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). 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. 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. 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? 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. Lepospondyl taxa must be included, and to do this, postcranial elements will need to be included. Indeed, it seems very inappropriate to test a topology without including the key taxa upon which it is based. 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. 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.

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.

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

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. 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.

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).

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.

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.

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.

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.

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.

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.

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 explicit focus is testing existing hypotheses of relationships. 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.

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

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.

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

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.

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.

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.

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?? 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. 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).

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.

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.

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.