Abadi, Azouri, Pupko, & Mayrose, 2019; Chan, Hutter, Wood, Grismer, & Brown, 2020). The resulting topologies were identical to those performed under unpartitioned analyses and there were insignificant differences in branch lengths and branch support, further reinforcing our hypothesis that incongruence in Rhacophoridae is due to incomplete lineage sorting and not systematic/methodological bias. Results and Methods have been updated accordingly.

4) The authors make statements such as "correct topology" and "wrong topology". What was

the criteria used to determine which is the right and the wrong topology?
My suggestion is to carry out AU tests for all topologies and all datasets to discard that the other topologies retrieved are not equally good descriptions of the data. I think this is particularly important to support such statements.

>We thank the reviewer for pointing this out. We consider the correct topology to be the one derived from a total evidence approach; i.e., the largest and most comprehensive dataset (all-combined). In our opinion, a topology test would not help in this case, because the test is dependent on which alignment was used as a reference (intron, exon, UCEs etc.). This inherently introduces user selection bias. To avoid confusion, we have changed the terminology from "correct topology" to "optimal topology" whenever referring to our results and explicitly stated that:

"*Using the All-combined dataset as a proxy for a total evidence approach, we consider T1 to be the optimal topology*." (Last line of section 3.2)

5) The chi-square test used by the authors "We then used a chi-square test to determine whether the frequency of gene trees (gCF) and sites (sCF) supporting the two alternate topologies was significantly different. Insignificant p-values (p > 0.05) indicate a failure to reject the hypothesis of equal frequencies, indicating that discordance among gene trees and/or sites is likely due to ILS."
It seems to me a very large leap from topologies having all equal frequencies inferring that this is due to ILS. It's an assumption that is not clear.

>We acknowledge that this test does indeed require assumptions to be made (as of most tests). Unfortunately, to the best of our knowledge, this is the only method that can statistically test this hypothesis. We have provided additional references that use a similar approach, including a recent paper in *Systematic Biology* (Burbrink et al., 2020; Green et al., 2010; Huson, Klopper, Lockhart, & Steel, 2005).

6) I feel that the authors have missed on the opportunity to discuss further on types of data and how these can improve phylogenomic resolution. I think this would be one of the most informative and helpful things for researchers doing these type of analyses to know. But instead, only a couple of short sentences in the discussion were given for this. I would like to see the authors explore the impact of the types of data, link this to phylogenetic informativeness, to topology support and if resolution is actually improved or not and what could be happening. In my opinion devoting more space to discussing this would be more appropriate for a general audience, and will distinguish it from a paper that seems to be more focused for amphibian taxonomists.

>We thank the reviewer for this suggestion and have expanded and re-structured our discussion on the different types of data/analysis in relation to phylogenetic inference, branch support, and overall concordance. We created a new subheading for this (section 4.1) and positioned the more general and widely applicable discussions up front (sections 4.1 and 4.2). Consequently, the more taxon-related discussions have been moved down (4.3 and 4.4).

Minor observations:
- Line 64-67: "Discordance between gene trees and the species tree can also result from biological processes such as introgression, horizontal gene transfer, and incomplete lineage sorting [35,38–44]". The authors have forgotten to list here one of the major sources of

discordance is gene duplication and gene loss. In particular, failing to identify orthologous from paralogous genes will affect phylogenetic estimation greatly. See point 1) of major observations.

>Added gene duplication and loss to the list and also added a reference to it: (Hahn, 2007). Additionally, we explain above how we addressed this possibility.

- Line 178-179: Authors assume that the only source of incongruence that they have is ILS without discarding other sources of incongruence. See point 1) of major observations.

>This was not meant to be an assumption, but rather, a hypothesis based on preliminary data exploration. We are aware that discordance can arise from many sources, but exhaustively testing for each of them would not fit into the scope of a single study, nor do we find it necessary. It is normal for large studies to conduct preliminary testing to refine the hypothesis-testing framework, which is what we have done here. Our exploration of tree discordance and gene tree frequencies indicated that ILS could be a factor, hence, we proceeded to perform more robust analyses to test that hypothesis, which our results clearly support.

- Lines 209-210: specify units, did you mean base pairs for numbers supplied?

>Added the units (base pairs)

- Line 221: congruent: do you mean that all topologies were identical here? Following from this, how do you reconcile here which is your optimal tree?

>Yes, congruent=identical topologies (slight differences in branch lengths). Added a clarifying statement at the end of the paragraph:

"*Using the All-combined dataset as a proxy for a total evidence approach, we consider T1 to be the optimal topology.*"

- Line 272: not clear what is the test carried out: what is the null hypothesis and what is meant by equal gene tree frequencies. Needs to be more clear in the main text.

>Added more details to explain this test

line 285: "anomalous gene trees" - can you be more specific, what do you mean by anomalous in this case?

>Provided a definition:

"*most probable gene trees that do not match the underlying species tree*"

- Lines 296-299: "Through systematic analysis of different classes of data, we were able to demonstrate that ILS caused by rapid diversification events was responsible for most of the deep-level discordances in Old World treefrogs." I don't think this is true since authors have not discarded other sources of discordance. See major observation point 1).

>We have changed the tone of that statement to reflect the fact that our results support the **hypothesis** that ILS is **likely** responsible for the discordances:

"*Overall, our systematic analyses of different classes of data support that hypothesis that ILS (caused by rapid diversification events) is likely the main underlying factor responsible for most of the deep-level discordances in Old World treefrogs*".

- Missing more details in the methods: especially in the design of the bates used. At least provide a short summary.

>Added additional details (see response above)

- Other details missing in the Supplementary Materials Methods: Needs to provide more details about parameters and thresholds used for it to be fully reproducible. In particular in the matching against probe reference, since this point is key for the identification of orthologs and paralogs.

>Added additional details (see response above)

- What constituted the "Legacy" dataset? from what publications? include references.

>The Legacy dataset consists of nuclear loci commonly used in phylogenetic studies of amphibians. They were not selected from specific publications, but were curated based on our experience and their prevalence on GenBank. The list of loci can be obtained from the original paper describing the FrogCap method (Hutter et al., 2019) and we have provided clarification in the Methods section. Additionally, all loci used here were also used in an influential and highly cited phylogenomics study of Gondwanan frogs (Feng et al., 2017) (although not all Feng et al. 2017 loci were used in this study).

- In Methods: specify software or code used to calculate gCF and sCF.

>Software and code have been specified:

"*Concordance factors were calculated in IQ-TREE* (Minh, Hahn, & Lanfear, 2020) *and the chi-square test was performed in R using scripts available here: http://www.robertlanfear.com/blog/files/concordance_factors.html*"

- Provide more details about software used to determine phylogenetic informative sites.

>Done:

"*Sampling completeness and PIS were calculated using the summary function in the program AMAS* (Borowiec, 2016)."

- Bibliography has several spelling mistakes which need to be double-checked and addressed.

>Checked and corrected.

- Figure 2 B seem a bit pointless. In Fig. 2B you are comparing values bootstrap values that are all 100 against branchlengths, it is obvious that you will not be able to see a correlation. I don't see the point of this graph.

>That is precisely what we are trying to depict. Although it may look slightly awkward, we feel that it drives home one of the main points of the paper, i.e. bootstrap support values from concatenation methods are poor and misleading indicators of branch support in phylogenomic datasets. This is a very important point to drive home because most, if not all phylogenomic studies still use bootstrapping as a measure of "confidence" for the "right" topology (Chan et al., 2020).

- Figure 3 caption needs to be more informative. What filtering was used for each type of dataset?

>There is a legend at the bottom of the figure explaining how colors correspond to different filtered datasets. In any case, we provided additional explanations in the caption.

- Supplementary Materials font seems to be different on the second half of the document. was that supposed to be like that?

>Corrected the font.

- All alignments and gene trees inferred should be provided in a repository to ensure results will be replicable.
>As requested, all alignments, gene trees, and consensus/species trees have been uploaded to Dryad. The link has been provided in the Data Accessibility section

Referee: 2

Comments to the Author(s)
RSPB 2020-1438

This was a co-review and overall we both enjoyed reading this manuscript. We thought that it used a large, novel dataset to demonstrate some important qualities of how 'short and deep' branches relate to gene tree discordance and bootstrap proportions. We both feel that it is a manuscript with potential broad appeal to the readership of Proceedings B. However, we thought several key analyses should be reconsidered (particularly the direct comparison of gene trees estimated using different model selection criteria) and in several instances that the methods lacked explanatory detail. We have listed below some major and minor comments that we hope will be helpful in constructing a revised version of the manuscript prior to publication.

Major comments:

1. Different introns can be under different selection pressures (e.g. intron with regulatory regions/intron without, selfsplicing intron/non-selfsplicing intron, etc.), might this affect the results? Would it be worth trying to ascertain whether the introns used can be binned into neutral/positive/negative/stabilising selection? This is not a request for the authors to do this, but it is worth considering the potential implications to the results.

>This is a great suggestion, but we feel like it would be beyond the scope of this paper. Characterizations of introns in a phylogenetic context have been done for bacteria (Toro & Martínez-Abarca, 2013), but we are not aware of the implications on vertebrates. Additionally, this level of annotation information on introns in frogs is not well-known.

2. Were non-Sanger and non-UCE datasets checked for paralogs? If these are present they can result in ILS-like behaviour.

>Addressed in response to first reviewer, above

3. The authors' worry about computational intractability is understandable, but means that the results of the concatenated analysis are not comparable to the gene tree analyses under the estimated models: The concatenated analysis was run under the GTR+G model with an unpartitioned alignment (why not GTR+I+G?); gene trees are necessarily "partitioned", and they were subjected to model testing. We suggest running all gene trees under the GTR+G model, and compare the results to the gene trees obtained under the selected models. Alternatively, rate heterogenous models can be used in lieu of the more commonly used models and would preclude the need for model testing. The concatenated tree should also be run under gene partitions. If there are no significant differences, then the comparison between the concatenated and gene tree analyses stands. If not, the conclusions will need to be re-interpreted and revised.

>As suggested, we have performed additional ASTRAL analyses using gene trees estimated with GTR+GAMMA and the resulting topologies were similar to those derived from gene trees estimated via model testing. This is in agreement with previous studies that have shown that model testing is not mandatory and has insignificant effects on species tree inference (Abadi et al., 2019; Chan et al., 2020). This is an important point that will be widely applicable to many readers, hence, we have emphasized it in the Discussion (section 4.1).

We did not use GTR+I+G because combining invariant sites with a discretized gamma distribution (I+G) can potentially result in non-identifiability (Borges & Kosiol, 2020; Chai & Houswörth, 2011). IQ-TREE has developed an analytical workaround (Nguyen, Von Haeseler, & Minh, 2018). However, theorical proofs have yet to be developed. In any case, we have performed a variety of analyses using different parameter settings and none of them have changed the resulting topology. Hence, we do not expect that including invariant sites in the substitution model will do so.

4. Were quartet scores checked in the Astral trees, or just local PP?

>The DiscoVista analysis (Fig. 4) uses branch quartet scores to calculate gene tree frequencies, hence, this has been accounted for. We have added the maximum normalized species tree quartet score to Table 2. This provides a rough idea of overall discordance among gene trees. One important pattern emerged: filtering by taxon completeness does not improve quartet score, but filtering by PIS does, indicating that the former strategy does not improve congruence, but the latter does. We have added this to the Discussion.

5. Which IQTree output was used as input for ASTRAL? Binary best ML tree, greedy consensus or 50% majority-rule consensus? The level of input tree resolution affects the resolution of the tree summarised by ASTRAL. All binary trees can lead to erroneously well resolved trees, while the 50%MRC can lead to more conservative estimates of topology with

the output tree being less resolved than the "true" tree. Check "5.1.1 Gene tree uncertainty" in Mirarab 2019 arXiv:1904.03826v2

>The best ML tree was used, which was suggested by the authors of the program to be the most straightforward choice. This has been added to the Materials and Methods (section 2.4)

6. While an explanation about the anomaly zone is given in the supplementary materials a bit more detail would be helpful to understand the manuscript. At the very least, the assumptions of the approach should be stated explicitly, and how they may affect the results discussed.

>We have decided to move some of the information from supplementary material to the main text as added clarification on this analysis.

7. Given that the paper describing concordance factors has only recently been published, it would be helpful to have more a more detailed description of the method in the manuscript. It is quite likely that people that do not regularly use IQTree are not aware of concordance factors, even if the preprint has been around for at least two years.

>We have provided additional details on this analysis. The paper describing concordance factors was published this year in MBE and we have updated that reference accordingly.

8. Parsimony Uninformative and Constant Sites contribute to branch lengths, but not topology. Thus, it makes sense that their removal would not affect site concordance factors. Might be worth saying explicitly how the BLs changed between the PIS filtered and unfiltered analyses. Would make for easier reading.

>Our rationale for removing markers with low PIS was not to remove uninformative sites per se, but rather, to remove markers with low numbers of PIS as these markers could produce potentially produce "noise" due to insufficient phylogenetic signal. This was done to determine whether these noisy markers can adversely affect phylogenetic inference.

Minor comments:

Page 2, line 49: "...high bootstrap support values..." would be better as "... high bootstrap proportions...", previous clause already mentions branch support

>Changed "values" to "proportions"

Page 3, Line 67: "...ancestral polymorphisms fail..." would be clearer as "...ancestral polymorphisms don't..."

>Changed "fail" to "don't"

Page 3, Line 68: "...failure of lineages to coalesce..." is confusing. Coalescence and lineage sorting might be considered different temporal concepts; one backwards in time (coalescence) the other forwards in time (lineage sorting). Please clarify.

>Removed the confusing part of the sentence as it is not really necessary. We have made those points more concise and clear:

“*Incomplete lineage sorting (ILS) occurs when ancestral polymorphisms do not reach fixation between successive divergence events. This can occur during periods of rapid diversification, particularly when effective population size is large relative to its associated branch length*”

Page 4, line 116: "i.e." should be substituted for ":", "i.e." is used to explain the same information with different wording, not to signify enumeration
>Changed “i.e”. to “:”

Page 4, line 118: "...holdings of University..." reads better as "...holdings of the University..."

>Added “the” as suggested

Page 5, line 138/9: "Capella-gutiérrez" should be "Capella-Gutiérrez". Also should it be referenced as a number in Proc B style?
>Corrected

Page 5, line 139/140: consider using "...at leat 50% of..." instead of "...at least 50 percent of..."

>Corrected

Page 5, line 145: "...missing taxon representation and..." reads better as "...missing taxa and..."
>Changed to “missing taxa” as suggested

Page 5, line 160: ASTRAL-III was not cited in the manuscript
>Added missing reference

Page 6, line 185: Does "polytomy" refer to a hard or soft polytomy? Is the polytomy due to uncertainty in the data (soft), or is the "true" tree non-binary (hard)?

>We believe that the program does not differentiate between soft and hard polytomies. It merely detects zero branch lengths.

Page 6, line 188: "...and will generate..." would be better as "...and can generate/yield...", even though the branches are short there is still p=1/3 that the correct topology is recovered

>Changed “will” to “can”

Page 6, line 190/191: "...ancestor-and-descendent..." should be "...ancestor-and-descendant...", this is also present in the supplementary materials

>Corrected

Page 9, line 295: "...soft polytomy...", How was hard vs soft ascertained? Links to Page 6, line 185
>We thank you for pointing this out. At this point, we are unable to determine whether it was a hard or soft polytomy. We added this sentence for clarification:

"*However, we were unable to determine whether the node constituted a hard or soft polytomy*"

Page 9, line 296-299: "Through systematic ... treefrogs." should be re-worded to show that it is a hypothesis. Until time machines are available the causes of ILS in frogs will probably remain hypothetical. Also the datasets were not tested for the potential causes of ILS, only whether there was ILS.

>We have re-written this sentence to reflect that uncertainty (also see our response to a similar comment by Reviewer 1)

Page 10, line 326-329: "This shows ... before." To our current understanding of the manuscript, the possibility of hard polytomies was not discarded, so the lack of resolution at short internodes may not be artefactual. Also, we are aware of several other studies that have demonstrated short internodes remain challenging with large 'phylogenomic' datasets (e.g. Streicher et al. 2016. Syst. Biol. 65: 128-145; Burbrink et al. 2020 Syst. Biol. 69: 502-520; Roycroft et al. Syst. Biol. 69: 431-444; Singhal et al. Syst. Biol. In Press), so it would be appropriate to indicate via citations that this is a wider discussion in the literature at the moment.

>Thank you for pointing this out and referring us to the additional references. We have added them to the discussion.

Page 10, line 330: "total evidence" While multiple datasets/types were used, we do not feel that the term total evidence is warranted. There were no morphological data used, and there was only a passing reference to life-history traits.

>We use total evidence in the context of this study, which refers to a combined analysis of all available data. This is a reasonable and accepted use of the term (Lecointre & Deleporte, 2005). This has been added to the section 3.2. Total evidence can never refer to "all" bodies of evidence, because that is in essence, innumerable. Adding morphological data or life-history traits only adds two additional lines of evidence, which can hardly be considered "total" as well.

Page 11, line 357: "... Pyron and Wiens [2001]) and ..." should be "... Pyron and Wiens [2001] and ...", there is an extra ")"

>We have re-formatted that entire section so that citations are in ProcB format.

Page 12, line 376: "... funded NSF..." reads better as "...funded by NSF..."
>Corrected

Figure 1: Adding a cladogram for topology T1 would make comparison to the other cladograms easier. Suggest keeping phylogram as is and rearrange cladograms to show all 5 topologies.

>We do not feel that this is necessary as it will be redundant with the phylogram. We have

clearly delineated genera in the phylogram so that it is easy to read and compare with the other cladograms.

Figure 4: Might be easier to interpret if each topology has a different colour globally, rather than just each topology with each node.

>We actually feel that isolating each node makes it easier to interpret. Having full phylogenies will actually introduce a lot of unnecessary noise to the figure.

Also changing the y-axis label to "Cumulative percentage" might make the graph easier to interpret.

>Done

Citations: In-text citation style is inconsistent, some have Author (date), Author [date], Author date or [#]. Under the journal's style they should be: Author [#] at the start of a sentence, or [#] in all other instances. Also, update Minh BQ, Hahn MW, Lanfear R. 2018 New methods to calculate concordance factors for phylogenomic datasets. bioRxiv , doi: http://dx.doi.org/10.1101/487801. It has now been published.

>Done

Literature cited

Abadi, S., Azouri, D., Pupko, T., & Mayrose, I. (2019). Model selection may not be a mandatory step for phylogeny reconstruction. *Nature Communications*, *10*(1), 934. doi: 10.1038/s41467-019-08822-w

Borges, R., & Kosiol, C. (2020). Consistency and identifiability of the polymorphism-aware phylogenetic models. *Journal of Theoretical Biology*, *486*, 1–6. doi: 10.1016/j.jtbi.2019.110074

Borowiec, M. L. (2016). AMAS: a fast tool for alignment manipulation and computing of summary statistics. *PeerJ*, *4*, e1660. doi: 10.7717/peerj.1660

Breinholt, J. W., & Kawahara, A. Y. (2013). Phylotranscriptomics: Saturated third codon positions radically influence the estimation of trees based on next-gen data. *Genome Biology and Evolution*, *5*(11), 2082–2092. doi: 10.1093/gbe/evt157

Burbrink, F. T., Grazziotin, F. G., Pyron, A. R., Cundall, D., Donnellan, S., Irish, F., … Zaher, H. (2020). Interrogating genomic-scale data for squamata (lizards, snakes, and amphisbaenians) shows no support for key traditional morphological relationships. *Systematic Biology*, *69*(3), 502–520. doi: 10.1093/sysbio/syz062

Chai, J., & Housworth, E. A. (2011). On Rogers’ proof of identifiability for the GTR + Γ + i model. *Systematic Biology*, *60*(5), 713–718. doi: 10.1093/sysbio/syr023

Chan, K. O., Hutter, C. R., Wood, P. L., Grismer, L. L., & Brown, R. M. (2020). Larger, unfiltered datasets are more effective at resolving phylogenetic conflict: Introns, exons, and UCEs resolve ambiguities in Golden-backed frogs (Anura: Ranidae; genus *Hylarana*). *Molecular Phylogenetics and Evolution*, *151*, 106899. doi:

10.1016/j.ympev.2020.106899

Feng, Y.-J., Blackburn, D. C., Liang, D., Hillis, D. M., Wake, D. B., Cannatella, D. C., & Zhang, P. (2017). Phylogenomics reveals rapid, simultaneous diversification of three major clades of Gondwanan frogs at the Cretaceous–Paleogene boundary. *Proceedings of the National Academy of Sciences*, 201704632. doi: 10.1073/PNAS.1704632114

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., … Pääbo, S. (2010). A draft sequence of the neandertal genome. *Science*, *328*(5979), 710–722. doi: 10.1126/science.1188021

Hahn, M. W. (2007). Bias in phylogenetic tree reconciliation methods: Implications for vertebrate genome evolution. *Genome Biology*, *8*, R141. doi: 10.1186/gb-2007-8-7-r141

Huson, D. H., Klopper, T., Lockhart, P. J., & Steel, M. A. (2005). Reconstruction of reticulate networks from gene trees. In *Proceedings of the Ninth International Conference on Research in Computational Molecular Biology* (pp. 233–249). Heidelberg: Springer.

Hutter, C. R., Cobb, K. A., Portik, D. M., Travers, S. L., Wood, P. L., & Brown, R. M. (2019). FrogCap: A modular sequence capture probe set for phylogenomics and population genetics for all frogs, assessed across multiple phylogenetic scales. *BioRxiv*, *825307*. doi: 10.1101/825307

Lecointre, G., & Deleporte, P. (2005). Total evidence requires exclusion of phylogenetically misleading data. *Zoologica Scripta*, *34*(1), 101–117. doi: 10.1111/j.1463-6409.2005.00168.x

Minh, B. Q., Hahn, M. W., & Lanfear, R. (2020). New Methods to Calculate Concordance Factors for Phylogenomic Datasets. *Molecular Biology and Evolution*, 1–7. doi: 10.1093/molbev/msaa106

Nguyen, L. T., Von Haeseler, A., & Minh, B. Q. (2018). Complex models of sequence evolution require accurate estimators as exemplified with the invariable site plus gamma model. *Systematic Biology*, *67*(3), 552–558. doi: 10.1093/sysbio/syx092

Toro, N., & Martínez-Abarca, F. (2013). Comprehensive Phylogenetic Analysis of Bacterial Group II Intron-Encoded ORFs Lacking the DNA Endonuclease Domain Reveals New Varieties. *PLoS ONE*, *8*(1), 1–7. doi: 10.1371/journal.pone.0055102

Appendix B

Dear Editor,

We have carefully considered all the comments and addressed them to the best of our abilities. In particular, we have performed tree topology tests to provide statistical support for what we consider to be the optimal tree topology. The results and methods have been updated accordingly and additional details regarding this analysis are in Supplementary Material. In addition to mentioning the caveats of hidden paralogy in the Discussion, we also took the initiative and performed additional analyses to further investigate the possibility of hidden paralogy. Our results strongly indicate that hidden paralogy is unlikely to be present. However, as correctly pointed out by Reviewer 2, we were not able to completely rule this out without complete genomes. Results and discussion on this analysis have been added to the main text and details on the methodology have been added to Supplementary Material. In addition, all minor comments by both reviewers have been addressed. Included in our revision is the main text with all changes tracked. Below is our point-by-point response to each of their comments (our response in blue).

Associate Editor
Comments to Author:
Two experts in the field have reviewed your revised manuscript, and both agree that it is improved with respect to the previous version. However, reviewer #2 has identified several weakness in the methodological part, e.g. hidden paralogy and the AU-test that should be addressed. Regarding the AU-test, following the reviewer's comment, it is sufficient to carry out this test with the optimal alignment and testing all different retrieved topologies against this.
In conclusion, considering these comments and the methodological issues, I cannot recommend the MS for publication in its current status.

Reviewer(s)' Comments to Author:

Referee: 2

Comments to the Author(s).
This was a co-review of the revised manuscript. We were listed as Referee 2 of the original submission. Overall, we thought the authors did a great job of addressing the reviewer comments. Below we have provided a few minor suggestions for improving the main text prior to publication.

Multiple places in manuscript: When referring to statistical significance please use "non-significant" instead of "insignificant".

As suggested, we have changed insignificant to non-significant whenever referring to statistical tests. This was done in section 2.4 and 3.2

Page 2, line 36-37: "We showed ... internal branches." Please clarify, not clear whether causal relationship is between ILS and short branches, or discordance and short branches. Also, short branches result from ILS, not the other way around.

We have changed that sentence to:

“We showed that incomplete lineage sorting was detected at all nodes that exhibited high levels of discordance, which also resulted in those nodes having extremely short internal branches”

The first clause makes it clear that the relationship is between ILS and discordance, whereas the second clause indicates that the short branches result from ILS

Page 2, line 53: Remove the word "support"
Removed

Page 3, line 100: Substitute "UCE's" by "UCEs".
Changed to UCEs

Page 4, line 132: "each unfiltered dataset" maybe remove "unfiltered"
Removed “unfiltered”

Page 4, line 133: "filtered at 75% sampling completeness", worth clarifying whether it is taxon sampling or length of sequence, wording isn't clear

Changed to “filtered at 75% taxon sampling completeness” for added clarity.

Page 5, line 165: Consider substituting "(respectively)" for ", respectively,", based on the nearness to the use of parentheses introducing abbreviations and linking support value and software
Changed "(respectively)" to ", respectively,"

Page 6, line 173-174: "indicate" and "indicating" within the same sentence, consider swapping one of them for a synonym

Changed “indicating” to “implying”

Page 6, line 179: Consider "we tested ... for the presence of polytomies" or just "we tested ... for polytomies"
Changed to "we tested ... for polytomies"

Page 6, line 184: Switch "ancestor-and-descendent" for "ancestor-and-descendant". This persists in the Supplementary materials in section 1.3.
Changed to descendant

Page 7, line 235: "larger proportion" instead of "larger portion"?
Changed to proportion

Page 8, line 249 and 251: Add comma before "but"
Added comma

Page 8, line 266: "efficacious" Was the desired use related to efficacy or efficiency?
Changed "efficacious" to "viable" to avoid confusion

Page 9, line 277: "fewer markers" not "less markers" Markers can be counted
Changed to "fewer"

Page 9, line 291: Place "such as ... bootstrapping" between commas
Added commas

Page 9, line 294 and 297: Substitute ";" for ","

Changed to comma

Page 10, line 311: Maybe add some citations after "other phylogenomic studies"
Citations are provided at the end of that sentence.

Page 10, line 321: Use "distinct" instead of "differential"

Changed to "distinct"

Page 11, line 341-348: Rearrange/revise paragraph, the opening sentence is a bit jarring/confusing as it sounds like the authors established the subgenus Stelladerma etc. in the present study.

Upon further consideration, we have decided to remove section 4.4. This entire section is very taxon-centric and will only be of interest to a small fraction of readers who work on this group. Furthermore, it detracts from the main goals/theme of the paper, which is to elucidate sources of discordance, not resolve taxonomic uncertainty. We found that this section is mostly a summary of the results and thus, does not need a dedicated Discussion. Removing it will also prevent the paper from exceeding the page limit.

Page 11, Line 359: Maybe replace 'further corroborated by' with 'consistent with'
Changed to "consistent with"

Referee: 1

Comments to the Author(s).
Overall the manuscript has improved and it's more readable. The authors have made an effort to make it of more interest to a general audience. The explanation of the equal frequencies test in the context of the Multispecies Coalescent is more clear and now the results from these make sense.
But, following from the issues raised before, even though they have tackled several of these, there are still 2 major ones that I do not feel the authors have satisfactorily answered or tackled in their paper (see below Hidden Paralogy and AU Test). I have also added 2 more (Data and Different Data and topologies) that won't take too much for the authors to address and I think it would enhance the paper and the discussion greatly.
As for the numerous minor observations, I also hope that the authors consider making these changes, which will help improve the quality and detail of their paper. I apologise in advance for the lengthy text.

Major observations:

- Hidden Paralogy:
In their response the authors claim that "Namely, the target markers were matched to the Nanorana parkeri and Xenopus genomes, where markers that matched to more than one location in each genome were initially removed altogether from the probe set. It is also possible that lineages may have lineage-specific paralogs not originally found, so during the bioinformatic processing of each sample, we removed any assembled contigs that matched substantially to more than one of the target markers".

Having only 1 match to a probe alignment hardly constitutes a convincing way to determine orthology (and I am not referring to in paralogs either). I will explain why:

As all Amphibians (indeed vertebrates) have undergone 2 whole rounds of genome duplication and multiple gene losses, for 3 species A, B and C, it would be expected that they will have 4 copies of a same locus/gene: A1, B1, C1, A2, B2, C2, A3, B3, C3, A4, B4, C4. Please see attached file with a figure explaining this. In this example, true orthologs will depict the following species tree: ((A+B)+C);
Now imagine the case where after gene losses, in particular cases of late gene loss, you only end up with copies of loci A1, B3, C3, which will appear in single copy in your sequencing (and you will think it is ortholog), but these will yield the topology: (A(B,C)); which is not the species tree topology. This is a case of hidden paralogy, and this is something that you cannot discard in your method, even using probes. This pattern, it would be very difficult to discern in your data, as it will look like ILS, when it is not.
Because of the taxonomic level at which you are working (inter Genera relationships within a Family) it is likely that most of your discordance issues are due to ILS and that hidden paralogy may be a small proportion of your data, but still you cannot discard hidden paralogy. This is an issue that is not unique to this study and it will take a long while to resolve.
Following from this explanation, I would feel more comfortable if the authors added an explicit caveat in their discussion that they cannot discard that there is hidden paralogy in their data in those sentences where they claim that the source of discordance is ILS. You can

use as citations the references I gave before to support this (https://doi.org/10.1093/molbev/msz067 and https://doi.org/10.1371/journal.pone.0062892)

In addition to this, I invite the authors again to read those papers because some of these issues are explained there more clearly and it will also come in useful as they state somewhere in their FrogCap pipeline that they intend to include a code/software to identify orthologs in this type of data in the future.

We thank the reviewer for raising this issue. To address this, we have performed additional analyses to detect hidden paralogy (details and results in Supplementary Material section 1.4 and Table S4). This is summarized and discussed in section 4.3 as requested by the reviewer.

- AU-Test and "optimal tree": Upon the suggestion that the authors perform an AU test, the response was: "We consider the correct topology to be the one derived from a total evidence approach; i.e., the largest and most comprehensive dataset (all-combined). In our opinion, a topology test would not help in this case, because the test is dependent on which alignment was used as a reference (intron, exon, UCEs etc.). This inherently introduces user selection bias. To avoid confusion, we have changed the terminology from “correct topology” to “optimal topology” whenever referring to our results"

I disagree with this explanation, in fact, choosing a "total evidence" as the "optimal" (which is the same as saying "best") topology is also subjective and biased. Why is T4 worse than T1? or T2, T3, T5? only because it doesn't match what they have chosen as their preferred topology? I still am of the opinion that the authors should carry out the AU test, or any hypothesis testing to discard that the alternative topologies are not equal descriptions of the data. Ideally, they should carry out an AU test for each alignment they tested against all topologies. This is something that can be carried out very simply in IQtree. At the minimum, they should carry out this test with their "optimal alignment" and testing all different retrieved topologies against this. This analysis would also add support to the authors' taxonomy suggestions at the end of the Discussion (section 4.4).

Bottomline is, unless the authors can prove that one topology is better than another through a test as explained before, using the terms "best", optimal", "worse", "suboptimal", "right", "wrong" is not correct and not supported.

As requested, we have performed a suite of topology tests (including the AU test) comparing all inferred topologies (T1–T5) with the optimal/largest alignment (All-combined dataset). As anticipated, and in agreement with our previous inference, the T1 topology was found to be the optimal topology. Methods and Results have been updated accordingly and details are provided in the Supplementary Material.

- Data displayed and available: Although the authors have now included the supporting alignments and most of the resulting trees, they still need to include some that I was not able to find (namely the results of the SVDQuartet) and also it is my opinion that the paper would benefit enormously if the authors provided each of the resulting topologies in the Supplementary information (see comments for "Additional results below").

We have included the SVDQuartets tree (SVDQuartets_T5) in the Dryad repository. We do not think that it would be beneficial to include all resulting trees in the Supplementary Material because of the numerous tips that will make for poor visualization. Additionally, there are metadata within the trees that users may be interested in and this will not be able to be shown using an image. That is why we provide the original tree files for downloading.

- Different data (exon, intron, UCE, Legacy) yield different topologies because these data are likely undergoing different selection pressures and it is something that has been already observed in phylogenomic studies. I think it warrants a mention in the discussion (I'd say it would go well in section 4.3 Causes of discordance). You should have a look at Shen et al. 2018 https://www.nature.com/articles/s41559-017-0126 and Bravo et al. 2019 https://peerj.com/articles/6399/ .

We have added this to section 4.3 and adjusted the language of our conclusions to reflect the other possible factors (other than ILS) that were not assessed in this study.

Minor observations:

- Trees deposited in the Data Dryad: I only found the trees for ASTRAL and IQtree, but couldn't find the SVDQuartet trees. This is either missing or the labels need to be made more clear.

Added the SVDQ tree to Dryad. Only one representative tree was added because they were all identical.

- Methods and Results: 4 or 5 outgroups? In the main text for the methods it says 5 outgroups (which should be all fully listed in the main text). But then table S1 only has 4 outgroups listed. Also there is a major inconsistency with the outgroup species listed: Table S1: missing Scaphiophryne marmorata and Arthroleptis_variabilis. Also said table has Abavorana nazgul which doesn't appear anywhere in the alignments. Please correct this.
Also following from this point, you should add a sentence in the Methods main text of why you chose those 5 outgroups, and how distantly related are they from the ingroup.

We apologize for the oversight. The errors are in Table S1, which we have corrected. A total of 5 outgroup taxa were used and these are now accurately represented in Table S1. *Scaphiophryne marmorata* and *Arthroleptis_variabilis* have been added and *Abavorana nazgul* (incorrectly included) has been removed. Outgroup taxa have been listed in the Methods, as well as justification for their usage.

- Materials and Methods: Bioinformatics: Line 146. Alignment information should be mentioned here in the main text, also add the software used.

We are already at the word/page limit and do not think that this needs to be in the main text. All details, including software used are provided in the Supplementary Material.

- Lines 190-205: it's missing information on which datasets were Bootstrap carried out and needs to specify the bootstrap calculated (parametric Bootstrap or the Fast bootstrap which are different and both available in IQTree).

The reviewer's reference to line numbers are not in sync with our uploaded manuscript, hence it is hard for us to track down. Nevertheless we have added bootstrapping information to section 2.3.

- Line 282-286: Where are the analyses that support the statements made? can you supply the distance between the gene trees and the "large" trees or how did you check for this. These statements need reference to the results from your analyses to back it. These should be supplied for both topology and branch lengths.

We have edited to sentence for clarity. No analyses performed to validate this assumption, that is why we made it clear that it was just a possibility. We have provided a reference to back up our hypothesis:

"Surprisingly, the SVDQuartets analysis inferred a novel topology across all datasets (T5) which was not supported by the topology test, possibly due to high ILS and large numbers of sites per locus that is known to affect the accuracy of this analysis [58]."

- Outgroups missing. I was surprised that I couldn't see any of the outgroups in any of the supplied figures. Figure 1a is the only tree you show and it does not depict the full tree of your results. In the spirit of being fully transparent with the information, it would be better to show the outgroups somewhere, such as resulting trees in the Supplementary information (see point "Additional results")

Added the outgroup to Fig. 1

- Figure 1: Also the way the groups are summarised in T2, T3, T4 and T5 are very confusing and one has to spend ages to make sense of the comparisons and the positions of Nodes N1, N2, and N3 across all 5 topologies. I would suggest that the authors label N1-N3 in each of the 5 topologies.

Added the positions of N1-N3 to the alternate topologies

- Table 2: Needs to add more relevant data to support why certain nodes from the different analyses fall in the anomaly zone. You state in the methods and in the Supplementary Methods that you performed "Anomaly Zone calculations". The results of these calculations and the branchlengths should be provided in this table.

Due to the nature of the calculations, the results are binary, I.e if the length of the descent internal branch is less than a(x), it is in the anomaly zone and vice versa. Hence,

the calculations will only tell you if it IS or IS NOT in the anomaly zone; that is why we have used “y” and “n” in the table. There is no other meaningful way to represent these results in a table. There are also two sets of branch lengths associated with each calculation (ancestor and descendent branches), which makes it difficult to summarize in table form. That is why we have provided the original trees in Dryad that have branch length data for reproducibility.

- Fig S2: missing labels on the left hand for all vignettes. Also it would be good if you could supply a sentence or 2 in the caption for this to state what is the point you want to highlight, the major result from this.

Added label for the y-axis and expanded the caption.

- Line 380-381: "Quartet scores for UCE datasets were also higher compared to introns and exons, indicating that UCE markers may be less affected by ILS" add reference to results for this: Table 2.

Added reference to Table 2

- Additional results: Also I would like to ask the authors to provide a visualisation of the resulting trees for each of the 11 datasets and all the different methods in the supplementary information, as this will help any reader to compare the underlying results. I personally had to copy and paste each of them in a tree viewer, why make life harder to your reviewers when it is better to show all your results in an explicit manner? Please add in each tree a label of Nodes N1, N2 and N3 and which topology they result in (T1, T2, T3, T4 or T5). And also add the support values calculated for each branch based on whatever method used.

If we were to include all trees from all datasets and analyses, we would have to show a total of 32 trees. This will inundate the Supplementary Material section and make it very tedious to read. Most readers will be interested in the additional methods and results, but will not be comparing all trees simultaneously. Hence, in our opinion, adding 32 pages of just trees to the supplementary material will actually be counter-productive. We have already made all trees available for download, which we think is sufficient. Readers will be able to specifically choose which tree they want to examine, instead of scrolling back and forth through 32 pages of trees, which they won’t be able to manipulate. The raw tree files also contain additional metadata (e.g. branch lengths) that we won’t be able to show as images

- Figures 1 and 3: Node 1: very confusing this "Stelladerma" appearing out of nowhere (not shown in Fig 1, not appearing in any alignments) until you reach the end of the paper and even then, it's not clear the clade of Theladerma that the authors refer to. Anybody who is not familiarised with Rhacophorid or even frog taxonomy are going to find this very confusing. Names for clades and species should be consistent in all figures.

We thank the reviewer for pointing this out. We have reverted *Stelladerma* back to *Theloderma* for consistency and Figs. 1 and 3 have been updated accordingly.

- Figure 3: I noticed that you labeled some of the topologies T1, T2, T3 in some cases next to some of the alternative topologies. But nowhere in these appear T4 ot T5. Is this because of an error? I actually think it would help to illustrate some of your points in the discussion if you labeled in this figure which topology matches which of these hypotheses (for example take Line 425-426 you mention at Node 1 you retrieve T4. Label T4 in the hypothesis that corresponds here for this node in this figure). I think you should do these for all Topologies (T1-T5).

This is not an error. The different topologies (T1–T5) correspond to discordance at different nodes. For e.g. topology T2 and T5 correspond to discordance at N2, while T3 and T4 correspond to discordance at N3 and N1. Some topologies like T5 are not shown in Fig. 3 because none of the alternate gene trees supported that topology. For added clarity, we added the nodes to all the alternate topologies in Fig. 1

- Phylogenomic relationships (4.4): I find most of this whole section confusing and difficult to follow because the authors do not point to which parts of the trees they refer to (Node 1, Node 2, Node 3). Also, can you name/find any other supporting evidence for your proposed hypotheses? like morphology, ecology, etc in disfavour of the "Stelladerma hypothesis". It would also help that you define in a very brief way what is this hypothesis.

This has been addressed in our response to Reviewer 2 above. We paste our response to that comment below:

“Upon further consideration, we have decided to remove section 4.4. This entire section is very taxon-centric and will only be of interest to a small fraction of readers who work on this group. Furthermore, it detracts from the main goals/theme of the paper, which is to elucidate sources of discordance, not resolve taxonomic uncertainty. We found that this section is mostly a summary of the results and thus, does not need a dedicated Discussion. Removing it will also prevent the paper from exceeding the page limit.”

- Lines 391-392: "Bootstrap values were not correlated with topological concordance and routinely produced strong support for highly discordant nodes (Fig. 2). " You should add this is in agreement with what has already been explain in Minh et al 2020 (https://doi.org/10.1093/molbev/msaa106).

Added citation to text

- Line 407: "However, higher concordance values can also be an artefact of small datasets" same as previous observation, this is exactly what was explained in Minh et al 2020 (https://doi.org/10.1093/molbev/msaa106), so it would be good to add this as reference here.

Added citation to text