Paris, July 25, 2019

Dear Editor, dear Robert,

We are pleased to resubmit our paper entitled “What do ossification sequences tell us about the origin of extant amphibians?” to PCI Paleontology. It took us nearly a year to revise it because we collected many additional data and performed several more analyses. We were able to accommodate most comments by the reviewers, and in the very few cases in which we could not, we justified it thoroughly. We are very pleased with the result and we hope that you will find it suitable to be recommended by PCI Paleontology. Below, our responses are in **bold type** (if the way in which I can transmit this letter to you allows this formatting).

Best wishes,

Michel Laurin, on behalf of the three co-authors.

Dear Michel,

As of 22 Aug I've got three reviews for your PCI manuscript, and have made some further comments of my own, all pasted below. There are some teething issues with the PCI website, hence the delay in getting this info to you. Ideally the whole process would go via the website, but until this happens I've taken the liberty of just emailing you the reviews & recommender (my) decision here so you can get started on the revision sooner rather than later.

Sincerely, Robert Asher

Dear Robert,

Thank you. Below, I respond in bold type to the various comments.

Best wishes,

Michel

***
Recommender comments on Laurin et al. PCI-Paleontology by Robert Asher
22 Aug 2018
I've just got a third review in today and will make that available to the authors. (I don't think it shows up yet on the PCI-Paleo site). My comments below were written after having just the first two reviews; the third doesn't change my decision to "recommend revision", but does provide further constructive critiques that the authors should consider. My editorial comments are pasted below, followed by plain text comments from each of the three reviewers, R1-R3 separated by "***".

***
21 Aug 2018

Overall I like this manuscript and am keen to see it as a formally accepted paper in PCI-Paleontology. Both reviewers raise a number of issues which need to be addressed. R2 in particular

argues that taxon and character sampling is not quite sufficient to reject all hypotheses besides monophyletic origin among lepospondyls, or at least not as strongly as the authors do in this manuscript. I welcome a revision taking these critiques into account, and if possible increasing the scope of taxa and/or characters sampled as per reviewer critiques.

I believe that we incorporated all comments as best we could. For the most part, this was not problematic and we feel that the paper is much clearer and more thorough now. Thus, we are very happy with the revised draft, and we hope that you will find it suitable for PCI Paleontology. The few cases in which we could not comply with requests are fully justified, both in the reply letter, and in the draft itself. The taxonomic sample has been improved by including lepospondyl sequences, though these can only be incorporated in the analyses of the new appendicular dataset, which provides results of a lower quality than those of the skull. Thus, we succeeded in expanding both the taxon and character samples.

Unfortunately, as explained below, adding the aïstopod (*Phlegethontia*) sequence is not possible because its resolution is too low. The method that we used is for continuous data; the approximation of sequence position improves with the number of developmental stages represented in the sequence, and that one is just too low (three stages, but just two relevant for our study). In addition, two of the seven bones that we study are not represented by separate ossification centers, so this would generate additional missing data (which our method cannot handle), and there is a controversy about its affinities: stem-amphibian or stem-tetrapod, and lepospondyl or not. Obviously, the referees who suggested adding *Phlegethontia* had not understood these points (their comments show this plainly enough). We added much new text to explain the method (and associated constraints) better.

Additional, minor comments of my own are pasted below. Please respond to all of these and the reviewer comments in your revision.

In the caption for fig. 1 write out the "LH" abbreviation (and wherever possible minimize acronyms in the text)

We did. We wrote “LH: Lissamphibia nested among “lepospondyls””. That is the explanation of “LH”. On the figure itself, this acronym appears too, as well as repeatedly in the text, but we added in several locations, including in the legends of some tables, as reminders of its meaning, so it should be clear to all readers.

line 103: depending on your response to the reviewers, and given that you're looking at cranial sequences, it would be more informative here to note "...extensive database on cranial ossification sequences..."

We now include also analyses of appendicular data. We distinguish between both datasets throughout.

line 111: state what the software was.

Done. This was specified later in the draft, but we repeated it here.

line 122: Characters missing for a given fossil, but present in other taxa, can have an impact on phylogenetic estimation for that fossil by one or both of the following interrelated effects:
- changing placement of taxa to which the fossil may be related
- changing number of steps (in a parsimony context) on a given tree of other characters that are

known for that fossil.
Given the above (detailed in Asher et al. 2005 JVP 25(4):911-923), are you sure that characters "could not be scored for the temnospondyls Apateon and Sclerocephalus, so they could not have helped resolve the main question examined in this study"?

In this case, there can be no effect because we did not perform phylogenetic analyses to obtain our reference trees. Instead, we compiled them from the literature and placed the few extinct taxa on them in the positions predicted by the six hypotheses. Thus, our characters do not influence the placement of the taxa in the tree; they just yield a different fit, globally, of the data to the trees. In any case, the software that we used cannot handle missing data, so it is simply impossible to analyze characters with missing data in this case.

line 144: I don't think many readers will remember these acronyms, but I recognize the benefit of not having to write each one out at every occurrence. Perhaps add more frequent references to your fig. 1, for example here, and remind readers that acronyms are defined & figured in your fig. 1.

Good idea! Done. We also now explain the acronym choice in the legend of Figure 1 better, which will serve as an essential reminder of the six tested hypotheses.

line 168: It would be straightforward to apply an optimality criterion to these sequence data and actually test if they are indeed "unlikely to provide a well-resolved tree". You wouldn't need to figure anything or write at length, but a note simply that---assuming you're correct--- method X (e.g., parsimov or others you prefer) "...yields an unresolved tree so instead we tested likelihoods of the competing hypotheses in Fig. 1 ..."

Good idea! We did this and discuss the results, which are approximately what we expected.

The parenthetical on lines 186-187 sounds a bit too informal & personal and I'd recommend deleting it.

Done.

line 203: replace "consensual" with something like "consensus" with relevant citations of the papers/phylogenies behind this consensus.

We reformulated, but the source papers are already listed, taxon by taxon, in the few paragraphs that follow that line (see lines 396-452).

line 209: I think the term "databases" is more familiar written as one word.

Replaced.

line 234: Fabre et al. 2012 (BMC Evol Biol) present a well-sampled rodent phylogeny that includes both P.melanophrys and M.auratus; ensure that Wilson & Reeder 2005--- a more taxonomic than phylogenetic reference--- are consistent with their estimate.

Indeed, it was a bad idea to base this on Wilson & Reeder 2005. We looked into this and updated our text. We now cite two recent papers for this:

Lu T., Zhu M., Yi C., Si C., Yang C., Chen H. 2017 Complete mitochondrial genome of the gray red-backed vole (Myodes rufocanus) and a complete estimate of the phylogenetic relationships in Cricetidae. Mitochondrial DNA Part A 28:62–64.

Zhuang L., Bluteau G., Trueb B. 2015. Phylogenetic analysis of receptor FgfrL1 shows divergence of the C-terminal end in rodents. Comp. Biochem. Physiol. B 186:43–50.

line 243: this may seem like a trivial point but it's quite important: unless you're interested in gene trees, phylogenies aren't "molecular" or "morphological" but rather the data used to reconstruct them can entail one or both. So write "molecular phylogenetic analysis" or "... most recent phylogenies based on genomic/molecular data" rather than "molecular phylogenies" (as you've done elsewhere in the manuscript, e.g., lines 15, 50, 82 …)

Done.

line 245: can you better justify your disagreement with Irisarri et al. 2017 (or cite someone to this effect)?

Done. The paragraph was greatly extended to explain why we disagree with their dates (and what was wrong with the analyses).

line 261: you might add an "and" before "in case", and/or better explain the connection between the "continuous evolutionary model" and why equal branch lengths are important to minimize bias against a particular hypothesis.

Done.

line 266: remind readers why inclusion/exclusion of Sclerocephalus & the squamosal is relevant here.

Done. It is simply that *Sclerocephalus* is a second Permian temnospondyl genus, but it cannot be scored for the squamosal, so we can include either *Sclerocephalus* or the squamosal. Hence, the analyses with *Sclerocephalus* include only six bones instead of seven.

line 291: "previous attempts" sounds pejorative, and agree or not, Anderson's conclusions are not simply "attempts". So delete "attempts" and just write "previous phylogenetic conclusions from ossification sequences..."

Done. Though we also view our work as an attempt (like all science, in fact), so we did not intend a pejorative connotation here.

line 299: by "untenable" you mean broadly regarded as false, as in amniote or mammalian non-monophyly, right? Please clarify.

Yes. We now omit all judgment from that sentence.

line 356: I'd start a new sentence after "LH" and delete "especially" & start the new sentence with "Similarities ..."

Done.

***

R1
This paper addresses a fundamental problem of vertebrate phylogeny, one that concerns the relation

of a major extant group with extinct ones that were quite diverse and important in the Paleozoic. The authors examine ontogenetic data in the form of ossification sequences of the skull, and integrate information in a large-scope analysis that is sound in the critical examination of the data and of the limits and power of the method. The alternative hypotheses are clearly laid out and the previous attempts to use the kind of data investigated are properly revised. The study benefits from published data on exceptional fossils and makes the best out of those data - yet sampling is limited to relative few species – that is the nature of the data. The authors took steps to account for biases that could be introduced by stratigraphic (time) provenance of fossils (e.g., lines 261-263). At the end it is tree lengths and few fossils what provide the tests, but given the importance of the subject and the critical and thorough approach used, I find much merit in this paper and its conclusions to advance discussion of temnospondyls / lepospondyls / lissamphibia relationships. I suggest the authors revise a couple of potentially relevant references that contain much data (see below) and clarify some points below. One issue that is left largely ignored is that of intraspecific variation, which can be significant in extant amphibians. The data used is in many cases a studied optimization or rather consensus of that variation, but this needs to be mentioned at least, and its potential effect on the study discussed.

Good idea. We now incorporate two sequences representing consensus from two localities for two species each: *Apateon caducus* and *A. pedestris*, Separate analyses are performed using the data from each locality separately and jointly (three analyses), which seems like the most natural way to combine and segregate these data. This is introduced on lines 135-149 and discussed elsewhere in the paper.

For the lepospondyls *Microbrachis* and *Hyloplesion* which we have added now, Olori (2013) has addressed the problem of intraspecific variation (sequence polymorphism) at length and provided consensus sequences which we have used. For *Sclerocephalus*, *Archegosaurus*, *Micromelerpeton* and *Eusthenopteron*, data on variation are lacking; at least for the first two of these, the number of known individuals is almost certainly too low to contain much variation.

The paper concerns skull data, but valuable insights on published, postcranial data on ossification sequences are presented. Abstract, line 17, perhaps because the diversity of methods used hampers comparisons‘. I would rephrase this, as this paper conducts specific analyses and points into a direction that does not make this clause here fitting

Done.

– I would write ‘integration of data’ as opposed to ‘comparison’.

Done.

As formulated, there is a contradiction in the Introduction in terms of the use of molecular data – please rephrase.

We looked and did not find a contradiction, but perhaps the text was not sufficiently clear. We added more explanations in this section. We stated that the topology obtained by analysis of molecular data cannot distinguish between several hypotheses, but molecular dating can, because the age of nodes can be compatible with the fossil record or not, under various hypotheses.

Line 105 – for mammals, the most comprehensive dataset of cranial ossification is that of Koyabu et al. 2014 Nature Communications (not Weisbecker 2011) Weisbecker and Mitgutsch 2010 presented a comprehensive summary and analysis of anuran cranial ossification patterns. Besides

citing it for obvious reasons, please make sure you have considered all the data mentioned there. Weisbecker V, Mitgutsch C (2010) A large-scale survey of heterochrony in anuran cranial ossification patterns. J Zool Syst Evol Res 48, 332-347.

Done. This took weeks of intensive work, but it allowed us to include more taxa, so we re-did all the analyses. This is the main reason why it took so long to resubmit this draft.

Lines 245-246 – there is a statement on the disagreement of the authors with Irisarri et al. 2017 calibration dates, but no justification or discussion of why. Some explanation/argument, would be fitting here. This paper is otherwise cited in line 582 – I guess there for the topology but not for the divergence times, correct?

Indeed. We have clarified by explaining in detail why we don't take dates from that paper.

Figure big phylogeny: given that the divergence dates of the placental clades is far from being universally accepted, based on empirical work of alternative groups, it would be good to add some statement to this effect.

Good idea. This was mentioned to an extent, but we added "and are controversial" in this sentence: We caution, however, that all available molecular dates for Paleogene and earlier mammal nodes are controversial and may be overestimates (Berv and Field 2017).
***
R2 (Olori)

Overview: Major revisions necessary (or a more explicit re-focus regarding what is actually being tested)
This paper is an important review and analysis of the growing number of data sets about skeletal development in early tetrapods and their potential descendants. Although such individual data sets often are published today, comparing them to one another in a comprehensive fashion, and within a phylogenetic framework is cumbersome and rarely done with any breadth. The authors of this paper attempt to do just that while also answering lingering questions in vertebrate paleontology.

In general, the many data sets are brought together with care and thought – something difficult to do given all the different ways that ossification sequences can be put together and interpreted, and the different morphologies across tetrapods. The authors also make a neat statistical assessment of comparisons across taxa, although by using just a single method. That approach is fine, but the paper would be much strengthened by including other methods as well (PGi, other methods of ranking to standardize, etc.), and comparing results across different analyses. Not only would this provide another measure of "confidence" regarding the results, but it would allow the authors' work to be more easily compared to the work of others, who may have used different methods to assess the evolution of skeletal development (hardly anyone seems to use the same methods these days).

PGi would provide results of a different kind, not directly comparable to Akaike weights, and lacks the ability to provide the equivalent of Akaike weights, or probabilities to compare support for various hypotheses. More importantly, this method rests on an edit cost function that is contrary to our working hypothesis (that the timing of developmental events can be modeled with a bounded Brownian motion model). More specifically, Harrison & Larsson (2008: 380) stated that their function attempts to minimize the number of sequence changes, regardless of their size. In other words, the magnitude of change in timing of an event is deemed irrelevant. We believe that this is unrealistic, as shown by the fact that Poe's (2006) analyses rejected that model (which he called UC, for unconstrained change) in favor of the model we accept (AJ for adjacent states). Strangely, Harrison & Larsson (2008: 380) miscited

Poe (2006) when they wrote that "There is currently no evidence to suggest that in most data sets, sequence heterochronies of larger magnitude are any more likely than smaller magnitude shifts (Poe, 2006)." We suspect that they saw Poe's (2006) paper too late and realized that it invalidated the premise of their method; rather than rewrite their cost function, they seem to have tried to sidestep a major objection. Or, possibly, Poe's (2006) results go against their personal preferences (which they may hold for purely theoretical reasons). No matter what their motivation, we strongly disagree with this approach.

Among the remaining alternatives, the main one would be event-pair cracking with Parsimov. But that method is even more objectionable, for reasons that were detailed thoroughly in Germain & Laurin 2009, but that include the unnecessary decomposition of sequences into event pairs and the fact that the method cannot incorporate absolute timing information (in the form of time, developmental stage or body size, for instance) or branch length information (this latter objection turns out to be irrelevant here because the data fit a speciational model better than a purely gradual one). More importantly, the simulations performed in that paper showed clearly that event-pair cracking with Parsimov yields more artefactual change and has lower power to detect real shifts. This would create problems when trying to compare the fit of the data on various phylogenetic hypotheses. That method is also problematic when trying to infer ancestral sequences, as had been documented previously.
Also, please note that the continuous analysis and related methods based on analysis of ontogenetic sequences using techniques for continuous data has been increasingly used recently. Thus, the issue of comparability with other studies is quickly dwindling.

To sum up, if there were an alternative method that we considered equally appropriate to analyze our data, we would be willing to seriously consider the option of analyzing our data with these methods. But this is not the case, to the best of our knowledge. It is true that comparisons between studies would be facilitated by the use of the same method. But for this standardization to benefit the scientific community maximally, it should be achieved by all scientists adopting the most appropriate method of these data, not for us to fall back on a suboptimal method simply because it is well-established. This is why we used only the continuous analysis. We have added a paragraph summarizing these arguments at the end of the section on "Analysis methods".

We looked into other ranking methods, such as size, but that gave lower resolution, so it had no advantage in this case; on the contrary.

It may also help future workers select particular methods, if the authors could provide some review and comparisons regarding the strengths and weaknesses of each, and whether results are repeatable across different methods.

This would be worth doing if we had simulated data on a known phylogeny; we could then determine which method behaves best. But here, we have empirical data, and the true phylogeny remains unknown. We would no doubt find differences, but it would be impossible to conclude which method is best. This would make our paper more difficult to understand. This is an interesting topic that would require a full paper to treat thoroughly, and indeed, we may do this in the future. In the meantime, we have outlined our objections against the two main alternative methods; this does offer a contrast between these methods, though of a more limited scope than what the referee had in mind, probably.

The authors, in fact, bring up the issue of all these different methods in their abstract, but then make the same mistake they lament, by using just one.

The reference to multiple methods has been removed from the abstract because it was misunderstood by at least two reviewers; we actually meant that various combinations of data types and methods had been used, and even the choice of taxa and characters may influence

conclusions. Using a single method was a choice, not a mistake (we now use a second: direct phylogenetic analysis of the data, at the request of referees, but as we expected, this does not yield useful results). We used the method that we consider the best. The problem with several other methods is that they are not as statistically sound, or they cannot handle large datasets (this is a problem with PGi, as Harrison & Larsson 2008 acknowledged). Our point is not that all authors should use all available methods, but that we should all strive to use the most appropriate method.

An even larger issue, however, is taxon sampling, discussed in detail below.
Major issues:

The authors compared ossification sequences for cranial elements only. To my knowledge, in most lepospondyls for which we have ossification sequence data, the skulls are already ossified in all preserved material. Occasionally data exists for one or two elements, but not for all seven scored by the authors. Perhaps this type of work would be better focused on postcranial ossification material, so that more lepospondyl taxa may be included?

We now incorporate appendicular characters (vertebral ones are unsuitable as mentioned in the manuscript) and perform separate analyses of these data. But as we expected (and as statements in the literature suggest), they yield relatively poor results. Nevertheless, they are now discussed extensively in the draft.

In fact, the lepospondyl taxon for which the most cranial development information exists, is an Aistopod, and that group in the last few years has been supported as a stem tetrapod rather than a lepospondyl (see work by Pardo, Anderson, etc.). That is a major concern for a study that turns up a result of lepospondyl ossification sequences best aligning with those of modern amphibians.

It is true that Pardo and various coauthors (Pardo et al. 2017a; Mann et al. 2019 – all now cited in the draft), as well as Clack et al. (2019), have found the aïstopods far down the tetrapod stem. However, even in these phylogenies, the other lepospondyls – the few that are included – remain close to the amniotes (or even inside). Unlike any of Pardo's work, Clack et al. (2019: fig. 3) included *Hyloplesion* and *Microbrachis* in their dataset and found them to lie as close to the one included amniote (*Paleothyris*) as ever, even though the aïstopods (and the adelospondyls and the urocordylids) wandered off to join the colosteids. Thus, aïstopods may simply not be closely related to "microsaurs", and in that case, they are not relevant to our problem (lissamphibian origins).

It would of course have been great to include aïstopods in our analyses. We tried, and even scored *Phlegethontia* in our master database. Alas, we are unable to use it because of its low resolution and too high amount of missing data, two features that would generate so much noise and limit the taxonomic sample so much that results would be meaningless. Furthermore, vertebral characters are of little if any use for our problem (Danto et al. 2019), and aïstopods don't have appendicular characters.

Lepospondyl taxa must be included, and to do this, postcranial elements will need to be included.

This has now been done, though the results are poor. They are extensively discussed.

Indeed, it seems very inappropriate to test a topology without including the key taxa upon which it is based.

There were very good reasons for this, but this comment is inapplicable to the revised draft, which includes both lepospondyls and appendicular data.

What is really being compared is a situation in which amphibians and amniotes are widely separated from one another by any extinct tetrapods, rather than whether amphibians specifically share a relationship with lepospondyls, to the exclusion of amniotes (ie what is implied in the LH topology). As discussed below, actually including Lepospondyl taxa with data changes the whole pattern of character tracing, which affects ancestral reconstruction, number of evolutionary steps/events, etc. The answer may be completely different, and a different topology supported.

Any time that the data are changed, whether this is by changing taxa or characters sampled, the results may change. Now, lepospondyls are incorporated into the analyses of appendicular data (for cranial data, this is impossible because, precisely, there are no data for lepospondyls, except maybe the insufficient data for *Phlegethontia* as discussed above).

Another more minor issue may be the proportion of extant vs. extinct taxa, wherein the "pull of the recent" may be dictating early tetrapod evolution in terms of pattern of character evolution. Why are we still using living taxa to explain the evolution of their ancestors? It should be the other way around.

This is simply a matter of data availability. Developmental sequences are much easier to get from extant than from extinct taxa. But note that our analyses actually emphasize extinct taxa because the phylogeny of the extant taxa is constant (excepted for the position of the caecilians in some tests); only extinct taxa are mobile. More importantly, in the revised draft, we present some analyses with many more extinct taxa (up to 11, instead of 2 or 3).

Specifics, by line number:

55 – More recent work suggests that salamanders (and maybe caecilians) have lost a tympanic ear that would have been present ancestrally (Anderson et al. 2016). That renders the point here mostly irrelevant, and somewhat more supportive of temnospondyl origins.

We strongly disagree with this interpretation; we have seen no convincing evidence that urodeles or gymnophionan ancestors ever had a tympanic middle ear. In fact, Anderson et al. (2016) miscited our work, and we may respond to that paper in the future (to the extent that Christensen et al. 2016 haven't already done that). However, we do not believe that this is the place to debate such topics here. A large review paper would be required to cover this debate; last time we did that (Marjanović and Laurin 2013), it took 65 printed pages!

Christensen C.B., Lauridsen H., Christensen-Dalsgaard J., Pedersen M., Madsen P.T. 2016 In defence of comparative physiology: ideal models for early tetrapods do not exist. Proc. R. Soc. B 283:20160716.

84 – Substitute "among" for "between" because this refers to more than two hypotheses being compared.

Reformulated.

108-109 – authors need to be more forth coming in the methods about the sources of data and taxa included. Most readers won't access the sup data, and given my reservations above, they need to be honest about which extinct taxa were used (especially among lepospondyls), and the proportion of early tetrapods and outgroups to extant tetrapods.

We have documented this better, and there are more extinct taxa in the new analyses, though

there is still a need for the supplements. PCI Paleontology is an online publication venue, so readers who access it will be familiar with accessing the supplementary data as well.

Not including enough extinct taxa will cause a bias of the "pull of the recent", in which the simply more common conditions of the living groups will outweigh, or even mask, the ancestral conditions present in extinct taxa. That would mean very little could actually be said about the evolution of skeletal development, and invalidate the authors' results here.

There are more extinct taxa in the revised draft. Given that the position of extant taxa (except for the caecilians in some analyses) does not change, only that of the extinct taxa changes, there can be no pull of the recent of the kind evoked above. Finally, note that a "pull of the recent", if it affects such analyses (this has not been demonstrated for the method we use), should actually cause a bias in favor of the TH, DH or PH, not for the LH, in the cases in which only temnospondyls are included.

111- It seems a little unreasonable to choose a method that cannot handle missing data, given that this study focuses on comparisons between fossils and living animals. Most fossil data are incomplete in some way, and this is particularly true for lepospondyls vs. temnospondyls (the latter have a much better fossil record, and more complete ossification sequences).

No method is perfect. Of course, the referee could easily reuse our data, once our paper is published, to perform another study using a method that accommodates missing data. We would have preferred using a method that handles missing data, but unfortunately, we found none that had nearly as good statistical properties as the one we used. At least, we perform quantitative and statistical analyses, unlike many recent papers that have drawn sweeping conclusions!

122- yes another big point in trying to do these comparisons is that some taxa are simply very different. Temnospondyls as a whole, but especially Apateon show early ossification of postcranial material and late ossification of cranial. That is extremely hard to compare with lepospondyls, which generally have a completely ossified skull before the postcranial ossifications. By leaving out either postcranial or cranial elements from the analysis (or, just many other cranial elements, as in this case), the results will be very biased; some taxa that are otherwise wholly different in their total ossification sequence, make look more similar when only a subset is analyzed. This should be done with much more caution, and much more warning to the readers. A lot of information in the methods is left out.

Indeed, it is difficult to compare the ontogeny of *Apateon* with that of lepospondyls, which is much more poorly known. But we did not try to do that. Anyway, inclusion of lepospondyls and appendicular data in the revised draft makes these points moot.

133- this is incorrect. Firstly, Schoch 2006 used the actinopt Amia with fairly few homology problems. Secondly, some part of the development of Eusthenopteran were published (Cote, 2002; Schultze 1984), though admittedly little about cranial development. It would provide some data about postcranial though.

We reformulated the statement about homology problems (Schoch 2006 has a different opinion than us about some homologies, but we need not comment on this big problem further here), and more importantly, we have now incorporated appendicular sequence data from *Eusthenopteron*.

138 – The authors themselves bring up one the major concerns noted above, and honestly state that

no lepospondyls were used. How can their results be valid? With no actual lepospondyls, and no non-tetrapod outgroups, it seems fairly impossible to test their hypothesis directly, let alone confidently place living amphibians with a taxon not even present in the study.

See our responses above; we abundantly explained our method. Outgroups are not needed when assessing character fit onto a tree. The referee evidently forgets that outgroups are used to root the tree, when performing a tree search. But when mapping character evolution onto an already rooted reference tree, outgroups are superfluous. And now that lepospondyls and *Eusthenopteron* are used in some of our analyses, this point is moot.

157 – size already was shown to not correlate well with developmental stage nor ossification sequence, although my own work suggested that because fossil data are missing so much, using size as an approximation for fossil cases, only, doesn't really change our results too much, given that they are so poorly resolved anyway.

Actually, we don't necessarily think that size is bad in principle; we had refuted an example that had been used to show that size was poorly correlated with developmental stage (Laurin and Germain 2011: fig. 1). But here, the problem is that body size data are either missing or not precise enough to provide as precise (resolved) a sequence as the simple order of events.

169 – statistical tests are not my strong skill, so an additional reviewer may be helpful to assess the appropriateness of CoMET and AIC for this application. However, I would add that other authors have compared sequence data in a phylogenetic framework (PGi for example, by Harrison and Larsson), so why aren't those methods also used and compared to CoMET's output? It isn't even discussed why more recent methods are not used.

There is a factual mistake here. PGi is not more recent than the continuous analysis that we used; PGi was proposed by Harrison and Larsson (2008) and we proposed the continuous analysis in 2009. More importantly, PGi is appropriate for small datasets. It would be difficult to use it here. But mostly, it lacks the ability to provide probabilities or an equivalent of AIC weights. We have provided a detailed response in the draft that explains why we believe that these methods are inappropriate for analyzing our data:

"It could be argued that using other methods (in addition to the method outlined above) would have facilitated comparisons with previous studies. However, the two main alternative methods, event-pair cracking with Parsimov (Jeffery et al. 2005) and Parsimov-based genetic inference (PGI; Harrison and Larsson 2008), have drawbacks that caused us to decide against using them. Our objections against event-pair cracking with Parsimov were detailed by Germain and Laurin (2009) but can be summarized briefly as including the unnecessary decomposition of sequences into event pairs and the fact that the method cannot incorporate absolute timing information (in the form of time, developmental stage or body size, for instance) or branch length information. More importantly, the simulations performed by Germain and Laurin (2009) showed that event-pair cracking with Parsimov yields more artefactual change and has lower power to detect real sequence shifts. This would create problems when trying to compare the fit of the data on various phylogenetic hypotheses. That method is also problematic when trying to infer ancestral sequences, as had been documented previously. The performance of PGI has not been assessed by simulations, but it rests on an edit cost function that is contrary to our working hypothesis (that the timing of developmental events can be modeled with a bounded Brownian motion model). More specifically, Harrison & Larsson (2008: 380) stated that their function attempts to minimize the number of sequence changes, regardless of their size. In other words, the magnitude of change in timing of an event is deemed irrelevant. We believe that this is unrealistic, as shown by the fact that Poe's (2006) analyses of thirteen empirical datasets rejected that model (which he called UC, for

unconstrained change) in favor of the model we accept (AJ for adjacent states). Furthermore, analyses of ossification sequence data using techniques for continuous data as done here (see above) have been performed by an increasingly large number of studies (e.g., Skawiński and Borczyk 2017; Spiekman and Werneburg 2017; Werneburg and Geiger 2017, just to mention papers published in 2017), so the issue of ease of comparisons of our results with other studies is not as serious as it would have been only a few years ago. "

186- perhaps the paper was a bit rushed? Why not wait for the corresponding consultant to reply, before abandoning some of the models? The paper would be strengthened by just waiting a little to see if these can be done, and if they cannot, explaining why more thoroughly.

No, we have waited more than a year. There is no telling when or even if the colleague who programmed this will ever have time to look into it. If it has not been done by now, there is little chance that it will be done in the next 12 months. Note that during the revision process, the programmer answered; he just suggested to let an analysis run overnight, but that did not work. The paper was not rushed; it took years to prepare, and many months to revise...

192- true, but this is primarily character mapping with a more refined and modelled approach. That is different from phylogenetic analysis. In the former case, the authors are mapping characters onto existing hypotheses for check for best fit (more in line with objectives anyway, given that the goal was to test those specific topologies). Doing a phylogenetic analysis would have a different goal: see if the signal from development data agrees or disagrees with topologies based on adult phenotypes. That is a different type of analysis with a different type of goal. It doesn't need to be included here if the explicitfocus is testing existing hypotheses of relationships.

That is correct. But at the request of others, we added direct phylogenetic analyses, with results broadly congruent with our anticipations (poor results that don't help to discriminate among competing hypotheses).

However, the two approaches should not be conflated in the methods. They are not alternative approaches because they do not accomplish the same thing, as misconstrued earlier in the methods and repeated again here, though implied rather than stated outright.

They are different approaches because if these cranial developmental sequences were much more abundant (say, for instance, 50 or 100 characters), they could be analyzed directly and might well yield a tree fairly congruent with the established consensus and showing directly where the extinct taxa fit. We do not conflate the methods here. To clarify this, we have added this text, at the same place: "This quantification is another reason to prefer this approach over a phylogenetic analysis (performed below, but with the poor results that we anticipated), which would at best yield a tree showing where the extinct taxa most parsimoniously fit (if we had dozens of characters, this might be feasible). Comparisons with other hypotheses through direct phylogenetic analysis are not possible."

203 – use a different phrase because "consensual relationship" in English means something of a romantic or sexual nature.

Done! Thanks for spotting this!

206 – this is a bit puzzling, because molecular divergence estimates often include fossil calibrations anyway. Those gaps cannot be completely avoided. Also what is the purpose of the time tree? It is not explained in the methods. If developmental sequences are being mapped onto existing typologies already, why introduce yet another tree, and do stratigraphic data really add anything to

the analysis?
This is unclear as presented currently. It seems a time tree is unnecessary, given that so few extinct taxa are included, and as the authors note, there is so much disagreement regarding molecular divergence times anyway. With ossification sequence being so limited, the time tree feels a little redundant/unable to be fully utilized.

We now explain, in the section on the reference phylogeny: "The tree had to be time-scaled because many of the evolutionary models that we fit on the tree in the first series of tests (to determine which evolutionary model can be used to compare the fit of the hypotheses) use branch lengths to assess model fit." We also expanded the explanation, in the beginning of the results section: "This simplifies the discussion, because it means that the original branch lengths are irrelevant (under that model, all branch lengths are equal); unfortunately, the branch length (evolutionary time) data were needed to reach this conclusion. Thus, the only remaining variable is the topology."

Of course, we are fully aware that molecular dating requires the fossil record, but molecular dating can "smooth out" divergence time estimates to compensate for the patchy nature of the fossil record (some clades have a richer fossil record than others). Anyway, the use of timetrees for model-based comparative analyses is standard, and for good biological and statistical reasons.

238 – no mention is made regarding the horrid state of squamate relationships. Which toplogy is used, the one based on morphology or the one based on molecular data? Certainly most of the citations favor the molecular tree, but that is not stated,

Yes, that was stated, lines 388-390 (unchanged): "For consistency and to avoid the effects of gaps in the fossil record, we used molecular divergence dates whenever possible."
The references used were cited lines 426-428: "Three references were also used to integrate squamates in the phylogenetic tree and for the calibration of divergence times: Brandley et al. (2005), Rabosky et al. (2014), Reeder (2003)."

Between both, it is clear that the source trees are mostly based on molecular data. All major taxa have controversies (birds, mammals, crocs, etc.), so we don't feel we need to mention every time that there is controversy. This is well-known, and squamates represent only 4 terminal taxa in our dataset, and only scincids are represented, so this is really an unimportant point for our study.

and the disagreement/issues are not mentioned. The disparity would probably affect divergence estimates for squamates.

Not in this case, given the very small squamate sample size.

245 – no reasons are provided for "disagreeing" with Irisarri's dates. Please elaborate so that the reader is informed and the choice may be assessed.

We greatly expanded that paragraph to give a thorough explanation. This serves as an example of the care that we took in calibrating the tree. This is one of our specialties, so we did not make these choices lightly.

254-255 – this is not really true. Software will test any hypotheses given to it, with any data set of scored characters. However, the lack of lepospondyl taxa in the analysis means that the program is filling in missing data for the taxon, or if the taxon is just left off completely, the character evolution

may not be correct, even if the remaining toplogy can computationally be assessed. In other words, adding in those missing taxa could change which pattern of character evolution is the best match, and thus which toplogy best explains the data.

Obviously, the referee did not understand our procedure here. Referee 3 has similar problems, and we added detailed information to better explain our procedure. It does not involve filling in missing data. Here is the explanation, pasted from above:

In fact, our study assumes nothing about lepospondyl ontogeny. It only shows that the data on *Apateon* are more compatible with a position of temnospondyls as stem-tetrapods than as stem-amphibians, and this is compatible only with the lepospondyl hypothesis. We have added this clarification in the abstract:

"Among extinct taxa, only two or three temnospondyls can be analyzed simultaneously for cranial data, but this is not an insuperable problem because each of the six tested hypotheses implies a different position of temnospondyls and caecilians relative to other sampled taxa. For postcranial data, more extinct taxa can be analyzed, including some lepospondyls and the finned tetrapodomorph *Eusthenopteron*, in addition to temnospondyls."

We have also inserted a similar explanation on line 196:

"However, as explained below, the absence of lepospondyl sequences in our cranial dataset does not preclude testing the six hypotheses (TH, PH1, PH2, DH1, DH2, LH; see above or Figure 1 for the explanation of these abbreviations) because each of these six hypotheses makes different predictions about where temnospondyls and caecilians fit relative to other taxa. Thus, in the absence of lepospondyls in our dataset, the tests of these hypotheses are somewhat indirect and inference-based, but they remain possible. Our tests based on postcranial data include two lepospondyls (*Hyloplesion longicostatum* and *Microbrachis pelikani*), but the absence of caecilians in that dataset proves more limiting than the absence of lepospondyls in the cranial dataset because the TH, DH1 and DH2 become indistinguishable (Fig. 1 c, g, h). However, the presence of lepospondyls allows us to test two variants of the TH/DH distinguished by the monophyly (e.g. Ruta and Coates 2007) or polyphyly (e.g. Schoch 2019) of "branchiosaurs" (the temnospondyls *Apateon*, "*Melanerpeton*" and *Micromelerpeton*)."

Note that this was already explained, in other words, on lines 453–462, and we have left these lines largely untouched, hoping that the other explanations that we inserted in the abstract and on line 196 clarify this sufficiently.

263- it is unclear why branch lengths would all be made equal in the end, after all the methodology regarding the different evolutionary models that the authors implemented earlier in the methods section. Were those other models used and tested? Perhaps this just needs to be explained better.

This was explained briefly in lines 265–268 (now 518-520). Perhaps part of the problem is that the fact that the final model has equal branch lengths was in a parenthesis, which the reviewer may not have taken the time to read. Thus, we have removed the parenthesis and inserted that statement between commas instead. We have also developed the explanation on several lines, so it should be much clearer now.

277 – the LH topology minus the actual lepospondyls might be best supported when lepospondyls also are not included in the other topologies, but what happens where their ossification data are included??

For the skull, this is impossible to know given that these data don't exist (at least, not in sufficient quantity to allow our test). For the appendicular data, this is now tested.

As noted above, that changes the whole pattern of character tracing, ancestral reconstruction, number of evolutionary steps/events, etc. The answer may be completely different. It seems very inappropriate to test a topology without including the key taxa upon which it is based.

It is impossible to know what the impact of missing data might be. But for appendicular data, this is now tested.

What is really being compared is a situation in which amphibians and amniotes are widely separated from one another by any extinct tetrapods, rather than whether amphibians specifically share a relationship with lepospondyls, to the exclusion of amniotes (ie what is implied in the LH toplogy).

This point is clarified in the text, as mentioned above (at least twice, so we won't repeat the explanation here).

314- the data are unpublished, but I did do this in my dissertation (Olori, 2011), which might be a good starting point, at least in terms of source material. I never published those results because of all of the concerns and problems regarding ossification sequences well discussed by the authors here.

Good point. We added a long paragraph that discusses the analyses in Olori (2011), towards the end of the discussion.

Note, by the way, that the trees intended to represent the TH and what we call the DH1 in Olori (2011: fig. 5.2, 5.4) lack caecilians, so they cannot distinguish the TH from the DH1 or for that matter the DH2; instead, these two trees differ only in the position of *Micromelerpeton*. This is exactly parallel to the fact that our analyses of appendicular data, obviously lacking caecilians, test two trees that both represent the TH, DH1 and DH2 and differ instead in the position of *Micromelerpeton*.

352 – clever subtitle, but first we need to revisit whether lepospondyls are monophyletic (unfortunately this problem seems to keep recurring every few years). The following discussion is weird, given that no data actually exist for lepospondyl cranial development, other than the fact that it is very early relative to temnospondyls.

As explained above and in the text, it is the position of temnospondyls that indirectly gives the most favored hypothesis, simply because under the LH, unlike under any of the alternatives, temnospondyls are stem-tetrapods. And that is their most parsimonious position according to our analyses of cranial data. We have reformulated the section headline and developed the text to emphasize this.

I am happy to review future versions if the authors plan to continue work on the study. I think with the major issues addressed the paper would be a nice contribution to the literature and a great jumping off point for future use of sequence data in phylogenetic studies, as the authors suggest. I definitely agree with their assessment of the potential for these types of data

***

R3

The conceptual basis of this manuscript is indeed very interesting, especially in light of several studies that concluded ossification sequences don't appear to contain phylogenetic signal.

Thank you! Please note that we had already established, based on simulations and theoretical considerations, that such data should contain a phylogenetic signal and that the main problem that had raised doubts about this was linked with event-pairing (Laurin and Germain 2011).

It remained possible that ossification sequences could in fact contain such signal, but the taxonomic level of this signal has yet to be fully explored. I'd first like to congratulate the authors on compiling such an exhaustive list of extant ossification sequence data sources. This appendix alone will be a useful tool for many future research projects.

Indeed, we hope so!

I have several questions and found areas of ambiguity that in their current state render this manuscript unready for publication. Based on the short explanations of the methodologies, I would find myself unable to be able to repeat the work – the key attribute of reproducible science. However, if these issues can be explained and justified in the text this would make an interesting contribution.

My main concerns are:

1) Clarity of methods

2) Assumptions of the models and tests being deployed (ie., continuous characters, using branch lengths in a composite reference tree)

3) The strength of conclusions based on largely inference alone

For the first, my recommendation would be to more clearly describe the data and methods. I appreciate the text is concise, but some questions remain and the readership and utilization of the approach would be increased if methods could be explained a little more in depth for non-experts to be able to deploy them in their own work, and to fully understand the present work.

We have expanded the text (43 pages, as opposed to 25 in the original version, without the figures, tables, and their legends) and believe that it should be much clearer now.

Line 111, I am uncertain what 216 characters this original matrix is derived from. All sequence data? Or are these characters from a previous phylogenetic analysis that includes non-sequence data characters?

We clarified on the lines just before that only ossification sequence data are included and that the tree is compiled from the literature. All are ossification sequence data.

After the missing data criterion was established, 7 characters remained and these are listed as bone names in the text. What are the actual characters? Their position within the sequence relative to one another? Please clarify what exactly these characters are and amend the text to explain this in the applied order.

Most of this was indicated a few lines below. Perhaps part of the problem was that the seven remaining characters (bones) were listed in a parenthesis. We have now removed the

parentheses to make the text more visible. The only information that was missing is the order in which the bones appear, and we have added this at the end of that section. But to make sure, we added an explanation that these are ossification sequence data, standardized between 0 and 1, using the method of Germain and Laurin (2009).

Line 136, the absence of lepospondyls (and that only 3 fossil taxa in general are included?) is alarming. The obvious question is, how can a relationship between lissamphibians and lepospondyls be supported by ossification sequence data if no ossification sequence data is available for lepospondyls? More on this below.

We have added detailed clarifications about this, and we present new analyses of lepospondyl postcranial data. See our responses to reviewer 2.

Also, why not try using Phlegethontia in a stem tetrapod position? It seems it's position down there is pretty well accepted. Might serve in lieu of a 'fish' basal taxon?

Two reasons render this difficult or undesirable. First, its position near the tetrapod stem should be tested further; our reanalysis following thorough rescoring of the matrix of Ruta and Coates 2007 suggest aïstopods are lepospondyls and closely related to lissamphibians (Marjanović and Laurin 2019). Including it would require, ideally, doubling the number of topologies tested, because for each of the six topologies that we tested, *Phlegethontia* could fit in two positions, so we would end up comparing twelve topologies. Second, and more importantly, its sequence is too poorly resolved: three stages are known, and only two of those are even relevant for the seven bones considered here. Given that our method assumes Brownian motion (describing continuous characters), binary data would simply not match the expectations of the method, meaning that the results would be unreliable. When several states are present, the approximation is acceptable (see Laurin and Germain 2011). In addition, many bones, including the parietal and exoccipital (two of the six bones in our smallest cranial analysis), are not present as separate ossifications, which would generate additional missing data; in fact, *Phlegethontia* has 89.2% missing data! And our method cannot handle missing data, which means that very few characters would be left to perform the test on. For all these reasons, adding *Phlegethontia* would presumably add much noise and might actually decrease the reliability of the results, rather than improve it. Thus, including it would raise more problems than it would solve.

Line 153, Just to be clear, these are the position of ossification events in the series of 7 bones, correct? Could an example using the current data be provided?

We have added explanations just before the formula, and an example after the formula, to make this clearer. Indeed, this is a rather important step in the analysis. We had not explained it in detail here because this had been done in Germain and Laurin (2009) and Laurin and Germain (2011), but obviously, it is useful to include a reminder here, especially because our data filtering procedure (due to the inability of the method to handle missing data) is new (though simple).

Line 157, I feel the philosophy of the reasoning as to why skull length was not used to standardize the data is not sound. Just because results are less clear doesn't mean the test isn't working. What seems most likely is that the vast size differences of the organisms at comparable developmental stages would cause problems. Perhaps exploring this justification would make readers feel less like this was being discarded as an option simply because it didn't give a clear answer.

We do not imply that standardizing by size does not work, or that it leads to incorrect results.

However, it decreases data availability because body size data are not available for all sequence positions and for all taxa, and hence, standardizing by size reduces information content. Given that we could retain little data (only seven characters), it is important to use the method that discards the least amount of data, and this is what we have done. We have also explained that this is not an inherent problem with body size data but only a practical limitation of data availability.

Line 160, These are the seven characters, correct? Perhaps restate that these are the seven characters that can be found in SM2 (with the definitions there also?). It sounds a bit like these are other data from the seven characters mentioned previously, and I am not sure which interpretation is correct.

Good point. We added the mention that this is the reduced dataset of 7 characters.

Line 172, I wonder if these are truly continuous data. The methodology renders the data continuous-like values, but I feel they aren't actually continuous in the real world (they are discrete events). Does this factor violate the assumptions of the models being fitted to the data? Perhaps a little explanation can clarify this so I don't wonder if the tests are all invalidate by this interpretation.

The events are discrete, but they occur in continuous time, so using a method for continuous data is appropriate. We have added an explanation about this in the text at that place. However, if we added *Phlegethontia*, these data would indeed be far from continuous, as explained above. This, combined with the uncertainty about the position of *Phlegethontia*, would probably create statistical artifacts.

Also in this section, I wonder about the treatment of branch lengths. Since the reference tree is a composite, the original branch lengths are no longer relevant in the composite tree. An analysis would need to be rerun with all the taxa to get those original branch lengths. So I believe any test involving original branch lengths from separate analysis whose trees were stitched together are invalid. Those characters were not given a chance to participate in the branch lengths of parts of the trees not included in the original analysis (e.g, mammal characters do not get to contribute to branch lengths in the amphibian part of the tree).

The branch lengths are not established from our data; they represent geological time and are taken from the literature. Our procedure consists of estimating divergence times between all taxa (geological ages of all nodes). When taxa are pruned, branch lengths are adjusted automatically. This is now explained, above and below that passage.

Line 254, I like this logic, however, it is not caveat free. That is also ok, but a detailed inspection of what is actually being tested, rather than what is the stated goal leaves me very cautiously accepting the conclusions. I will explore this now, and where my interpretations are incorrect, let this guide the author to clarifying the text to justify the conclusions.

We have clarified this point; see our responses to reviewer 2. We have added elsewhere that the test is indirect and that we could not incorporate lepospondyl cranial data because the little bit we have is of too low resolution to be useful and there are substantial uncertainties about the affinities of the only potential lepospondyl with a however partially documented skull ontogeny (*Phlegethontia*).

It seems that the actual variable being used to determine the correct topology for lissamphibians is the position of Apateon (and Sclerocephalus in some analyses).

This is correct, for the cranial data, and has been emphasized more in the revised text.

This implies what is actually being tested is how similar Apateon's ossification sequence is to either salamanders, batrachians, or all lissamphibians with nothing known of variation among fossil taxa (surely there is enough variation among mammals such that sampling only 1 animal could yield drastically different results).

Indeed, among mammals, substantial variations exist, and these are documented in our dataset. For temnospondyls, much less is known, but the inclusion of *Apateon* from two localities for each of the two sampled species (this is new to this version of the draft), as well as additional temnospondyl taxa in some analyses, helps assess how variability in temnospondyls may affect our results.

In order to test what is described as being tested, a true phylogenetic signal in ossification sequence data needs to be demonstrated

We have demonstrated it; this phylogenetic signal is very strong in the cranial data. It is weaker in the appendicular data, and we discuss this caveat extensively. See the new discussion.

and Apateon needs to be demonstrated as representative of a temnospondyl, or at the very least an amphibamid, condition. Basically, when Apateon is better fit on the crown tetrapod stem, I can't help but think this may be due to species specific patterns of ossification or even neoteny dependent patterns of ossification. Based on the exceedingly limited data from fossils at hand, there is no accommodation of variation.

In the original dataset, the only other included temnospondyl was *Sclerocephalus*. In the revised dataset, we incorporate additional temnospondyls, and these additions decrease the risk of *Apateon* greatly impacting the results if it turns out to be atypical of temnospondyls generally. However, ossification sequences are only known from a tiny fraction of temnospondyl diversity, so we have pushed this about as far as the state of the art allows.

I understand the approach, however, much more caution in the results needs to be expressed.

We have explained that the test is indirect, in the absence of lepospondyl data, and added extra temnospondyls.

Furthermore, the conclusions of a lepospondyl-lissamphibian link are ultimately entirely left to inference. Simply that if lissamphibians are placed between Apateon and amniotes is interpreted as meaning that they share a similar ossification sequence to lepospondyls. However, this is entirely not observable. This, in essence, is not testing the LH, since there are no data allying them with that clade at all. The results simply show that lissamphibians do not have a sequence more similar to Apateon than they do to amniotes. That this is consistent with a LH is an inference alone.

Yes, all science is based on inferences (therefore, we did not add this to the paper because this is trivial). Even the result of a phylogenetic analysis (like a parsimony analysis) is an inference. For instance, it implies that all nodes (hypothetical ancestors) have a given character list or value, and this is all inferences. Yet, this is routinely done, and many papers rely on inferred nodal values. Molecular dating, for instance, relies heavily on this (all dates are inferences), yet this forms the bulk of the research program of several systematics labs (Blair Hedges' to take a famous example). We clarified this point (but a bit more briefly), by stating that the test using cranial data was indirect in the absence of lepospondyl data (see replies to reviewer 2). However, we now have postcranial data on two lepospondyls, which alleviates this problem.

I'd lastly just like to verify that the argumentation being presented is not circular. It seems that early on we aren't sure if ossification sequence data carries phylogenetic signal.

There is no circularity. On this dataset, the presence of a phylogenetic signal was tested using a tree that is not derived from these data, so independence (lack of circularity) is ensured.

An analysis was performed that searches for a best fit of the sequence data based on phylogenetic congruence. This inherently means an assumption of <yes there is signal> is applied to the analyses.

No. The test would have been circular if it had been performed on a tree obtained from these data. But our tree is based on the literature, mostly on molecular sequences. Thus, there is no circularity. And we performed two tests of phylogenetic signal, both of which indicate that there is a strong signal in the cranial data (but much less in the appendicular data).

Finally, it was concluded that this best fit means ossification sequences are phylogenetically informative.

Yes, that is correct.

However, the best fit model is the one that was attempting to maximize the phylogenetic signal.

Yes, but that is not circular given that the phylogeny was not obtained from these data – or any ossification sequence data at all.

I find for the main goal and all analyses presented, that some discussion should be made about the actual sequence data analyzed and what about it might be phylogenetically significant. From my experience with development, I find ossification sequences can be strongly influenced by function (e.g., the timing of usage of an element). As such, I don't expect there to be much phylogenetic signal, and as a result I am not surprised that Apateon is different from lissamphibians. I do not take that result to mean there is not a close relationship between Apateon and lissamphibians. Much of the discussion is spent on topics not related to the study at hand, and while interesting and useful in a broader context, take away from the main findings of the study.

Our analyses demonstrate a strong phylogenetic signal in the cranial data, with $p < 0.0001$ for the squared-change parsimony-based test, and AICc weights $< 10^{-11}$ according to the model-based test. We discuss extensively the phylogenetic significance of the findings. However, the idea to discuss the potential influence of function is good and we implemented it by adding a new paragraph at the end of the first section of the discussion. As this paragraph explains, we use bones from the dermal skull, and they are apparently developmentally tightly linked to each other (Laurin 2014). Thus, we don't expect functional constraints to blur out too much the phylogenetic signal. But it seems to be different for the appendicular data.