Dear Dr. Asher,

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. 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.

I have several questions and found areas of ambiguity that in their current state render this manuscript unready for publication.

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. 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 and useful contribution.

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? 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.

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.

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?

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?

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 method 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.

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.

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 invalidated by this interpretation.

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). Please explain if this issue is being taken into account somehow.

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.

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 means 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). In order to test what is described as being tested, a true phylogenetic signal in ossification sequence data needs to be demonstrated **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. I understand the approach, however, much more caution in the results needs to be expressed in light of the data at hand.

Furthermore, the conclusions of a lepospondyl-lissamphibian link are ultimately entirely based on inference. Simply that if lissamphibians are placed between *Apateon* and amniotes and the models are best fit to that topology 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 direct data supporting 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, and the text should reflect that the results are consistent, nothing more.

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. 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. Finally, it was concluded that this best fit means ossification sequences are phylogenetically informative. However, the best fit model is the one that was attempting to maximize the phylogenetic signal. Just clarify if this isn't the case.

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.