Dear Dave,

I've had a read through the paper, and I think it represents a really big step forwards—toward understanding both the true diversity history of diatoms, and where Rabosky & Sorhannus went wrong. Intuitively it makes a lot of sense that sampling an uneven community versus an even one with similar effort will recover less diversity in the uneven community—even if the true diversities are actually the same.

Here are a few thoughts that occurred to me as I read through the paper. They're not terribly well organized, but I hope they're of some use.

From the Material and Methods section:

> *Although such compilations are not immune to biases due to sampling intensity (Alroy 2001),* ***the diversity values are less directly controlled by numbers of samples than occurrence databases*** *(Supplement, figures S1-S4), and the data has been filtered by experts for taxonomic and stratigraphic error (Lazarus in press).*

The statement I've put in bold is a pretty strong one and fairly important for the paper. First off, although I understand you're saying that the BDR curve is more reliable because it isn't as unevenly sampled as the Neptune database, the chosen wording makes it sound like you're suggesting compilations from the literature are *per se* less susceptible to sampling problems than occurrence databases. This I would take issue with. My sense is that any data set of species ranges will be affected by sampling intensity, and I would think the major difference between occurrence databases and literature compilations is that the latter are *missing* information on sampling intensity. (Just because that information is not there doesn't mean it's not just as important.)

Anyhow, if I've understood correctly that the suggestion is that the BDR data are more evenly sampled through time than the Neptune data, and that this is what you're showing in the supplementary figures S1-S4, I think these figures (or some other representation of that evidence) are sufficiently important that they might like to see the limelight of the main text rather than the obscurity of the supplement!

About the supplementary plots: I think I see your point here that in figures S1 and S2 the two lines are much more closely correlated than in S3 and S4. However, it's a bit of an apples-and-oranges comparison, because your metric of sampling effort is # of occurrences in NSB and # of publications in BDR. Not to say that it's true, but I think it wouldn't be hard to imagine a scenario where the poorer fit to diversity in BDR is caused entirely by variations in the sampling intensity (# of occurrences) used in each of the references. I'm not sure how you can get around this, because of course your BDR data doesn't have occurrence information. The best thing I could suggest for a more convincing comparison is to plot # of references for the NSB data instead of using occurrences—then you'd really be comparing apples to apples. (I assume there is a publications table in Neptune?)

From the Subsampling… section:

> *Diversity has two components - total, and relative.*

Completely agree with the sentiment, but the terminology here seems unfamiliar. It may just be that I haven't read widely enough, but I have heard the second one referred to as evenness, often described by RADs (rank-order abundance distributions, i.e. analogous to your Figure 3).

> *This is not true of the Cenozic Neptune diatom data (Figure 3).*

I'm a tiny bit confused about this part, and I'm not sure if it's a problem—my understanding is that, strictly speaking, the canonical sampling-standardization methods assume a constant abundance distribution through time in the underlying population. What you are looking at is the distribution of occurrences across taxa in the data, which might be subject to biases. That said, incomplete sampling of the underlying population is more likely to make the data look uneven (I think—that's just an intuition!!), so this would imply the opposite effect on the abundance distributions with time than is seen in your Figure 3. So I think that this is actually OK.

> *The effect of changing relative skewness in occurrences is shown by 'A80': the number of species in* ***abundance ranked*** *species lists needed to reach 80% of the total occurrence data in a bin*

Minor point—later on you call them occurrence ranked, which I think is more accurate than abundance ranked.

> *'Fraction Div-80': the fraction of total diversity found in occurrence ranked species data at the 80% occurrence level*

This part I get, and is really convincing—the fraction of all the observed species in a time bin needed to make up 80% of all the occurrences in that time bin goes down (way down!) over time. To the point that in the last few million years, a quarter of the species make up 80% of the occurrences.

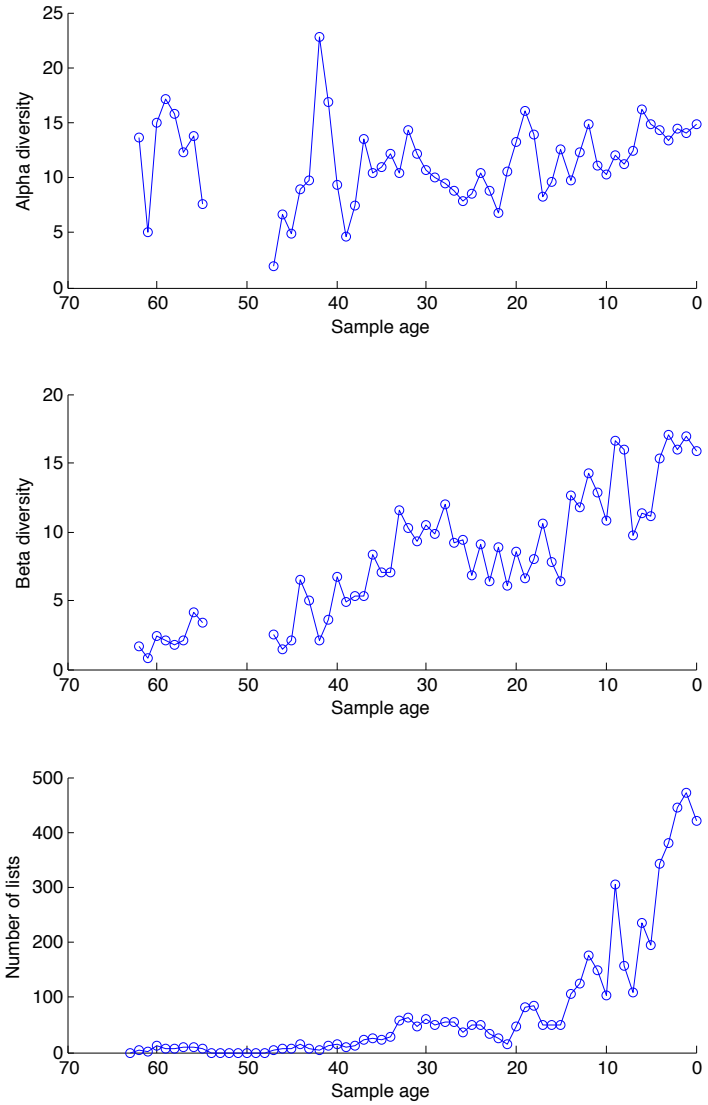
> *'A80': the number of species in abundance ranked species lists needed to reach 80% of the total occurrence data in a bin*

So this is basically the same, except it's the raw number of species needed to make up 80% of occurrences, rather than the proportion. It's really remarkable how similar this trajectory is to the NRS curve!

> *The changing pattern of ubiquity could be due to…artifacts of data collection.*

(I have heard it called evenness or dominance rather than ubiquity). It might be worth addressing this further—because if that's true, you'd want to be sure that whatever data collection problem causes this change in ubiquity doesn't also affect your BDR curve. I think this is important because even though changes in the evenness within the data set will affect subsampling and not the range-through methods you're pushing in this paper, if the differences in the evenness in the data are indicative of difference in underlying population evenness, this could be a source of bias in your observed range-through diversities, too.

Another thing to consider here whether this evenness/dominance pattern is the accumulation of local changes in evenness (how local communities are constructed) versus changes in the geographic distribution of diversity (beta diversity, basically). I've looked at this a little bit and it seems that, in the Neptune data, average list length stays about constant through time, while the beta diversity increases. Here's a plot:



As Bush et al 2004 document in their paper, this makes for real problems in subsampling. If diversity is becoming increasingly partitioned (and you mention this in the MS, talking about provincialism), this presents an additional (and different) challenge to subsampling algorithms. As you subsample very strongly, you will (statistically) begin to miss some biogeographic provinces altogether, and thus underestimate true diversity. You can attempt to address this by tweaking the subsampling algorithm to draw occurrences/lists in a way that's not geographically blind but makes it more likely to draw subsamples from geographically disparate localities. This is all stuff I'm hoping to work on in my own project on diatom diversity!

> *Data collection methods however may also be systematically different for different geologic time intervals. Most Neogene diatom (and indeed most microfossil) occurrence data are collected for biostratigraphic purposes and*

report primarily a small number of well established biostratigraphic marker species (Lazarus in press). Paleogene diatom studies by contrast have tended to be of a more exploratory nature and report (more uniformly) occurrences for many species whose stratigraphic utility is not well established.

I think this is a really good point that might deserve further attention. If this is true, what are the expected effects on the observed pattern, and what does it suggest for true underlying diversity? Let's assume for example that the relative abundance distributions of standing diatom assemblages were constant over the Cenozoic, but there was a change in data collection from broad-survey in the Paleogene to mostly-biostratigraphic-taxa in the Neogene. We'd expect to see in our samples more even RADs in the Paleogene and more uneven samples in the Neogene. [I think this is the point you're making!] If that's the case, the problem is not really with the subsampling exercise (or at least that's not the only problem!), but with the data collection—a point you make well in your last paper… However, if that's the case, I think you'd need to demonstrate that this doesn't also apply in the BDR data set!

From the "Biases in existing…" section:

Thus at the present time there is not enough appropriate data to permit a robust quantitative reconstruction of Cenozoic diatom diversity history.

I don't think the Chicago database crowd would agree with you there. I think they would argue (if they were reviewing this MS!) that it's not that there aren't enough data, but that the data are biased and that bias needs to be corrected for. You show very convincingly in this paper that the *way* in which they try to correct for that bias is wrong—because it throws out *too much* data in trying to correct for the bias, since abundance distributions vary over time. However, my own hunch here is that simply because the method of correcting for the bias is wrong, doesn't mean it's impossible to correct for it, or that it's OK to ignore it—perhaps we just need a better method to do it.

From the "Evidence for true increase…" section:

First, if total diversity were truely constant, the shift in ubiquity observed in the Neptune data in the early Neogene would result in a decline in subsampled diversity, when in fact a modest diversity increase is calculated (NRS curve, Figure 1). This suggests that the underlying diversity has increased by more than the negatively biased subsampling curve indicates.

OK. So you're saying that if true diversity had stayed constant over time, but the relative abundance distribution of that diversity had changed in the way documented by the Neptune data, subsampling should recover a decline in diversity because of the bias against time intervals with uneven RADs. However, since we observe an increase, it can't be true that diversity stayed constant, and it must have increased.

I agree with this in principle, but I think the crucial question then becomes: what are the relative magnitudes of the bias in sampling intensity versus the bias in the subsampling method? You're showing that there's a bias in the subsampling, so the answer provided by that approach is wrong; but Rabosky and Sorhannus show that there's a bias in the raw data, so that answer is wrong, too. The question now is, in a sense, which is less wrong?! I don't think it's quite sufficient to argue that because the subsampling is wrong, the raw data must then be right, because that still doesn't address the problem of uneven sampling.

One way to investigate this, I think, would be to try this out in a model—make up a dataset where you have a known "true" diversity and a known abundance distribution varying in a known way over time. Now subject it to the levels of uneven sampling documented in Neptune and try to recover the true diversity in both ways (range-through versus sampling standardization)—which comes closer to the true answer? I don't think I'd be able to answer that question without going through that exercise.

Second, all non-subsampled estimates of global diatom diversity, which are not affected by changes in ubiquity, suggest an increase in diversity, particularly in the mid Neogene-Recent, whether computed from the Neptune occurrences database or independent compilations from the primary literature, using range-through or within-bin methods.

I'm not sure I agree with the first statement *(non-subsampled estimates of global diatom diversity are not affected by changes in ubiquity).* The non-subsampled datasets are, in a sense, also subsamples—of the underlying populations. Therefore, they should also be subject to biases introduced by changing abundance distributions. However, I think that works in favor of your argument (I think I tried to say this somewhere above already): as abundance distributions become more uneven through time, this should make total diversity look *less* than what it really was, because it becomes ever harder to sample further out along the tails of the hollow curve. Therefore such a change in relative abundance distribution would actually work to flatten diversity curves, and suggest the true increase in diversity was *greater* than what's suggested by the raw data—not less. (But—and this is a big but!!—this is under the assumption of constant sampling effort, and again we don't know the relative magnitudes of these biases). Again this is just a gut feeling and it would be worth going through a simple modeling exercise to see how bad (or good, for your argument!) that bias could be.

Third, there is substantial positive evidence for increased provincialism and the development, particularly near the Eocene-Oligocene boundary, and in the mid Neogene-Recent, of diverse high latitude diatom floras in both the northern and southern hemispheres (figure 2).

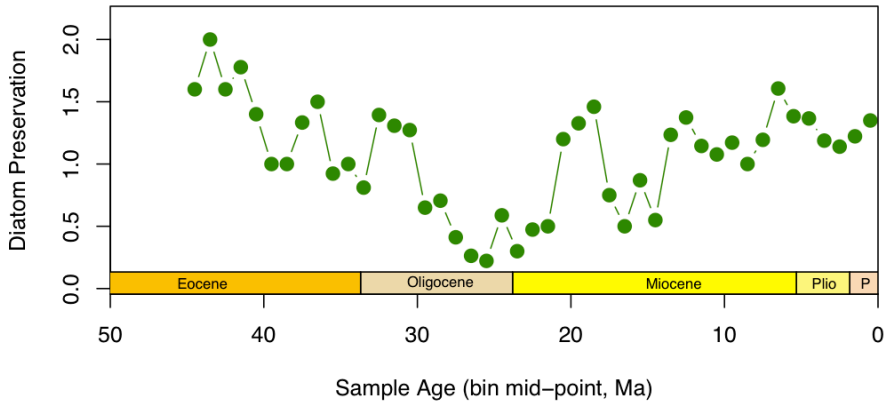
Totally agree with this. However I'm not sure I see right off the bat why figure 2 shows an increase in provincialism. I would think of provincialism essentially as beta diversity—how different floral lists are from one locality (or one ocean basin) to another. So if there was no provinciality at all, I'd expect the diversity of each ocean basin to be the same as the total diversity—basically, the same stuff's found everywhere. As provinciality increases, each ocean basin should make up proportionally less of the total diversity. I'm just not sure I can easily see that in the graph—looks to me like total diversity is always a lot bigger than individual diversities—maybe you can plot the diversity of each ocean basin as a % of total diversity to make that point more persuasive?

The development of provincialism introduces a separate, though similar, bias to the one brought into play by the change in RADs over time. This is the sort of thing Bush et al 2004 talk about. Again, I've mentioned this above. If you have diversity spread out into lots of different provinces, rather than evenly distributed spatially, subsampling is going to start missing diversity because—when strongly subsampled—a few of the provinces and their diversity are going to be thrown out in each subsample. This is clearly a bad thing, and a flaw in the subsampling algorithm that leads us to underestimate diversity of those time intervals with high provincialism. This speaks against trusting the results from such subsampling efforts.

However, I would again hesitate to turn around and thus suggest that the sampling bias should be ignored. Again, we don't know the relative magnitude of the subsampling bias in throwing out good diversity versus the effect of not sampling earlier diversity enough. It could be the case that diversity really went up (and I suspect that you are right and it is true!!), but I don't think it has necessarily been shown that it must be true. Rather than abandoning subsampling altogether I think I would advocate developing a subsampling algorithm that takes provincialism into account and can handle it—that way we could address both biases from the provincialism and from the differences in sampling over time.

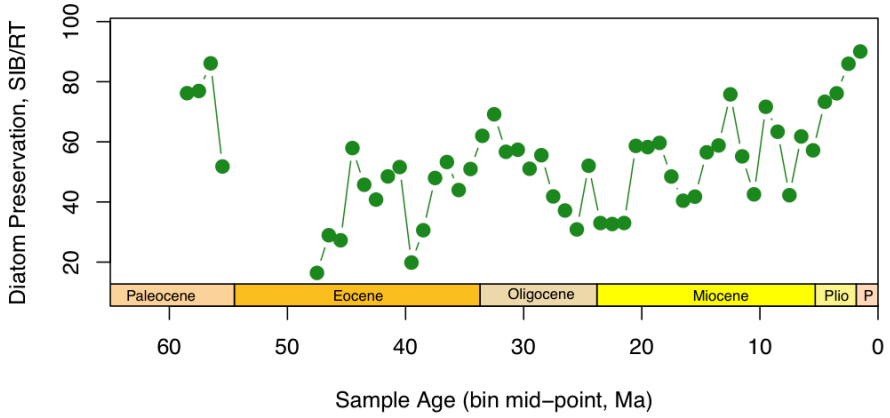
Lastly, preservation bias is thought to have been less strong in the geologic past.

Yes, this should help the argument!! Again it would be nice to see some more specific supporting data. Neptune has some, albeit very qualitative data, that might lend some insight—the G/M/P classification for ODP slides… I took a look at that at one point (G=2, M=1, P=0) and found it interesting:



There does seem to be a bit of a decline over time, but preservation looks particularly crummy the late Oligocene—incidentally also where we have the most major dip in diversity in all of the diversity estimates (RSD and BDR included!).

We should be able to get a similar indicator of what preservation is doing by looking at the Neptune range-through process—i.e., in which time bins are we having to infer the presence of species by range-through because they're not showing up in Neptune? Looking at the ratio of sampled in bin/range-through diversity gives something of a similar answer (apologies for slightly different timescale):



Now, this statistic has some other problems—edge effects that plague range-through (which is one of the reasons range-through fallen so out of fashion with the Chicago crowd), making preservation look U-shaped, really good at the beginning and end, and crummy in the middle… Alroy uses a localized version of this, comparing each time bin only to the one before and the one after, but which solves the edge effect problem… But I won't bore you with endless plots. This is interesting stuff.

In the Supplementary info, "Prevalence of marker species…" section:
This whole section strikes me as a little gold nugget that I would be sad to see relegated to the supplement. I understand that it might be necessary because of the limitations on space in Nature, but this really underscores and supports the point you make in the last paper (Model A vs Model B vs Model C in occurrence data collecting!) and this is a really major issue for the Neptune data. The fact that you have actually calculated a number here for biostrat/water mass vs. floral documentation taxa is significant—I think you should at least allude to that in the main text!!

In Summary!
I realize I've gone on a lot and I should stop and send this before it's too late and all for nothing. I think it's a great paper and the insight about RADs is a crippling blow to the Rabosky and Sorhannus view of the world! I don't think I'm completely convinced by the evidence for your BDR data being as free from the sampling biases as you suggest they are (at least in the way it's presented right now), and I think there is a lot of room for some simple models to see how bad the various biases might be relative to one another. I think there are also ways of addressing problems you point out and *also* addressing variations in sampling bias—e.g. by applying Alroy's new SQ algorithm to the data, and by subsampling with awareness of oceanic provinces. Since these are things I'm actively working on myself and hoping to publish at some point in the not-too-distant future, I won't pressure you too strongly to include that in your MS!! :-)