Reviewer: 1

The authors present a new model for the fluctuation and diversification of Phanerozoic biodiversity. The resulting analysis shows that fat-tailed distributions in overall genus diversity can arise plausibly from Gaussian fluctuations in the sizes of orders (counted in genera), so long as the distribution of Gaussian variances across orders follows an appropriate distribution. In fact, the authors show evidence for both the Gaussian fluctuations of genera sizes, and for a Gamma distribution of variances in this system, which when put together would be consistent with fat-tailed distributions of diversity fluctuation. The authors go on to discuss the interpretation of this analysis in terms of the overall tempo and mode of diversification.

This analysis, summarized in Fig 1 is intriguing, non-trivial, and to me convincing—though I am not as convinced that the subsequent analysis fully gets to the bottom of where this pattern comes from. I have some questions and comments which I hope are useful. Overall, I think the manuscript would benefit from a clear description of the model assumptions, some more discussion of possible explanations for the observed patterns, and a clearer unwrapping of the proposed connections to punctuated equilibrium (and other evolutionary hypotheses discussed).

** If I understand correctly, the authors assume that all orders undergo Gaussian fluctuations around an equilibrium size. While the evidence for this presented in Fig 1 is convincing, do the authors rule out other distributions for the way orders change through time? For example, is it clear that these orders do not undergo net diversification or extinction over time? I did not see this ruled out in preference for the equilibrium hypothesis, and yet it seems like it would be reasonable to include a term for net diversification or extinction over these timespans. If the authors did allow for this, would the equilibrium model still be favored, and would this change any of the subsequent analysis?

I suppose another way to ask the question would be: what happens if the authors do try to fit independent birth and death rates to the time series for each order?

**Related, am I understanding that the authors use the fit of beta_k for an order as synonymous with the diversification rate of that order? For example, discussion on p11:

“A Gamma distribution, while consistent with multiple processes (e.g., (40)), could result from evolution of diversification rates across an adaptive landscape that promotes niche conservatism and punctuated exploration of niche space.“

This might be confusing, because readers may think of diversification rate as being a net diversification rate, which here is assumed to be zero, IIUC.

** I would assume it matters when plotting fluctuation sizes in e.g. Fig 1 what is the timescale between measurements. There is some discussion of this in the SI, but I'd like to understand if the variation in timeslices is enough to affect these analyses.

** Do the authors acknowledge that there is any room for bias in the classification of orders? Mayr's principle of balance has been invoked before to explain power law patterns in evolutionary history (e.g. in the number of species per genus, etc). I found myself wondering if this might influence where this Gamma distribution for the fitted beta's comes from. For example, I would like to have seen if there is a clear relationship between fitted beta and the mean order size. Is it possible that the distribution of mean order sizes underlies the distribution in order fluctuation sizes?

** As stated in the text and shown in S4, the analysis when performed at the level of classes and not orders, seems to fail. I'd like to see an equivalent of Figure 1 for classes, even if placed in the SI. Specifically, does this break down because the distribution of class fluctuation sizes is not Gaussian (maybe itself is already fat tailed, before aggregating across classes?). Or does this break down because the distribution of betas at the class level is not Gamma. The authors show us the resulting P(x) doesn't work, but it isn't clear which ingredient breaks down. (Apologies if I missed this analysis)

** I found a few statements related to niches and evolutionary process a bit opaque, and think that readers would benefit from unwrapping/explaining the authors' thinking further. E.g. from L115

"specifying the probability that a given clade, chosen at random, will occupy a region of niche space

characterized [by] #k"

It is not obvious to me why niche space is characterized by beta, the size of fluctuations. I wouldn't really know how to start interpreting it in that way. Do the authors mean in some sense the potential number of niches within an order is characterized by beta, and orders are occasionally saturating this potential, and at other times (for unstated but probably plausible reasons) dropping below it? Or do the authors regard beta as a trait, that applies to the whole order, and the group of orders together fill out a niche space? It seems like an unusual interpretation of niche, as surely the species within an order are occupying a huge variety of niches. But maybe the authors have some very aggregated definition of niche in mind. This to my mind is all left vague, and it comes up again in discussion:

"Specifically, if #k values are associated with a clade's physiological and life history traits, and those traits evolve via Ornstein-Uhlenbeck-like exploration of an adaptive landscape,the resulting stationary distribution of #k will be Gamma (40,41)."

I guess it is not clear to me what it means to say `a clade's physiological and life history traits', as within an order surely these will be highly variable. DO the authors mean that beta_k might be related to an average of various important traits of lineages wihtin a clade? Again, this connection is not adequately spelled out for readers to form an opinion.

** In summary, I think the authors have found a very striking pattern, which has the potential to aid our understanding of fluctuations in diversity. But I would like to understand better what are the potential explanations of this pattern, and I did not find the connections to evolutionary processes invoked clear enough to be convincing yet.

Reviewer: 2

The present paper looks at fluctuations in generic diversity within higher taxa in the fossil record. These fluctuations were found to follow a fat-tailed distributions and analyzed using “superstatistics” to suggest that they can be usefully treated as shared within taxonomic orders but with variation across orders. This finding is interpreted to support niche conservation within orders, which leads to coherent macroevolutionary dynamics within orders, with shifts occurring when new orders arise.

Overall, I found the paper to be interesting and the analysis clearly explained. I came to it not knowing anything about superstatistics. Although the analysis is tidy and the approach may well be promising, to be honest, I did not see strong advantages over existing approaches. Fundamentally, the analysis dissects variation in diversity fluctuations across groups, which can be accomplished, for example, through a random-effects ANOVA. You’d get a test of significance of the grouping and could use standard approaches to assess relative support for grouping at order, class, or other levels. To be clear: I don’t object to the approach taken, and it may well have advantages as applied in other contexts. From this manuscript, however, I was not able to discern them. And, the finding of macroevolutionary differences among higher taxa is uncontroversial.

Perhaps some of the benefit of the superstatistics comes from the prediction that fluctuations should be Gaussian within groups, whereas the fluctuations become more fat-tailed with increasing heterogeneity in rates? On this point, I did wonder about the error component of observed fluctuations. My sense is that such errors will be substantial, especially for intervals with few extinctions/originations, and also Gaussian in form, which potentially confounds process interpretations. In any case, it is also possible to assess rate homogeneity under conventional approaches, for example by comparing a model with uniform rates within groups to a model with rate heterogeneity (there are various ways one could do this).

I would also say that I think there are places in which the statistical results are interpreted too strongly in terms of macroevolutionary process. One example is in the claim of “equilibrium” within orders. This sounds biologically significant, until it becomes clear that the term is being used in its sense in statistical physics, which can be generated by a wide range of macroevolutionary processes, including constant-rates birth-death, diversity dependence, and neutrality. These span most of of the scenarios considered seriously in macroevolutionary studies.

Another example concerns their interpretation of uniform regimes within orders, and thus that it is only rare events – the formation of new orders -- that break out of these regimes. This general dynamic may be at play, but it seems perhaps to reify the model too much. Alternatively, it is possible that biological and macroevolutionary similarity decrease smoothly with phylogenetic distance and this explains why orders and other higher taxa are more coherent than randomly formulated taxa. It is not clear to me that the analyses done can distinguish these possibilities. It is noteworthy that no analysis is presented with families as the focal higher unit (only orders and above). This may be for practical reasons (some families have low generic diversity), but it showing that families do worse

than orders would support the scenario favored in the manuscript (although there may be issues of sample size and statistical power).

Reviewer: 3

I reviewed an earlier version of this manuscript for another journal, and I was ill-pleased to see that my review was largely ignored. My primary comments were and remain:

1. The hypotheses that the authors wish to test are best framed in terms of richness: i.e., numbers of taxa. However, what they really are examining are shifts in diversity: a more abstract concept (see Hurlbert 1971 Ecology 52:577, which I probably should have cited in my prior review.). Subsampling routines such as SQS compare the diversity of two collections, not the richness. As Hurlbert showed 40+ years ago, if two collections have the same richness but different distributions of commonness, then subsampling routines will find more taxa from the collection with the more even distribution of finds. Ecologically, that collection is more diverse because there are more relatively common taxa. However, the richness is no different: and the models being tested here concern richness. (There are other concerns with the SQS model: see Hannisdal et al. 2017 Proc. R. Soc. B 284: for a discussion of some; but also see Close et al. 2018 MEE 9:in press)

2. Because of this, SQS is not the panacea for the sampling issue. I went on at some length in my last review about the importance of sampling for this. My suspicion is that the authors thought “we used SQS to standardize for sampling: why the reviewer miss that?” I obviously was not clear: the problem is that to get at differences in richness, we need to get at differences in sampling. One could modify Foote’s approaches (e.g., Foote 2001 Paleobiol. 27:602; 2003 J. Geol. 111:125; 2005 Paleobiol. 31:6, 2007 Paleobiol. 33:261) as I suggested in the last review. One could also use methods such as TRiPS (Starrfelt & Liow 2016 Phil. Trans. Royal Soc. London Ser. B. Biol. Sci. 371:20150219), although one should also be concerned that the exponential distribution of commonness leads methods like TRiPS to overestimate richness (see Wagner & Marcot 2013 MEE 4:703). Ultimately, to test the how/why of richness fluctuations, some sort of approach like this is necessary.

3. As before, I think that “niche” is the wrong word for the ecological level that the authors are discussing. They really are discussing something closer to guilds. As I noted before, there is a rich literature initiated by Richard Bambach about guilds and guild assignments. Much of this data can be procured from the Paleobiology Database, too. For example, to get information on the trilobite genus Ceraurus, go to: https://www.paleobiodb.org/classic/checkTaxonInfo?taxon_no=21472 and click on “Ecology and taphonomy.” This provides much of the information to which the authors refer to as affecting “niches” (really, guilds), and they form the bases of Bambach’s guild models. One can also use API: the link http://www.paleobiodb.org/data1.2/taxa/list.csv?base_name=Phacopida&show=parent,ecospace,taphonomy,etbasis will provide this information for all genera within the Phacopida. I have provided R-code for doing this below: I strongly encourage the authors to take advantage of these data! Alternatively, the data used in papers such as Bush et al. (2007 Paleobiol. 33:76) could be used.

4. Following point 3, I am very skeptical that orders are going to be homogeneous ecological units. Indeed, one of the explanations for why groups like orders initially radiated is that they were able to become ecologically diverse.

5. As before, I think that the authors do themselves a major disservice using the word “punctuated,” although they at least have taken “punctuated non-equilibrium” out of the

title. Paleobiologists have taken to using “pulsed” to describe events that are clearly distinct from “background” evolutionary time. Now, in plain English, there is not much difference: but given that “punctuation” has become so strongly linked to a particular speciation model (or set of speciation models), it is best to leave that word alone.

Some other comments.

Figure. 1. There still is no indication of what these orders are.

SQS Analyses: Again, there is no mention of what the coverage levels were, or how exactly they were chosen.

As I noted last time, I think that this paper offers a really interesting methodological advancement. However, I also think that this will be ignored because of objections to the theoretical parts of the paper. If the authors can redo the analyses in the ways that I suggested here and in the prior review, then I think that there is a much better chance of both the methods and interpretations gaining some traction.

Finally, here is some R-code that the authors can use to download ecological data from the Paleobiology Database. Just enter (say) accio_ecologic_guild_data(taxon="Phacopida") and you get a tab-delimited text file with the ecological data that the PaleoDB has for phacopid genera.

```
clear_na_from_matrix <- function(data, replacement="") {

for (i in 1:ncol(data)) {

if(sum(is.na(data[,i]))>0) {

duds <- (1:nrow(data))[is.na(data[,i])]

data[duds,i] <- replacement

}

}

return(data)

}

accio_ecologic_guild_data <-
function(taxon,taxon_level="genus,subgenus",output_type=".txt") {

http <- paste("http://www.paleobiodb.org/data1.2/taxa/list.csv?
base_name=",taxon,"&rank=",taxon_level,"&show=attr,app,ecospace,etbasis",sep="")

fetch <- RCurl::getURL(http)
```

```
taxon_guilds <- utils::read.csv(text = fetch, header = TRUE, stringsAsFactors=TRUE)

taxon_guilds <- clear_na_from_matrix(taxon_guilds,"")

return(taxon_guilds)

}
```


Bush, A. M., et al. 2007. Changes in theoretical ecospace utilization in marine fossil assemblages between the mid-Paleozoic and late Cenozoic. Paleobiol. 33:76-97. (10.1666/06013.1)

Close, R. A., et al. 2018. How should we estimate diversity in the fossil record? Testing richness estimators using sampling#standardised discovery curves. MEE 9:in press. (10.1111/2041-210X.12987)

Foote, M. 2001. Inferring temporal patterns of preservation, origination, and extinction from taxonomic survivorship analysis. Paleobiol. 27:602 - 630. (10.1666/0094-8373(2001)027<0602:ITPOPO>2.0.CO;2)

Foote, M. 2003. Origination and extinction through the Phanerozoic: a new approach. J. Geol. 111:125 - 148. (10.1086/345841)

Foote, M. 2005. Pulsed origination and extinction in the marine realm. Paleobiol. 31:6-20. (10.1666/0094-8373(2005)031<0006:POAEIT>2.0.CO;2)

Foote, M. 2007. Extinction and quiescence in marine animal genera. Paleobiol. 33:261-272. (10.1666/06068.1)

Hannisdal, B., et al. 2017. Common species link global ecosystems to climate change: dynamical evidence in the planktonic fossil record. Proc. R. Soc. B 284:20170722. (10.1098/rspb.2017.0722)

Hurlbert, S. H. 1971. The nonconcept of species diversity: a critique and alternative parameters. Ecology 52:577 - 586. (10.2307/1934145)

Starrfelt, J. and L. H. Liow. 2016. How many dinosaur species were there? Fossil bias and true richness estimated using a Poisson sampling model. Phil. Trans. Royal Soc. London Ser. B. Biol. Sci. 371:20150219. (10.1098/rstb.2015.0219)

Wagner, P. J. and J. D. Marcot. 2013. Modelling distributions of fossil sampling rates over time, space and taxa: assessment and implications for macroevolutionary studies. MEE 4:703 - 713. (10.1111/2041-210X.12088)

Reviewer: 4

Please attached pdf.

This paper offers some very interesting methodological advances on how students of long-term diversity patterns might quantify fluxes. This "superstatistics" approach was hitherto unknown to me, and I found it a potentially interesting tool for assessing diversification patterns that could supplant both equilibrium models (such as logistic or hierarchical diversification), and certainly represents an improvement over exponential models. Indeed, the basic results offer a strong refutation of the idea that diversification has been exponential (which, bizarrely, still has a couple of prominent advocates). Despite the novelty (at least to me: and, I suspect, to most evolutionary biologists) of the statistical methods, the authors do a good job of presenting them in a way that is fairly easy to follow. All in all, I think that there is great potential in this approach.

But…. (and, yes, I know what Ned Stark said about everything before the word "but.")…. I have one major issue with this paper in terms of methods and other major issues regarding the treatment of the data, and therefore with the theoretical implications. Finally, I also have a non-trivial issue with the basic terminology used.

My first issue is methodological. The paper is concerned with stage-to-stage fluctuations in diversity, and finding that they can be modelled well using a normal distribution. However, stage-to-stage fluctuations do not reflect simply changes in standing richness (= origination-extinction). Instead, they represent origination+extinction+sampling (Foote 2001 *Paleobiol.* 27:602; 2003 *J. Geol.* 111:125; 2005 *Paleobiol.* 31:6, 2007 *Paleobiol.* 33:261). Thus, we need to look at how much of the "wobble" is caused by changes in the probability of finding taxa. Here is an example using Foote's (2007) results. These are the ML sampling rates assuming continuous rather than pulsed turnover (Review Figure 1). Note that Michael only published these in figures; he provided me with the numbers upon request a few years ago). Note also Foote provides expected sample per stage; I am converting these to sampling probability based on one minus P[0 finds | expected finds]. Of course, the really big thing to note is that we get the same S-shaped distribution in stage-to-stage-shifts. That means if diversity were constant through time, then we'd still get some wobble similar to what Rominger et al. show in their Fig. 1B simply because of the "Signor-Lipps" effect, i.e., the shortening of stratigraphic ranges because of failure to sample the first/last units of occurrence (Signor & Lipps 1982 *Geol. Soc. America Spec. Papers* 190:291). Of course, what I present are sampling rates for all marine invertebrates combined: what

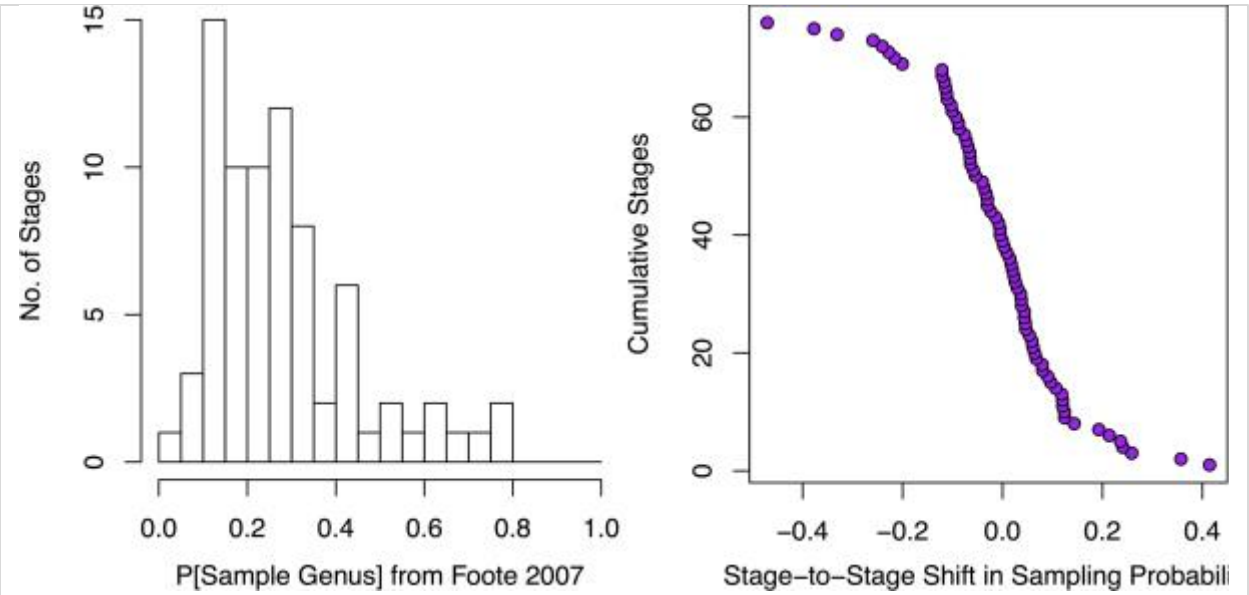


Fig. 1. A. ML estimates of per-stage sampling probabilities based on Foote (2007). B. Distribution of stage-to-stage shifts.

Rominger et al. would need to do is estimate them for each order individually. This is because sampling rates among different marine invertebrate groups differs: groups like brachiopods and molluscs have better sampling rates than groups like echinoderms (Foote & Sepkoski 1999 *Nature* 398:415). Moreover, given the availability of occurrence data in the PaleoDB, Rominger et al. might wish to estimate the distribution of sampling rates within those orders, as the mean sampling rates tend to be arithmetic means of exponential distributions (Wagner & Marcot 2013 *MEE* 4:703).

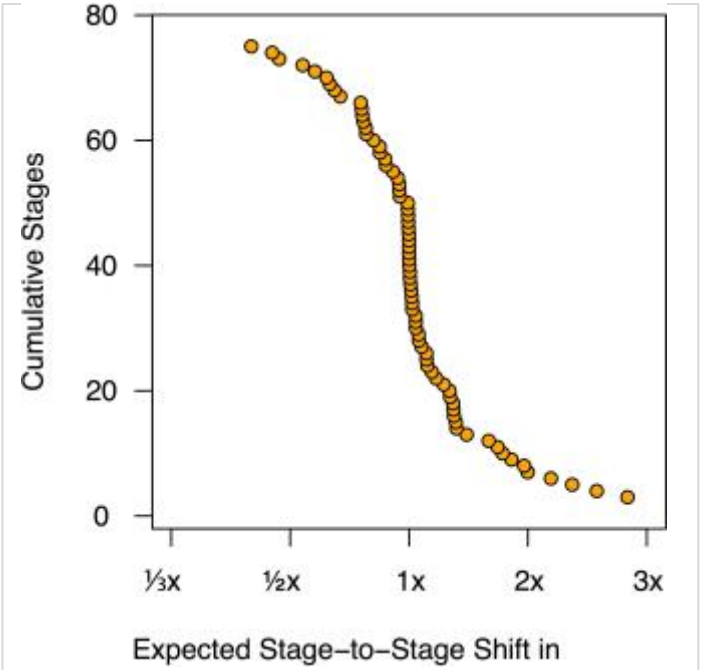


Fig. 2. Distribution of stage-to-stage shifts in expected turnover.

Fortunately, Foote's approach allows one to jointly estimate diversification and sampling. Given origination & extinction, we can estimate the expected shift in richness (see Raup 1985 *Paleobiol.* 11:42 eqn. 1). Review Figure 2 shows expected shifts in relative richness given ML

estimates of per-stage origination & extinction from Foote (2007). Again, this assumes continuous turnover. I also excluded the two biggest gains & losses (both between absolute 4.0-4.5), just to make the graph clearer: but those only exacerbate the S-shape. (They also reflect major extinctions & major radiations.) Now, we still see the S-shape that Rominger et al. document in Fig. 1B: but the advantage is that we have factored out the effects of sampling, so we have a much stronger case that this represents fluctuations in richness. However, I do think that it is necessary to do this, simply because sampling alone biases the fossil record towards showing the pattern that Rominger are attempting to explain biologically.

This leads me to my big issues with using orders as equivalents to some sort of ecological entities in their biological explanations. The authors' cite Holman's 1989 *Paleobiology* paper as a justification for this approach. However, Holman's suggestions never gained any sort of traction. I was a grad student doing paleobiological work at that time, and I had no memory of the paper. It has hardly become an accepted viewpoint, either: the paper itself has been cited only 18 times (excluding the preprint of this manuscript), and at least one of those papers have cited it to point out that evolutionary dynamics within orders are not that cohesive (Alroy 2004 *Evolutionary Ecology Research* 6:1). I would never consider that paper to have "demonstrated" what the authors claim it has: if there is a work that paleobiologists consider "canonical" with regards to higher ecological groups, then it is Bambach (1983 *Biotic Interactions in Recent and Fossil Benthic Communities*) instead!

Work's like Alroy's aside, I think that assuming that there is any sort of ecological cohesion to orders is very unsound. For one thing, Given Bambach's guild scheme, superfamilies probably would be more appropriate: most orders have members in multiple Bambachian guilds. But even if that were not the case, then an equally important issue is that ordinal classifications are not stable. I read this manuscript a couple of times, and I could not figure out if they used the ordinal classifications from the PaleoDB when analyzing those data and the ordinal classifications from Sepkoski's Compendium when analyzing those data, or just went with the classifications from one or the other. However, something that alarms me greatly is the inconsistency between the two datasets. Review Figure 3 shows this lack of consistency. I used only orders with 75+ genera in the Sepkoski database and API scripts offered from the Paleobiology Database to estimate these. One big thing to note is that this is not a "literal" match

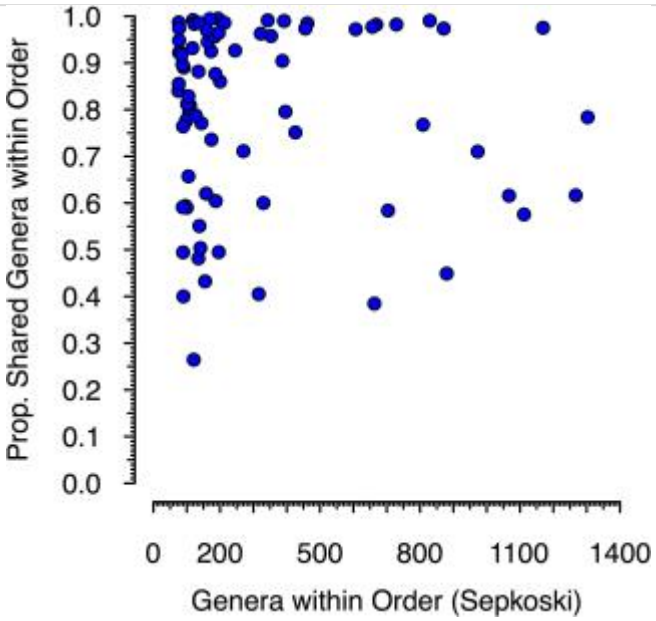


Fig. 3. Proportion of genera within a Sepkoski (2002) order that are assigned to the same order in the PaleoDB. Note that "same order" is means "same as other genera," not necessarily the same named order as used in Sepkoski. Thus, if 70% of the genera assigned by Sepkoski to the Archaeogastropoda are assigned to the Vetigastropoda by the PaleoDB, then there is 70% shared genera between the two databases.

because ordinal names and concepts change more easily than do species or genus names. Because Sepkoski's database uses ordinal names that no longer are used, I found the ordinal name to which the Sepkoski genera are assigned in the PaleoDB. For example, genera assigned in the Sepkoski database to the Archaeogastropoda now are assigned to orders such as the Vetigastropoda; so, if 70% of Sepkoski archaeogastropod genera now are in the Vetigastropoda, then this gives a 70% "shared" number. Regardless, what screams out here is that even the very genus-rich Sepkoski orders frequently contain a lot of genera now assigned to 2+ orders. This actually is not even as good as it sounds: a very large proportion of genera in the PaleoDB are still assigned to the genus used in the Sepkoski Compendium because we have not entered any assignments from taxonomic works yet! (The ordinal taxonomy used by Sepkoski already was becoming outdated when he died in 1998; one of his plans was to update it to follow the growing phylogenetics literature, but that obviously never was able to happen.)

Now, this might seem to be negated by Rominger's et al. finding that (lines 149 -152):

> "Repeating this permutation 500 times yields a null distribution of Kolmogorov-Smirnov statistics that is far separated from the observed value (Fig. 3) suggesting the good t at the order level is not merely a statistical artifact of classication but carries important biological information."

However, my bet is that this null hypothesis is *too* null. It has been well-documented that basic turnover rates vary among higher taxa: for example, trilobites tend to have higher turnover rates than do brachiopods or

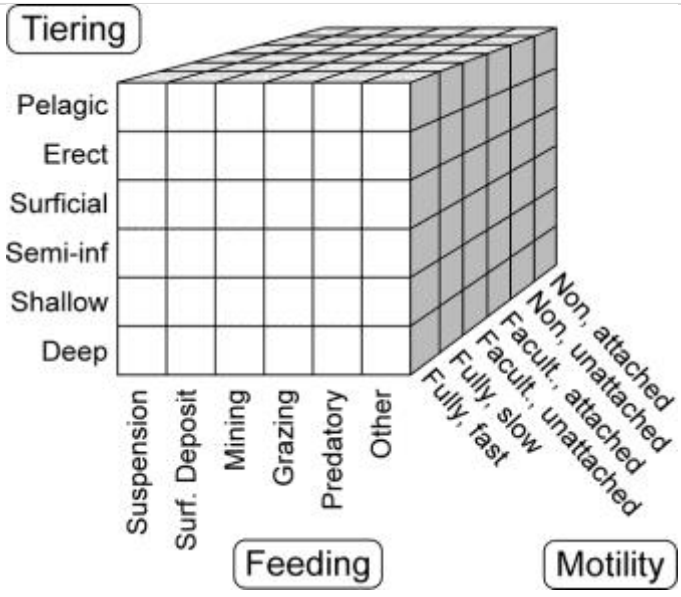
cephalopods, which tend to have higher turnover rates than do gastropods or bivalves (e.g., Sepkoski 1981 *Paleobiol.* 7:36). I think that the more appropriate test would be to permute genera among orders within classes: my bet is that this would greatly reduce the deviations from expectations.

My second issue concerns what the authors consider to be ecological traits associated with a niche. On lines 68 – 71, they state:

> "Larval type, body plan, body size, range size and substrate preference have all been shown to influence such rates. Thus, different regions of niche space, and the clades occupying them, will experience different magnitudes of stochastic fluctuation in diversity."

Of these traits, I would consider only one (substrate preference) to be an aspect of "niche" or even "guild." Range size is a macroecological feature: yes, niche is important in that taxa with commonly occurring requirements for their niches can be wide-ranging; however, what is equally (if not more) important is the ability of species to get to all of appropriate areas. Larval type's greatest effect is on how easily offspring can settle far from their parents, and thus is tied to range size. Moreover, larval type is very independent of adult ecology: we have examples of closely-related marine invertebrates with similar adult ecologies having very different larval types & larval ecologies (e.g., Wray 1996 *Syst. Biol.* 45:308). Body size is important at the extremes: very small and very large marine invertebrates can do things that normal-sized ones cannot; however, in the very wide range of intermediate sizes, it's probably not that important. And, of course, basic body-plans are associated with numerous basic ecological strategies for many taxa.

What I think would be much more appropriate is an approach such as what Bush et al. (2007 *Paleobiol.* 33:76) undertook. They coded genera in a three-dimensional cube, reflecting the basic "guilds" that Bambach (1983 *Biotic Interactions in Recent and Fossil Benthic Communities*) first



Bush et al. 2007 Fig. 1. Theoretical ecospace use cube. Each of the 216 boxes represents a "mode of life"—a unique combination of the tiering, motility, and feeding categories.

popularized. I've provided their figure 1 here. There are three basic axes to this: 1) tiering (how far above/below the substrate the animal extended); 2) basic feeding (suspension, grazing, etc.); and, 3) motility (from fast moving to stationary & attached).

And this is where I have a problem with the ecological interpretation of the results: most of the major orders are going to have genera in multiple cells in this cube. (As I note above, superfamily probably would be the best level to use and feel comfortable that most of the constituent genera are in that cube.) The basic idea of "niche conservation" is that, if only because of phylogenetic autocorrelation, a genus with surficial mobile grazing species is most apt to give rise to more species that are surficial mobile grazers: and when the shells are different enough to get a new genus label, then they still will have traits suited for unattached mobile grazing. Moreover, most of the traits really important for that will not be fossilizable ones. This sort of information is available in the PaleoDB, although most of it is based on assignments for superfamilies and families. Thus, what I would recommend is repeating the analyses using members of (say) classes that fall in the same cube.

Finally, another problem that I have with the paper (and which I had at the very outset) is that the title made me think that the paper was going to be on a very different topic. "Punctuated equilibrium" as a phrase has a specific meaning: a model of speciation in which anatomical change happens over a very short period of time (and usually coinciding with cladogenesis), with very little anatomical change afterwards. This sort of "speciational" change is the opposite of anagenetic or phyletic gradualism, where rates of change are fairly continuous and cladogenesis has no marked effects on change. The role of niches, including how they are conserved, is important to this because, under many models, the waiting time for a "successful" punctuation partially reflects the probability that a new morphospecies can occupy an open or under-exploited niche. Given the title, I assumed that this would be about an intermediate model, in which there is both anagenetic change and speciational change play a role and in which niche conservatism played some role in mitigating either within-lineage (anagenetic) or between-lineage (speciational) change.

Instead, the authors really are talking about evolutionary patterns one level higher in the biological hierarchy: the relative success of major taxonomic groups (orders). This is basically a case where standard terminology and

discipline-specific jargon get in the way. Rapid diversifications of major groups is "punctuated": but we can also call it "pulsed," a "burst," or simply "rapid." Equilibrium vs. non-equilibrium applies to a wide variety of systems, most of which are not even biological: but (for good or ill) Eldredge & Gould decided to refer to stasis as "equilibrium." ("Pulse then stasis" would be a better name for their model, but it didn't have quite the ring.)

The reason why this is a problem is that people seeing the title might assume that it's about speciation models: and thus, people interested in long-term biodiversity patterns might pass on the paper while people interested in speciation models will be disappointed by it. Of course, the (?only) great thing about discipline-specific jargon is that somebody probably proposed a term for a general phenomenon. Many workers use the term "early burst" (e.g., Harmon et al. 2010 *Evolution* 64:2385, Slater & Pennell 2014 *Syst. Biol.* 63:293) to describe punctuated radiations as "early bursts" (particularly when there is an explosion of new morphotypes, which is true for a large proportion of clades for which we have appropriate data; Hughes et al. 2013 *PNAS* 110:13875). So, what Rominger really are describing is a model of radiations with subsequent non-equilibrial "K" (from logistic models).

Finally, it would have been much easier to evaluate some aspects of the data if an electronic version (particularly showing generic assignments to orders) were available.

Minor comments.
Fig. 1 A. What are the orders here? The trilobites must be phacopids. I am uncertain as to which gastropod or cephalopod orders these are.

Lines 223-225: "However, subsampling cannot be applied to small orders (i.e. the majority) because SQS becomes increasingly unreliable as sample size decreases."

On a more general note, what was the threshold for the size (richness) of an order before it could be included? Also, SQS (like rarefaction) essentially assumes that the evenness of occurrences does not differ radically from one interval to the next. Indeed, given that SQS uses singletons as a measure of sampling rather than a measure of dominance/unevenness, intervals with

larger proportions of rare taxa will create the illusion of intervals with lower sampling intensity.

Alroy, J. 2004. Are Sepkoski's evolutionary faunas dynamically coherent? *Evolutionary Ecology Research* 6:1 - 32.

Bambach, R. K. 1983. Ecospace utilization and guilds in marine communities through the Phanerozoic pp. 719 - 746 in M. Tevesz and P. L. McCalls, *Biotic Interactions in Recent and Fossil Benthic Communities*.

Bush, A. M., et al. 2007. Changes in theoretical ecospace utilization in marine fossil assemblages between the mid-Paleozoic and late Cenozoic. *Paleobiol.* 33:76-97. (10.1666/06013.1)

Foote, M. 2001. Inferring temporal patterns of preservation, origination, and extinction from taxonomic survivorship analysis. *Paleobiol.* 27:602 - 630. (10.1666/0094-8373(2001)027<0602:ITPOPO>2.0.CO;2)

Foote, M. 2003. Origination and extinction through the Phanerozoic: a new approach. *J. Geol.* 111:125 - 148. (10.1086/345841)

Foote, M. 2005. Pulsed origination and extinction in the marine realm. *Paleobiol.* 31:6-20. (10.1666/0094-8373(2005)031<0006:POAEIT>2.0.CO;2)

Foote, M. 2007. Extinction and quiescence in marine animal genera. *Paleobiol.* 33:261-272. (10.1666/06068.1)

Foote, M. and J. J. Sepkoski, Jr. 1999. Absolute measures of the completeness of the fossil record. *Nature* 398:415 - 417. (10.1038/18872)

Harmon, L. J., et al. 2010. Early bursts of body size and shape evolution are rare in comparative data. *Evolution* 64:2385-2396. (10.1111/j.1558-5646.2010.01025.x)

Hughes, M., et al. 2013. Clades reach highest morphological disparity early in their evolution. *PNAS* 110:13875-13879. (10.1073/pnas.1302642110)

Raup, D. M. 1985. Mathematical models of cladogenesis. *Paleobiol.* 11:42 - 52. (10.2307/2400422)

Sepkoski, J. J., Jr. 1981. A factor analytic description of the Phanerozoic marine fossil record. *Paleobiol.* 7:36 - 53. (10.1017/S0094837300003778)

Signor, P. W. and J. H. Lipps. 1982. Sampling bias, gradual extinction patterns and catastrophes in the fossil record. *Geol. Soc. America Spec. Papers* 190:291 - 296. (10.1130/SPE190-p291)

Slater, G. J. and M. W. Pennell. 2014. Robust regression and posterior predictive simulation increase power to detect early bursts of trait evolution. *Syst. Biol.* 63:293-308. (10.1093/sysbio/syt066)

Wagner, P. J. and J. D. Marcot. 2013. Modelling distributions of fossil sampling rates over time, space and taxa: assessment and implications for macroevolutionary studies. *MEE* 4:703 - 713. (10.1111/2041-210X.12088)

Wray, G. A. 1996. Parallel evolution of nonfeeding larvae in echinoids. *Syst. Biol.* 45:308 - 322. (10.1093/sysbio/45.3.308)

Review of manuscript #aat0122: Non-equilibrium rate heterogeneity explains fat-tailed fluctuations in Phanerozoic biodiversity by Rominger et al.

This is a very interesting manuscript, worthy of consideration for publication in *Science Advances*. The primary finding is that diversity fluctuations with and between orders, but not within and between classes or phyla, follow a relatively simple compound statistical distribution the implements a superposition of statistics, called "superstatistics". The idea of superstatistics is relatively straightforward, effectively compounding the distributions of random variables from several different distributions to account for the observed data.

One of the reasons the manuscript is interesting is because one might reasonably expect no (relatively) simple statistical characterization would hold for such a complex process as the evolution of 500+ million years of marine animal life. The other primary reason it is interesting is that it holds at just one taxonomic level, but not for the other two tested. Thus, this suggests a higher taxonomic structure to biodiversity dynamics. There is some precedence for this idea (e.g., a coupled logistic equation can be made to fit Sepkoski's family level data, but not his genus level data), but this is the first time that it has been laid out concisely and strongly supported. The fact that the superstatistics only holds for orders is also important because had the superstatistical description held at all taxonomic levels it might have indicated that the finding is relatively trivial (in the way that finding that a whole bunch of morphometric measurements, say height, weight, and arm length, all followed the normal distribution, would be trivial). Thus, I feel this is worth publishing in *Science Advances*.

However, it appears to have been written in the flush of (deserved) over-excitement – it needs some work to maximize its impact, as well as its accuracy, especially given that none of the authors are well-versed in the paleontological literature (or at least so I infer, given the way the manuscript is written [see below]).

<u>Most important suggestions for improvement</u>

1) A major shortfall of the paper is the fact that the manuscript appears to have been written 4- 5 years ago, judging by the fact that the data download from the Paleobiology Database (PBDB) was in 2013, and that the most recent citation is from 2014 (1 only), with just one citation from 2013 as well. The PBDB is still growing at a reasonable rate, and so: (i) the analysis needs to be re-run with 2018 data. And, (ii) the text needs to updated to accommodate the steady rate of publications on equilibrial/non-equilibrial diversity dynamics. In my opinion none of these un-cited more recent papers undermine the fundamental contribution of this manucript, but a new consensus centering on time-varying equilibrial dynamics is beginning to emerge over the previous more polarized equilibrial/non-equilibrial debate (e.g., see the 2016 Phil Trans Roy Soc B volume edited by Quental and Ezhard).

2) The paper needs to make a 'bigger' deal of the fact that the superposition of statistics approach does not fit the phylum and class-level data. I strongly recommend: (i) highlighting this in the text, as well as (ii) bringing the phylum and class figure panels from the Supplemental Information into the main text.

3) Related to the point made immediately above, the analysis should also be performed at the family level, or if that is not possible, that needs to be explained why. Why stop at the ordinal level?

Intermediate suggestions for improvement

The paper tries, perhaps as it should, to discuss, or at least mention, all the related areas that their analysis *might* be relevant to. This is understandable, given the background of the authors (who I surmise come from a tradition where advances are considered important to the extent that they identify deep universal principles). But I feel they over-reach in many small ways that, rather than enhancing the paper, detract from it.

4) This 'over-reach' begins in the abstract, where they claim that "three universal properties of macro-evolution …. are sufficient to explain the previously unaccounted for fat-tailed form of fluctuations in the Phanerozoic. There are five issues with this statement: (i) The ability to describe diversity fluctuation with superstatistics does not constitute an explanation, even though it has implications for what an explanation might be. (ii) The results only pertain to orders, so it is not 'universal' in the way that the abstract and title implies. (iii) It only pertains to the marine fossil record – who knows what the terrestrial record would show. (iv) As they note, they assume (using the words "likely driven by") that "niche conservation" and "punctuated adaptive radiation" are sufficient, with the heterogeneity of diversification rates (do they actually mean *diversity* fluctuations, given that they do not measure rates, per se?) to explain the fat-tail form of the fluctuations – but they are simply *assuming* that the superstatistical fit to the ordinal data is driven by "niche conservatism" and "punctuated adaptive radiation", and offer nothing to support (not unreasonably) these claims. Thus, it is an over-reach to claim that they have explained three properties of macroevolution, as to me their sentence implies. They have found a surprisingly simple description of the fluctuations of just the orders, and while they can account for their data with the ideas of niche conservatism, with the assumption of non-overlapping niches across orders, they have not explained it. (v) The term punctuated adaptive radiation is problematic – later in the text they evoke Gould and Eldredge's 'punctuate equilibrium', but that theory was developed right down at the species level, and has nothing to do with adaptive radiations, per se. My guess is what they are trying to say is that the pattern they find is consistent with each order having a coherent ecological domain, with some sort of transition that corresponds to new orders that also places those new orders in a new ecological domains. However, the transitions need not be fast (punctuated), nor does the filling of the new ecological space have to be fast to explain their statistical result (i.e., the orders don't have to radiate, per se).

So, I think they need to re-word as much of the manuscript as they can, sticking to what the super-statistics results actually are and what they actually mean. And then, in the discussion, they can worry about positing connections between the phenomenology they have discovered to other ideas.

5) Lines 36-38. The casual (*sounding?)* dismal of escalation (Vermeij, [13]) is poor. The Madin et al. paper is weak, and has not gained traction. Moreover, the evidence of escalation is deeply supported, even if there hasn't been a successful attempt to describe it statistically (I suspect the

primary reason it has not been so described is that it consists of many, many, separate events, which, due to the nature of escalatory dynamics, leads to a long-term ratchet in the diversity dynamics). So, this needs to be re-written with a little more nuance.

6) Line 231. This section is labeled: "Three-timer and publication sampling correction", yet in the SM the authors claim that their results are pretty robust whether sample standardization is used or not. Further, the three-timer method of Alroy (date) has been 'improved' upon by Alroy (2014), and Alroy (2015), while there has been another approached developed by Foote (2016). So, I think section has to be labeled to encompass what the authors have actually done more broadly; the way it is written undersells the robustness of their results, and pays a sort of homage to Alroy that will undermine the credibility of the paper (there is nothing wrong with using the three-timer method, but the section has to be written with a broader view).

7) Line 344. It is stated that the data are available though the PBDB – but where are the data *you* analyzed available (especially, after all that cleaning)? To be frank, this statement comes across as lazy and dismissive.

Minor suggestions for improvement

I have not been exhaustive here, there are too many little places where the text could be improved, but here are some suggestions:

Lines 20-23. It is stated in the abstract that that "Its success opens up new research directions to better understand the universal nature of non-equilibrium dynamics across disparate systems of interest from societal to physical to biological." I don't see that the authors' (understandable) excitement about the greater potential of superstatiscal approaches (i.e., societal, physical) has any relevance with respect to this application to the paleontological record. It is distracting. If they really feel the need to say it, it should go elsewhere in the paper.

Line 23. I think the authors might mean "punctuated by innovations *that lead to new* orders".

Line 26. 'extant' tends to mean those living *today*, so maybe say "in the number of taxa alive at any given time'.

Line 30. In the line before, 'distribution' is singular, so should 'tail' be singular on this line?

Line 31. I'm not sure any one has tried to actually 'predict' unusually large fluctuations in diversity. This sentence feels like it needs a re-write.

Line 101-102. It is stated that "The β_k thus represent the inverse variances of homogeneous origination-extinction processes, …". Of what? Genus richness?

Lines 107-111. The discussion of the possible relationship between the Guassian fluctuations in genus richness and neutral-like process (in contrast to the largely dead idea of SOC [self organized criticality]) is distracting, and should be moved to the discussion.

Line 116. Insert "by" after characterized? And why not remind us of what β_k means in the paleontological context?

Lines 116-118. The speculation that "[t]he form of this stationary distribution could shed light on ….." is distracting, and should be moved to the discussion. Part of the reason it is distracting is because for an evolutionary biologist there is already a rich literature on the processes that lead to different clades exploring different regions of eco-space (adaptive landscapes), a literature that is usually couched in terms of the specific properties of the taxa and the environment. So, this statement, rather than advancing the field, I think just undermines the credibility of the authors, largely through the failure in their knowledge of the literature of the paleontological record and its interpretation.

Line 224-225. What exactly does latest Cenozoic mean? The Holocene, the Neogene, the last 10 million years?

Line 240–242: It is stated: "To eliminate further bias due to preferential publication of novel taxa we divide observed number of genera per order per time period by the expected number of genera given publications in that time period". I don't understand this hypothesized bias – please explain!

Line 257: Is some punctuation missing here?

Between line 348 and 349: I assume you mean *commutative* property of the Poisson distribution where the text says "communicative property of the Poisson distribution"

Lines 358–360. This is a sloppy sentence – in this paper the authors claim that much of the raw signal in massive fossil datasets, at least signals regarding fluctuations, are not artifacts of sampling, as has been proposed before [48]. But surely this depends on the type of signal being measured, so the conclusions of this paper need not be at odds with the conclusions of ref [48].

Figure

Figure 1: (i) Shown are three curves, for trilobites, gastropods, and ammonoids. The figure caption states that these are within order fluctuations, so please give the name of the exemplar order. This will help the casual reader to see that you have not done the analysis at the class-level, which is what the icons will evoke. (ii) I presume the red curve is a group of nautiloids, not ammonites, given that the groups persists into the early Paleogene – please confirm, or correct the position of the curve.