12-02-2021-RA-eLife-67509

Essential Revisions

Note: the line numbers reference the diff.pdf file, not the unformatted manuscript.pdf. eLife also required that I submit a formatted version using their LaTeX template (elife-manuscript.pdf); this only has minor stylistic changes compared to the manuscript.pdf (e.g. changes to the way supplementary files are references).

The reviewers appreciated the scholarship and extensive work that went into this manuscript. At the same time, they had several issues with the novel analyses. After some discussion, they suggested that if the paper is revised into more of a review, then it could be of interest to the broad readership of eLife. This would imply substantial streamlining, down-weighting and even removing some analyses, and addressing the reviewers comments on others.

Thank you for the feedback. At present (and discussed in more detail in the replies below), the three novel analyses and findings of this study (see reply to Reviewer 2) make this outside the scope of a review, so I am submitting it as a research article.

Find ways to validate the census size estimates (Reviewers 1 & 2).

Reviewer 1's feedback helped me discover an error in my previous analysis. The census sizes are now a few orders of magnitude smaller for cosmopolitan arthropod species like *Drosophila melanogaster*. I have also conducted a consistency check by comparing the implied biomasses from my study to previously-published estimates across phyla.

The reviewers had a hard time understanding the motivation for the phylogenetic analysis (e.g., Reviewers 1-3). One option is to clarify this motivation, as well as the interpretation of the results; another is to remove this analysis or parts of it. In addition, Reviewer 4, who is an expert on these analyses, had a number of concrete comments that should be addressed.

I have significantly reworked how this analysis was framed, in part thanks to the feedback of Reviewer 4. I also have reworked some sections of the discussion to discuss why such analyses are needed.

While the reviewers found the analysis of linked selection effects intriguing, they were unclear about its interpretation. Notably, to what extent is it simply affirming the conclusions of Coop (2016), as opposed to illustrating a much more dramatic effect (e.g., Reviewers 1-3).

I have reworked part of the discussion (lines 518-524) to discuss how the findings in this paper fits in, and builds off of the findings of Coop (2016). I discuss this in much more detail below.

Response to Reviewer 1

Census Population Sizes

The results are sometimes surprising. For example, Drosophila melanogaster is estimated to have a population size > 10^17 (Fig. 1); if the volume of an individual is 1 mm3, this implies a total volume > 1km x 1km x 100 m.

Drosophila melanogaster was listed in one previously-published dataset used in my study as having a body length of 1.5mm, which is 60% of the actual length, ~2.5mm. Consequently the predicted density was much higher than other Drosophilids. I have fixed this issue, which reduced the predicted population size of Drosophila melanogaster.

Additionally, I found an error thanks to this reviewer's intuition that the population sizes at the upper end of the range were too high. In the previous version of the analysis, I used the body mass and population density estimates from Damuth (1987) directly from the appendix. However, upon inspection of the regression estimates, I found my slope estimate was previously ~-0.9 compared to the expected result of -¾ (which is what is expected under metabolic scaling theory). The source of this problem was that Damuth (1987) has an adjustment for poikilotherms that was applied to his analysis, but not in the table in his appendix. I made this correction in my data, which led my slope estimate to be the expected -¾. This significantly reduced the predicted population densities, and thus population sizes of the arthropods —including *Drosophila*— at the upper range of the population size distribution.

Previously, my density estimate of *D. melanogaster* was about 10^9 individuals per km^2, or 10^3 individuals per m^2. While these densities (or higher) may be possible in a hospitable environment like an orchard, these are too high across the range. This is a known limitation of my method (I added sections on this on lines 187-191). However, applying Damuth's poikilotherm correction so the slope of density on body mass is -¾ reduced the predicted to $10^{6.7}$ - $10^{7.6}$ per km^2, or 5-36 individuals per m^2. This is consistent with the density estimated by Nei and Graur's (1987) of 5 individuals per m^2. I have included this discussion in lines 1198-1205 of the manuscript.

Additionally, some species classified as endangered have census estimates > 10^8 (Fig. 3). The author compares his area estimates with estimates for species in the IUCN Red List (focused on endangered species) to find that they largely correlate (although this is not quantified). I think further investigation of the quality of the census size estimates is warranted.

First, it is a known limitation of the estimate approach that recent decreases in population size due to anthropogenic causes will likely not be reflected in the population size estimates. This is because the population density is based on an empirical macroecological finding, that population density scales with the ¾ power of body mass. In cases where species have been hunted to extinction (such as *Canis lupus*), population densities are often much lower than expected based on this metabolic scaling rule. This suggests that my estimates are most likely biased upwards for species recently impacted by humans, and explains why some IUCN species have large population sizes (as well as why human population size is low, since we can live at higher densities than expected by a mammal of our size). I have expanded on lines 187-191 and lines 1210-1215. Note that if a species' range is reduced, but not their density, then the predicted population size would reflect this.

If the proposed method proves to work well, I imagine that the estimates of census size may be of broad interest in other contexts. In the context of Lewontin's paradox, it may be interesting to quantify the difference in the relationship between \pi and Nc suggested by the new estimates vs the proxies used in previous work (e.g., Leffler et al. 2012).

I agree that it would be a good check to look at how π relates to the range categories of Leffler et al. (2012). I added new Supplementary Figure S1 to show this; it recapitulates the point made by Leffler et al.'s Figure 3.

Biomass Analysis

Are there are other estimates of census size or biomass that can be used for validation, e.g., for species of economic and biomedical importance (e.g., herring and anopheles)?

I thank this reviewer for their suggestion to look at biomass (this is what caught the issue with the poikilotherm correction). I have added a section on my biomass analysis in the Supplementary Material (lines 1169-1197). My approach is to compare the total biomass by phylum implied by my census size estimates and the body mass of each species to the total carbon biomass by phylum groups from Bar-On et al. (2018). As a rough check of consistency, I looked at the ratio of the sample biomass to the biomass of species on earth by phylum, weighted by the fraction of species in the sample compared to the total count of that phylum (Supplementary Table S1, line 1179). Overall, the biomasses implied by the census sizes I have here seem broadly consistent.

Additionally, I also compared my census size estimates to other biomass and population size estimates that were available (lines 1198-1224). Overall, everything looks reasonable after using the same poikilotherm adjustment that Damuth (1987) did.

Phylogenetic Analyses

Quantifying the relationship between \pi and Nc controlling for phylogenetic relatedness: I am unclear about the motivation for this analysis. As Lynch argued (and the author describes), if TMRCAs of neutral loci within a species are smaller than the split time from

another species in the sample, its genetic diversity level was shaped after the split, and it could be considered an independent sample for the relationship between \pi and Nc. There may be underlying factors shaping this relationship that are not phylogenetically independent (e.g., similar life history traits) but it is unclear why that would justify down-weighting a sample. In that sense, I am not convinced by the authors argument that finding a 'phylogenetic signal' justifies the correction. Stated differently, it is not obvious what is the 'true' relationship being estimated and why relatedness biases it. One could imagine that the 'true' relationship is the one across extant species, in which case the correction is not needed (with the possible exception of species in which TMRCAs are on the same order or greater than split times). I don't know what an alternative 'true' relationship would be. Moreover, I am not sure how a more precise 'quantification' of the relationship between diversity and census size serves us. Regardless of corrections, it is obvious that the null provided by the standard neutral model is off by orders of magnitude. Perhaps once we have alternative explanations for this relationship then testing them may require corrections, but presumably the corrections will depend on the explanations.

I understand this point, and this gets to the heart of the issues raised by Westoby et al. (1995) and Uyeda et al. (2018). I have raised these same issues in the discussion, which I have expanded on (lines 437-473; see also lines 230-235), which should help explain why accounting for phylogeny is necessary. However, I disagree with this reviewer that PCMs are not appropriate here. Essentially, the phylogenetic signal I have estimated here suggests that there is phylogenetic structure in the residuals (not the traits, diversity and census size, themselves). A primary assumption of OLS is that the residuals are identically and independently distributed, which is not the case in these data; consequently, the OLS regression estimates are untrustworthy. Another way to think of this is with a thought experiment: imagine that I had a sample that only includes Drosophila species and carnivores (similar to the issue raised in Gillespie, 1991). Suppose there is no phylogenetic inertia in mean coalescent times, but only in the traits leading to census size and mutation rates. The OLS line between these two clade-level clusters is not strong evidence there is a relationship between census size and diversity, yet the OLS regression estimates treat each data point as *independent* evidence that contributes to the estimated relationship. Phylogenetic mixed-effects models account for this correlation in residuals, and weight observations accordingly. I should note too that I have not simply relied on PCMs without also fitting the OLS too (Figure 1), group-level means (Supplementary Figure S5), and PCMs on the relationship between diversity and the constituents of census size, range and population density, separately (Supplementary Figure S6). All of these results are consistent, and do not differ qualitatively. Finally, Reviewer 4, who is an expert in these methods, suggested that this analysis is correct in that it accounts for the phylogenetic correlation structure in the residuals.

One context in which phylogenetic considerations and quantification may be relevant is the comparison of the \pi - Nc relationship among clades. Notably, one could imagine that different population genetic processes are important in different clades (e.g., due to reproductive strategy) and a comparative analysis may highlight such differences. It is

less clear whether the corrections that are applied here are the relevant ones. Separating clades makes sense in this regard, but it is unclear why to correct for non-independence within a clade. Furthermore, it seems that in order to point to different processes one would like to control for the distribution of census population sizes in comparisons between clades (to the extent possible). Otherwise, one can imagine the same process shaping the relationship in different clades, but having a non-linear (in log-log scale) functional dependence on census population size (as in the case of genetic draft studied next). In this regard, I am not sure I follow the argument attributed to Gillespie (1991) and specifically how the current analysis supports it.

I agree with this reviewer that looking within particular clades is an important part of a phylogenetic comparative methods analysis. This was a central point in Uyeda et al. (2018) too. As such, I had previously fit the OLS relationships in phylum-level clades in Supplementary Figure S5. However, this reviewer argues, “it is unclear why to correct for non-independence within a clade”. Taxonomic-level groups like phylum, class, family, etc are arbitrary with respect to phylogenetic time, and thus under trait evolution models like Brownian motion, do not correctly weight the correlation between taxa when accounting for shared phylogenetic history. Additionally, even within a clade, there will be dependence between certain subclades that clade-level averages will not account for; this the same problem of shared history leading to correlation structure in the residuals, just at sub-clade level. Finally, clade-level averages do not make efficient use of the data; this can be seen in Supplementary Figure S5, as I do not have data points for many phyla to calculate a clade-level regression. The phylogenetic mixed-effects model I use solves these issues directly. I do agree with this reviewer’s point that there may be process heterogeneity among different groups; the node-height tests seem to indicate that this could be occurring. However, fitting such models that could account for process heterogeneity is an endeavour outside the scope of this present work.

In summary, I find the ideas of clade level analyses and of using phylogenetic comparative methods (PCMs) to look at census population size (and possibly diversity levels) promising. For example, as the author alludes to in the Discussion (bottom of P. 13), PCMs may be informative about the hypothesis that species with large census sizes have a greater rate of speciation. Yet I find the current analyses difficult to interpret.

The best way to frame these analyses, as Reviewer 4 suggested, is as a way to deal with residual structure. I have reworded this around lines 231-235.

As justification for using selection parameters inferred in D. melanogaster, the author argues that this is a "generous" assumption in that the effects of linked selection in this species are on the high end. One issue with this argument is that among reasons for the strong effects in D. melanogaster is its short genetic map length. This is not a substantial caveat, given that the analysis is meant as an illustration and it can be resolved by using appropriate wording. Perhaps more troubling is that the author's estimate of the reduction in diversity level in D. melanogaster is much greater than the reduction estimated in the inference that he relies on (several orders of magnitude and less than

one, respectively). This discrepancy is mentioned but should probably be addressed more substantially.

I have added a few sentences to address this good point raised by this reviewer (lines 402). However, I do not see how this is necessarily troubling; because the large predicted reduction in diversity is a result of using Nc in Equation (1), rather than an Ne based on π_0. For *Drosophlia melanogaster,* Ne ~10^6, whereas Nc ~10^{15}. By using Nc, we are seeing whether there is any chance that the linked selection model is capable of reducing π down to observed levels; indeed for large N_c species it is possible (because the draft barrier is reached), but this does not match the previously estimated reductions. Furthermore, this does not explain the observed levels of π for mid-Nc species (which have longer map lengths).

The results of the analysis are intriguing. The effects of linked selection `shrink' the ~13 orders of magnitude of census population sizes to ~3 orders of magnitude of diversity levels. This massive effect is largely due to the genetic draft (Gillespie 2001) and to a lesser extent to the decrease in map length with increasing census size: when the census population size becomes very large (Nc~10^9) and coalescence rates due to genetic drift decrease accordingly (~1/2Nc), coalescence rates due to sweeps, which increase owing to the smaller map lengths (and would otherwise remain constant), become dominant. In hindsight this is quite intuitive and aligns with Gillespie's original argument, but this is in hindsight, and using this argument in conjunction with data, specifically with census population size and map length estimates, is novel.

Thank you for this kind feedback.

As the author points out, the resulting relationship between diversity levels and census population sizes does not fit the data well. Notably, predicted diversity levels are too high in the intermediate range of census population sizes. Nonetheless, their analysis suggests that linked selection may play a much greater role than previous studies suggested (i.e., the analyses of Corbett-Detig et al. (2015) and Coop (2016) suggests that it cannot account for more than 1 order of magnitude). Maybe the poor fit is due to the importance of other factors (e.g., bottlenecks) in species with intermediate census population sizes?

I agree that bottlenecks are a likely explanation (though I think this is perhaps related to macroevolutionary processes, see lines 474-482). I have added a few sentences about how my work strengthens Coop (2016)'s findings and mentioned demographic fluctuations on (lines 517-524).

I also wonder whether the potential role of linked selection may be clearer if the different effects are shown separately, and perhaps with less reliance on the estimates from D. melanogaster. Namely, the effects of background selection can be shown for a few different values of Udel, e.g., between 0.3-3 (this range seems plausible based on many estimates). They can be shown both accounting and not accounting for the relationship

between map length and census size. Similarly, the effect of sweeps can be shown for several values of corresponding parameters, and perhaps even for different models for how the number of beneficial substitutions varies with census size (see Gillespie's work to that effect). I believe that such illustrations will be fairly intuitive and less restrictive.

XXX
Reviewer #1 (Recommendations for the authors):

Additional questions/suggestions/comments:

1) Substantial streamlining and editing would make the manuscript much clearer.

Thank you, following this and other feedback, I have edited some unclear sections.

2) The abstract is way too long: not on the intended resolution.

I have reduced the abstract significantly, to 200 words (the limit of eLife).

3) If you aim for a broad readership, you may consider having the background be less of a historical review (without sacrificing scholarship). Also, no need to recap the historical review at the beginning of the Discussion.
I have removed the recap at the beginning of the discussion. I have not cut down the historical context as, (1) I believe such a mini review (in the spirit of Coop and Ralph, 2012) is needed, and (2) other reviewers and readers did seem to like this aspect.

4) You recap what you do at length several times (e.g., L 130-156), which is repetitive.
I am not sure I follow this comment — this is the first time I describe what I am doing in the manuscript, which seems like an important part of the introduction?

5) It would be clearer to use consistent terminology, e.g., instead of "enigma", "anomaly", "paradox" and "explanation", "resolution"...
I have removed some synonyms for clarity.

6) The writing switches between acknowledging that several factors are plausibly at work and seeking "a solution".

I agree that “explanation” or “resolution” are better terms, and have made these changes.

7) It is claimed that "an ordinary least squares relationship on a log-log scale fits the data well" but I did not find a quantification to this effect. Namely, what proportion of the variance does it explain? Moreover, it is claimed that this relationship is homoscedastic. Can that be quantified as well? From looking at the figure it seems that the regression may explain ~1 of the ~3 orders of magnitude in diversity levels and the residuals explain ~2. It would also be helpful to say what we learn from this fit that we did not learn from

staring at the plot. Does one (or more) of the potential explanations for Lewontin's paradox posit a log-log relationship?

The log-log relationship is used because of heteroscedasticity on a linear-log scale, and because both axes vary over several orders of magnitude. I do not think it is too surprising that π varies a few orders of magnitude, since Nc varies over so many (and Ne a fraction of that). Regarding the proportion of variance, I have included an R^2 in the paper (the adjusted R^2 is about 0.25). However, it should be noted that R^2 has numerous statistical issues (see p. 181-182 of Cosma Schalizi's *The Truth about Linear Regression,* http://www.stat.cmu.edu/~cshalizi/TALR/TALR.pdf). Likewise, I don't think a formal test of heteroscedasticity is particularly helpful; a plot of the residuals versus x values is convincing enough (though I think unnecessary to include in the supplementary materials).

Minor comments (not exhaustive):

• L. 41: per bp per generation…

Fixed.

• Not sure a paradox can be quantified or have a scale (L. 173). Also, I think that "explaining a paradox" means to explain the puzzle not to resolve it.

The way I try to approach this is that Lewontin's Paradox is the shortfall between predicted levels of diversity and the expected levels if Ne = Nc. Around line 122, I discuss how the scale of the effect found by Corbett-Detig et al. is insufficient, as shown by Coop (2016).

• L. 85: "fecund"?

I am not sure I understand this comment; "highly fecund" is a common phrase, e.g Einum and Fleming (2000) and Hedgecock and Pudovkin (2011).

• L. 89: Why is the coalescent model used relevant to the question at hand?

I mention this because another reader suggested I point out these require non-Kingman coalescents.

• L. 113: "Other selection models cannot alone decouple...".

I have reworded this so it is clearer.

• L. 115: "proportional to".

I have made this change.

• L. 121: "and body size".

Fixed.

• L. 122: "They argued that this shows..."

Fixed.

• L. 136: "to provide"?

Fixed.

• L. 138: "Past work has ignored"?

• L. 139-40: "I account for this by using a synthetic..."?

I have reworded this section.

• L. 148 and elsewhere: "generously" is not the right word. You probably mean conservative, and this needs to be qualified.

I agree this wa unclear, so I have reworded it.

• L 215 and elsewhere: "with non-missing data" -> "without missing data"?

Fixed.

• L 216: "non-phylogenetic regression" -> "regression without correcting for..."?

I have reworded this.

• Maybe "phylogenetic dependence" rather than "non-independence", unless I am unfamiliar with specific jargon…

I think this is jargon; the term phylogenetic non-independence is more widely used.

• L. 439-40: "In contrast to demographic models, models of linked selection have comparatively fewer parameters and more readily permit rough estimate of diversity reductions across taxa." Is this obvious? Maybe there are some summaries that largely incorporate the relevant aspects of demography?

I agree; here I am trying to say that in terms of the complexity of possible demographic histories, the parameter space is massive. Seeing what aspects of demographic history are relevant to diversity to parameterize simpler models is a goal of future work.

• L 464: "giving linked selection the best possible chance...". Phrasing aside, this is not obvious to me. Perhaps certain functional dependencies between selection parameters and census and map sizes will yield better fits?

I have reworded this.

• L. 475: The Elyashiv et al. estimates actually apply to other sweeps if they result in substitutions (see their supplement).

This is a great point; I have added a sentence about this.

• L 501: No Barton (2000) in refs.

Fixed.

• L. 507: Not sure a paradox can show.

Fixed.

• L. 513: "tease apart"?

Fixed.

• L. 559: "surprisingly consistent relationship". Not sure I understand this statement or its basis?

This is described on lines 192-194; the relationship on a log-log scale has few outliers and is homoscedastic. For how variable these species are in their ecologies, etc the relationship is relatively clear.

• L. 629: Not sure "subset" can be a verb. More generally this sentence…

In this context, subset is a known, e.g. referring to the subset per phylum.

• L. 709/717: new paragraphs?

Fixed.

Response to Reviewer 2

My first and maybe most important comment is that the introduction, discussion and overall writing of the manuscript are really excellent. This might be the most lucid, extensive, balanced overview of Lewontin's paradox and the associated literature I've ever read.

Thank you for this feedback — this was one goal of the paper, as there are not any papers to my knowledge which cover the full context of Lewontin's Paradox.

My second comment, somehow counterbalancing the first one, is that the major point made here, that linked selection alone cannot explain Lewontin's paradox, has been made before, e.g. by Coop (2016) and Ellegren & Galtier (2016) commenting on Corbett-Detig et al (2015). The material presented here substantiates this point further, but is perhaps not a major advance per se, so that the manuscript lies somewhere between a review and research article.

I disagree with this point. This paper has several novel analyses. First, to date, there have not been any studies that directly quantify the relationship between genomic estimates of pairwise diversity and census population size. Previous analyses such as Nei and Gaur (1984) and Soulé (1976) had far fewer taxa and allozyme-based measures of heterozygosity. Second, neither of these analyses had used PCMs with a time-calibrated phylogeny to account for, and quantify shared phylogenetic history (more on this point below), nor did they explore these traits through a macroevolutionary lens as the present study. This analysis revealed the high degree of phylogenetic signal, and evidence of possible rate shifts in trait evolution (Figure 3C). Third, this paper finds and quantifies an undocumented relationship between census size and recombination map length (a relationship between body size and map length had been observed, e.g. Burt and Bell, 1987). This is an important covariant with population size, which suggests that linked selection may be stronger in high Nc species in part just due to their smaller map lengths (lines 328-338). This finding has not been made in previous papers like Coop (2016) and Ellegren and Galtier (2016). Finally, no previous work on Lewontin's Paradox has estimated Nc and addressed the question: is linked selection even capable of explaining the shortfall between expected and predicted levels of diversity? Figure 1 of Coop (2016) shows a hypothetical decoupled relationship, and does not consider the relationship between map length and census size. My analysis shows that while it is possible for linked selection to decouple diversity from census size for high-Nc species, the census size - map length relationship is such that current models of linked selection cannot fit the observed π-Nc relationship (Figure 4B). This is a finding that has not been made in Coop (2016) nor Ellegren and Galtier (2016), since (1) these works did not have Nc estimates, and (2) the census size - map length relationship had not been quantified.

I have a few additional, more specific comments below. I think this is a great addition to the existing literature, which clarifies and synthetizes many aspects of a complex question.

1. Phylogenetic inertia

I am not sure I get the point of the phylogenetic inertia analysis. It seems to be intended as a response to Lynch 2011, who, responding to a criticism by Whitney & Garland, stated that the coalescence time is not inherited across the phylogeny. That quote from

Lynch is mentioned several times, and as a motivation for performing this analysis. Yet the result reported here, i.e., that pi has some phylogenetic inertia, does not seem to contradict this specific statement, for at least two reasons. First pi might have some inertia via inertia on the mutation rate, not on coalescence time. Secondly, pi might have some inertia because it is in part determined by traits that have some inertia, such has body mass for instance. The text insightfully discusses these aspects (l399-407), but honestly I do think that this analysis invalidates Lynch's (somewhat trivial) point that coalescence time is not a trait that can be inherited. I still agree that the analysis is worth doing and publishing, but I would suggest putting less emphasis on the Garland/Lynch controversy. Also it might be fair to mention that Leffler et al (2012) and Romiguier et al. (2014) did attempt to correct for phylogenetic inertia when correlating pi to various traits, although they did not analyse the phylogenetic effect as thoroughly as it is done here.

I have changed the wording on lines 231-235, and lines 149-157, as I agree the analysis is not only a response to Lynch and has merit on its own. Additionally, I agree that I should highlight that Leffler et al (2012) and Romiguier et al. (2014) both do analyses to check for phylogenetic inertia, and have added this around line 389. Because the audience of this paper could include those in population genetics and phylogenetics, I have opted to keep the discussion of the Whitney and Garland, and Lynch disagreement around line 451. I believe mentioning that this is an unresolved dispute within population genetics frames my analysis well. It also gives an opportunity to clear up some misconceptions about PCMs and frame the analysis I have conducted.

I agree with this reviewer that the cause of phylogenetic signal in π may be inertia in the mutation rates, population sizes, or the reproductive systems. Lynch (2011) also makes this point (p. 1), but argues this is not an issue in his prior analyses because there is fast turnover in the genes that may be responsible for mutation rate, and that the sampling noise in π estimates is high. However, this latter point is not something he directly tests, but I directly quantify by estimating λ. In particular, λ is looking at the fraction of variance in the residuals attributable to phylogenetic shared history. Because I do not have multiple estimates of π per species, I cannot partition the variance into sampling variance and evolution on the tips, but this should not impact these results.

Finally, I think there is a lot of confusion too about whether PCMs are needed if one or both variables in a regression are free of phylogenetic signal, and if phylogenetic signal is dependent on the trait evolution rate. To clear up some misconceptions, I have expanded the discussion around these points on lines 468-473.

2. Range effect

I was surprised to read that species range alone has a significant effect on pi. The reason is that I suspected species range varied at a shorter time scale than coalescence time - e.g. think of what ranges were 20,000 years ago, when pi was probably, I thought, very similar to current pi; maybe worth discussing?

I agree this is interesting and future work could look at the interaction between diversity, ranges, and whether a species resides in a temperate or tropical region (which could be a proxy for how much ranges have changed in the recent past, e.g. due to glaciations). I agree this would be an interesting point to discuss, but given other reviewers were asking that the manuscript be shorter, I did not include this point.

3. IUCN categories

I found the result that endangered species have a lower estimated Nc and a lower pi than non-endangered species a bit trivial, knowing that lare body sized vertebrates are typically more threatened, and more of concern, than small body sized invertebrates. What would be more relevant to conservation biology is an analysis that controls for body size, e.g., are endangered large mammals less polymorphic than non-endangered large mammals. There is a fairly large amount of literature on this topic.

Another reviewer had a similar concern; I have moved this figure to Supplementary Materials.

4. The Methods section (l580-581) states that map length data are available in 41 species, but figure 5A shows a relationship with 131 data points; some clarification needed here

Thank you for pointing this out — this was an error. I have fixed it.

Reviewer #2 (Recommendations for the authors):

I would love to see a revised version of this piece published.

Response to Reviewer 3

I wasn't so convinced that the assessment of phylogenetic inertia (Lambda>0) really provides a way to assess Lynch's argument that coalescent times are too short to have a phylogenetic effect. For reasons outlined by the author in the discussion, it could well be that any phylogenetic inertia signal is due to inertia of life history traits correlated with effective population size rather than with diversity itself. The discussion raises this important point, but I think leaves us with the difficulty of really assessing how important that phylogenetic correction really is: if diversity has no direct phylogenetic non-independence, I am a bit unsure how much we have learned through this analysis alone (i.e. what is lambda telling us), without an explicit assessment of how often divergence times may actually truly be on the same order as coalescent times.

I agree that this is a complex issue; this relates to deeper points in the phylogenetic comparative methods raised by Westoby et al. (1995) and Uyeda et al. (2018). Reviewer #1 raised similar concerns (my response to those points is relevant to this point as well). Overall, the high

phylogenetic signal suggests that there is phylogenetic structure in the residuals, which is a violation of the assumption of independently and identically distributed (IID) residuals under the regression model. As noted in the replies to other reviewers, having phylogenetic structure in the Y or X variables alone is not a violation of the regression model, and not a problem (e.g. Revell, 2008). I have included more discussion on these points (lines 440-474).

That said, I think it's a very open question whether diversity actually has phylogenetic independence because of short split times relative to effective population sizes. The author mentions the possible effect of large Ne on causing this to be violated; but I also wondered whether many of the small Nc species are still retaining a fair bit of ancestral polymorphism, further homogenizing diversity levels.

This is a good point, and is indeed analogous to a point I raise in the discussion (around lines 474-482): if there is some interaction between the speciation process and population size, new small Nc species that have branched from large cosmopolitan species could have higher than levels of diversity. This could act to flatten out the diversity-census size relationship from both ends. However, following another reviewer's feedback that the manuscript is long, I have not added this additional point.

Overall a number of possible explanations (such as the effect of variable selected site densities, and variable recombination) were raised, and rather quickly rejected as 'unlikely to explain the qualitative patterns'. In a number of cases these statements were fairly brief, and I wondered whether in aggregate how likely a combination of these COULD explain the patterns. Looking at Figure 5B, it seems like the major effect of phylogeny (or correlated life history) is also apparent for the discrepancy between observed and predicted diversity- Chordates seem to have the largest discrepancy. With that in mind, I do wonder whether some feature of genome structure in Cordates, including a combination of the effects discussed in the paper that could account for the discrepancy (e.g. the effects of variable recombination rates/genome size and functional densities, variation in mutation rates, etc.) could collectively account for the paradox, even though individually the author rules them out as being able to explain the 'qualitative pattern'. Could the genome structure of chordates lead to a major difference in linked selection that's unaccounted for here? Mei et al (2018) (American Journal of Botany, Volume 105, Issue 1, p1-124) argued that species with larger genomes have greater 'functional space', implying a greater deleterious mutation rate in species with larger genomes. This could potentially be a factor driving those Chordates with intermediate Nc values furthest below the predicted line?

While these are very interesting points, I do not think I can address them with the current data. My dataset does not contain genome annotation level data necessary to address this hypothesis well.

Reviewer #3 (Recommendations for the authors):

My main suggestion is to expand a little more the arguments for why the author feels particular factors are unlikely to explain the patterns qualitatively, and to touch on some of the alternative explanations raised in the public review.

This is useful feedback that other reviewers have voiced as well. I have expanded on how my results connection to that of Coop (2016) on lines 517-524, as well as addressed some more points about the need for PCMs in the discussion (lines 440-473).

Response to Reviewer 4

First, in phylogenetic comparative methods (PCMs) there has been a persistent confusion as to what phylogenetic signal is relevant -- when applying a phylogenetic generalized linear model with a phylogenetically structured residual structure (which the author does here), one is estimating the phylogenetic structure in the errors and not the traits themselves. The comparative analysis are well-done and properly interpreted but at some points in the text, particularly when addressing Lynch's conjecture that PCMs are irrelevant for coalescent times and comments/analysis on the appropriateness of Brownian motion as a model of evolution, that there is some conceptual slippage and I suggest that author take a close look and make sure their language is consistent. Strictly speaking the PGLM approach doesn't assume that the underlying traits are purely BM -- only that the phylogenetic component of the error model is Brownian. As such running the node-height test on the both the predictors and the response variable separately -- while interesting and informative about the phylogenetic patterns in the data (including the shift points you have observed) isn't really a test of the assumptions of the phylogenetic regression model. It is at least theoretically plausible (if not biologically) that both Y and X have phylogenetic structure but that the estimated lambda = 0 (if for instance, Y and X were perfectly correlated because changes in Y were only the result of changes in X). To be clear, I am fine with the PGLM analysis you've done and with the node-height test; I just don't think that the latter justifies the former.

Thank you for the helpful feedback about PCMs. I agree that this was unclear in how I framed the node-height test, and have reworded this on lines 279-280.

One note about the ancestral character reconstruction: I think it is a fine visualization and realize you didn't put too much emphasis on it but strictly speaking the ASR's were done under a constant process model and therefore they wouldn't provide evidence for (a probably very real shift) between phyla. I think it was a good idea to run the analyses on the clade specific trees (particularly given how deep and uncertain the branches dividing the phyla are) but I just don't think you could have gotten there from the ASR.

This is a good point; I have reworded this section (lines 262-264).

I am not convinced that the IUCN RedList analysis helps that much here and in my view, you might consider dropping this from the main text. This is for two reasons: 1) species

may be of conservation concern both because they have low abundance in general and/or that their abundance is known to have experienced a recent decline -- distinguishing these two scenarios is impossible to do with the data at hand; and 2) there is of course a huge taxonomic bias in which species are considered; I don't think we can infer anything ecologically relevant from whether a species is listed on the RedList or not (as you suggest regarding the lynx, wolverine, and Massasauga rattlesnake) except that people care about it.

I agree with this point, and have moved this figure to Supplementary Materials

This is not really a weakness but I find it notable that recombination map length is correlated with body size. I realize this is old news but I was left really curious as to a) why such a relationship exists; and b) whether the mechanism that generates this might help explain some of the patterns you've observed. I would be keen to read a bit more discussion on this point.

This is a great point, and I have added a bit on lines 372-375 making it clearer that this is consistent with, but not evidence of, the hypothesis that genome size increases as Ne decreases due to non-adaptive processes (e.g. Lynch and Connery, 2003). However, this is also consistent with the hypothesis that map lengths are adaptively longer to more efficiently select against deleterious alleles (Roze, 2021).

Reviewer #4 (Recommendations for the authors):

Just a few minor points:

line 45 -- out of historical interests, it might be worth mentioning explicitly how well Lewontin's original estimates (that were the basis for his arguments) compare to modern estimates; i.e., did he overestimate or underestimate the amount of diversity.

I agree this would be interesting, but I hesitate to make this comparison in the text given the difference between (1) allozyme-based estimates of heterozygosity and the genomic estimates of pairwise diversity and (2) the different number of samples in these studies. Lewontin found about a factor of four difference in heterozygosity among taxa, compared to the two or so orders of magnitude I find here.

line 82 - the parenthetical in this sentence ("equilibrium") was hard to understand and took carefully reading to get the point. Consider expanding this to make your point more explicit.

I have clarified this.

line 84 - marine animals is extremely vague -- could you be more precise with some examples or delineate what types of organisms you are referring to (presumably not marine mammals or sharks...)

I agree with this point, but for the sake of clarity it is hard to be more specific, since many marine animals have these modes of reproduction. In looking for a way to clarify this without excessive examples, I have found that other others use this same phrase (e.g. Eldon and Wakely, 2006 say “marine organisms” and Hedgecock and Pudovkin, 2011 use “highly fecund marine animals”)

line 113 -118 - could you unpack this idea a bit ("Other selection models..) Had trouble following what you meant. Also just say "all else being equal"; ceteris paribus is IMO unnecessarily pedantic.

I have reworded this so it is clearer.

line 148 - you use the phrase "generously estimated" here and elsewhere. I think I understand from the context what you mean but I think it is worth being more explicit here
I have removed this phrase.

line 269 - i don't think this is important for explaining any patterns but it is also worth noting that the divergence time estimates are probably more uncertain that the estimates of the contrasts and are probably systematically underestimated deeper back in the tree; these could also create a negative relationship in these plots.

Thank you, this is a very interesting point — I have mentioned this (line 307).

line 288 - Please be explicit about what these parameters are
These parameters were introduced on line 48.

line 322 - Can you provide any evidence that people often think this

I have added a citation to Ohta (1992) here, which is a lovely review of her nearly neutral theory.

Again, I really loved this paper. Fantastic work. Hope my comments were helpful.