To the editors of *Nature Ecology and Evolution*,

We greatly appreciate the chance to revise our manuscript, and today we are returning a piece of work that has benefited greatly from the reviewers’ helpful comments. Itemized responses are given below.

Dear Dr. Carlson,   
  
Your manuscript entitled "Applying ecological network theory to re-estimate global viral diversity: host sharing matters" has now been seen by 3 reviewers, whose comments are attached. The reviewers have raised a number of concerns which will need to be addressed before we can offer publication in Nature Ecology & Evolution. We will therefore need to see your responses to the criticisms raised and to some editorial concerns, along with a revised manuscript, before we can reach a final decision regarding publication.   
  
In particular, Referee #2 feels that the presentation requires more clarity to distinguish it as a stand-alone piece of work, rather than a response to Carroll et al in Science. Referee #3 also feels that there are some issues with clarity of presentation, so we suggest re-formatting into our longer Article format (up to 3000 words for Intro+Results+Discussion), to give you room to address these comments.

We appreciate the opportunity to expand to a longer format, and have worked hard to produce a piece of broader relevance, that no longer reads as a direct response to the Carroll *et al.* paper.

Please note that we also do not permit punctuation in titles, so you may wish to revise this now. Titles should be short and informative (ideally no more than 100 characters, including spaces). For example: "Re-estimating global viral diversity based on host sharing" or "Global viral diversity estimates incorporating shared hosts", something like that.

We have updated the title of the manuscript to “Global viral biodiversity estimates accounting for host-sharing patterns.”

Reviewers' comments:   
  
Reviewer #1 (Remarks to the Author):  
  
Review of C. J. Carlson et al. Nature Ecology and Evolution  
  
This manuscript presents a new estimate of global viral diversity in mammals that reduces current estimates based on linear scaling between host and virus diversity by two orders of magnitude. This re-estimation of viral diversity is based on host sharing, applying a power law scaling relationship for host-virus species interaction networks.  
  
I think this is an excellent paper. In the absence of clear theoretical expectations, the authors use a simulation method to re-estimate viral richness, based on iterative resampling, curve-fitting, and extrapolation on bipartite networks. They apply the method to an excellent database that covers 10% of mammal diversity and their 511 catalogued viruses. Their approximation works remarkably well: they slightly overestimate virus diversity (568). Then, they extrapolate to the global host-virus networks, considering both RNA and DNA viruses. The upper bound found is only 2-3 % of previous estimates.

We appreciate the positive feedback!  
  
My only (relatively) major concern with this paper relies on the effects of sampling effort within individual host species on the estimation. The question is simple, but it is not clear in the manuscript: do the 511 catalogued viruses in the database represent the maximal sampling effort where diversity saturates for these hosts? In other words, can the authors provide rarefaction curves where viral diversity saturates, for at least some host species? In Table S1 they provide the degree for different orders, but it is not clear whether the degree depicted corresponds to the saturation phase of the rarefaction curve. If these curves do not saturate, hence, there is a “hidden” diversity not accounted for. If, on the contrary, sampling more individual hosts does not increase viral diversity, then, estimations are correct. In the methods, the authors state “…rarefaction curves were constructed for one or two individual species as a function of sampling effort”. But, do these curves saturate? Usually, in the study of other consumer-resource interaction networks, the number of individuals sampled should be large enough to provide a good estimation of the true degree of a species (see e.g. P. Jordano 2016, Functional Ecology <https://doi.org/10.1111/1365-2435.12763>).

In short, no, the 511 viruses do not represent the maximal sampling effort. This is the intention of using the two-step approach, and of the sampling correction (the 6% rate). We do use rarefaction curves to derive that 6% figure, which we explain in an entirely new standalone methods section addressing that approach. This is what the bat and macaque data are used for, as they are – according to Carroll’s assessment – the only two species with complete enough metagenomic data that those rarefaction curves can be run such that they do saturate, and can be treated as appropriate estimates of total one-host viral diversity estimates. (We also searched for other species during the course of our revision that we might be able to use, but found no appropriate grey data.) Moreover, the association data includes no repeat sampling (only the metagenomics data does), so the estimation procedure described in Jordano’s paper could not be applied here.

However, we found the Jordano reference useful and borrow from its terminology now, when describing the difference between Carroll’s approach and ours; Carroll’s assumes no forbidden links between host and virus groups, whereas ours makes no such assumption; and we cite Jordano there:

“This estimate is the most tenuous in our analysis, but uses much the same logic as the linear extrapolation used by Carroll *et al.*, without making their assumption that every host-virus family association is equally possible. In reality, there are disproportionate associations due to a combination of “forbidden links” (sensu [Jordano]) and non-random coevolutionary diversification.” (L286-290)

The writing is clear and appealing for a general audience. Also, the figures are very nice, simple, and greatly illustrate and complement the message of the paper.  
  
We appreciate the positive feedback!

Reviewer #2 (Remarks to the Author):  
  
In this study the authors estimate global virus diversity for mammals using a non-linear equation fitted as a power law to the number of hosts. The authors show that this method, in comparison with linear estimation used in a previous study earlier this year, provides order of magnitude lower estimation of mammal virus richness.   
  
Overall I find this paper timely, relevant and interesting. I agree with the authors that estimating richness of affiliate species linearly does not make much sense and I also find the estimation by Carroll et al way too high. So, I like the paper (and the fact that the code is already available!), but I do have comments (mostly since things were not explained well).

We appreciate the positive and detailed feedback.  
  
Good luck!!  
  
MAJOR COMMENTS  
  
1. The manuscript reads as a response to the recent study published in Science earlier this year by Carroll et al (ref 1). The language of the current manuscript is brief (too brief), the tone sometimes harsh. The writing style assumes that the reader is familiar with ref 1: it is impossible to understand the results and the introduction before reading the methods, to really understand what ref 1 did. In fact, I needed to read the Supplement of ref 1 to really get a really good idea of what they did. On the other hand, the authors make some broader analysis across network types in Fig 1 and show that estimating affiliate species typically produces a power law. They even give a package to perform this analysis making it broadly accessible. So, I think the editors and authors need to work together to decide if this is a brief communication as a response, or if it stand by its own and uses the Carroll paper as a mere example in which linear estimation has applied consequences. This would require some rewriting and re-formulation of the problem. Does not seem too difficult to do.

We appreciate this feedback, and are grateful for the chance to expand to a full-length paper. In the process of doing so, we have moved the section describing the methods of Carroll *et al.* to the supporting information, which has helped reframe our paper away from reading as a response to Carroll, while also expanding on the necessary information to understand discrepancies between their estimates and ours.

More broadly, we have reorganized and expanded text in the introduction and discussion to highlight the central issue of how diversity is estimated in macroecology, and to contextualize our results within that literature and within the broader virological implications. We are hopeful this has largely solved any outstanding tone issues, and also increased the relevance of our study.

2. I know this is a brief communication but I think the results and the intro are just too brief. I really did not understand from reading this what was the problem and how the analysis was done. I had to read this a few times. What I suggest is to move some description from methods to the intro. For example, the explanation on how Carroll et al calculated virus diversity could easily go to the intro in line 24. And some of the methods for estimation should be explained somehow before the authors “throw” all these results at the reader.

See response to point #1; we have also added more methodological overview to the introduction, to help guide readers through our approach, but did not move text on Carroll to the first paragraph, in an effort to reduce the appearance our paper was directly criticizing Carroll’s. We have also added two tables to the results, which make them much clearer, and remove the overwhelmingly dense numbers from the text.  
  
3. I was missing an explanation on why the analysis was done that way. Specifically, why was there a need to resample? Why don’t just fit the number of viruses to a power law? Why, and in which way is this analysis different than the classical approach of fitting a degree distribution to an exponential/power law/truncated power law (sensu Bascompte et al 2003 Ecol Lett)? Was resampling done just to produce CI? or is there some other logic behind it?

**Resampling**: At least some resampling is necessary to generate the curves (this is the method Strona and Fattorini developed independently), but for the confidence intervals we have updated our methodology using a new statistical 95% CI procedure to replace the second layer of subsampling (i.e., we no longer do a top layer of iterations to generate a confidence interval). The new estimates produce slightly wider confidence intervals, and make our approach more reproducible overall.

**Bascompte**: There is a tremendous amount of literature centered around fitting degree distributions for empirical and theoretical bipartite networks, but whereas those works investigate the shape of one distribution, here we are describing the pattern of a curve (generated by scaling in edge counts for the total network, not degrees). We suspect there are statistical underpinnings connecting the two, but after searching the network literature, we found no existing analytical solution to any variants of this problem, and so are unable to move past conjecture.

4. Why did the authors choose to fit a power law (that this was previously done by ref 7 is not a good-enough motivation)? A power law has a special property of scale-invariance. Why is scale-invariance important for affiliate diversity? what does it mean biologically for the viral diversity? Also, if we try to fit a different function would that have a worse fit (e.g., a lower AIC in model selection)? Worth trying no?

Given the number of reviewer comments centered around the power “law”, we have added a several-page supplement that includes several new figures and tables, as well as

* A brief discussion of macroecological precedent
* A model selection exercise that shows power laws outperform other simple methods, but there may be more nuanced power laws better describing the curves
* An explanation of how and why we applied the parsimony principle to use the power laws in the main text, and what the consequences of that are for our estimates

That supplement also includes a discussion of the scale-invariance property; as we explain,

“Power laws are widely seen as ‘special'’ in the natural sciences due to their property of self-similarity (scale-invariance), which facilitates extrapolation over several orders of magnitude (as is common, for example, in species area relationship studies). When power laws are recognized in ecology, this is also sometimes taken as evidence of scale-invariance in the underlying biological process. Regardless of whether this is true, we argue it is largely irrelevant in our present case (and our instrumental use of a power law should not be taken as support of this assumption). With ~ 500 mammals sampled of roughly 5,000 species globally, our models are only extrapolating over a single order of magnitude, and we would argue this largely circumvents the deeper question of scale invariance. Moreover, the tendency of the power law to overpredict at higher values as shown in Figure S2 (and the effect of predicting based on 10% of a network, shown in Figure 2) suggests that our estimates can be conservatively interpreted as an upper bound on possible diversity.” (L85-98 in Supp)

5. It is tricky to use the taxonomy we typically use for species on viruses. Designation of viruses to different ‘species’ (L.22) does not make much sense, and this is well known. The influenza virus itself has so many strains each with its own epidemiological properties. What is the taxonomical units of the data set of Olival? How does using ‘species’ as a unit limit our ability to really know what is out there? We should use whole genomes, or OTUs, etc.

We have added a new point to the discussion addressing this point, as well as parallel work unpacking this issue for more virus-tailored (but less ecologically-minded) taxonomic criteria:

“Our broader finding that viral diversity has likely been overestimated is congruent with the limited other literature on the subject. Parallel work focused on phage diversity has used rarefaction curves and the Pacific Ocean Virome metagenomic dataset to suggest that the size of the broader global virome (defined by genetic diversity rather than species counts, which are based in challenging species concepts) may have been overestimated in the early 2000s. [13]”

We ultimately chose to reuse the same species definitions set out in the Olival and Carroll papers so our estimates are directly comparable to previous work, and decided that deeper questions about the nature of species concepts for viruses are beyond the scope of our paper, which largely avoids the finer points of viral taxonomy.

6. It was not clear to me if the viral diversity was estimated for host orders and not species (L. 245)? I guess species because it does not make sense to me to fit to orders. But if that was the case then why?

This was not the case; we only show orders in the network figure to illustrate how broad connections span large phylogenetic distances. We have clarified this in the Discussion, when we discuss the role of sampling in Figure 2.

MINOR:  
  
7. L.28: change ‘severely’ with the exact quantity, or in terms of orders of magnitude. Also try to avoid harsh tone towards Carroll et al in general.

We have made this change, and have softened our tone throughout.  
  
8. L.24-25. Why? Why not considering host sharing a fundamental gap?

We are unsure how to interpret this review, as the original sentence in question reads: “A fundamental gap in such projections is the lack of attention to host sharing patterns.”

9. L32—33. Is this independent estimate a mean for the host species?

Yes; we have now clarified this.  
  
10. L37-39. And? Is this way of correcting the estimates helpful?

We have somewhat reorganized this text to better clarify the problem with the approach, and the fact that problem is most evident in the fact that non-linear scaling implies it is an inadequate approach.

11. L 43-44. Any ref for that? or is it just based on the Fig 1? If only based on the figure then I don’t think it is conclusive evidence.

We have rephrased this sentence to avoid this concern:

“This non-linear scaling between host and affiliate richness observed in host-helminth associations can be reproduced for several types of species interactions (Figure 1).”

We have also more broadly addressed this issue in the model fitting exercises in the supplement.  
  
12. L49: change “that the pattern would” to “to”.

We have made this change.  
  
13. L50-51 Data and computational tools for that kind of analysis exit at least 10 or more years. I don’t agree about that “recently”.

With the viral data presented in the Olival study only made accessible in 2018, it is hard to argue the data to find these types of patterns—at least, for viral diversity—has not recently increased; we have added two citations for this point as well. However, we have removed reference to computational tools:

“….the pattern may be subtle enough at smaller scales to not have been evident without the kind of large network data that is increasingly available in community ecology.”  
  
14. L.51-52: Why? Also see my major comment #4.

See response to major point 4; our new supplement goes into depth exploring our rationale.

15. L66-68. Using the 6.7% is also linear extrapolation (as in Carroll et al). Just saying...

While we agree this is a limitation, and state that as such, in this revision we have added a confidence interval on this rate to better represent our uncertainty, as well as a new section in the methods explaining why this rate is used.  
  
16. I Would include humans in Fig 2d,e. It will be cool to show with which orders humans share more viruses. In that regard, authors are missing an important reference for parasite sharing: Gómez JM, Nunn CL, Verdú M. Centrality in primate-parasite networks reveals the potential for the transmission of emerging infectious diseases to humans. Proc Natl Acad Sci U S A. 2013;110: 7738–7741.

We have added this reference and made this change to the figure, in addition to separating out zoonotic and non-zoonotic, which makes our discussion of the figure much clearer and more relevant to the rest of the paper. We appreciate this helpful suggestion!   
  
17. Why were edge weights proportional to number of species shared? How would another choice change the results? For example, Pilosof et al. (2015 PLOS One) used the jacquard index because two hosts that share similar parasite communities will be more prone to share a new parasite. Is it because you need the weights to fit the curves?

We have also made this change to the figure, which improves visual clarity of sharing patterns pictured. These changes are now explained in the methods as well. (Nothing used in the analysis for Figure 2D enters into estimates of viral diversity)  
  
Reviewer #3 (Remarks to the Author):  
  
This manuscript addresses the question of estimating the global diversity of viruses that use mammals as hosts, and the number that potentially use humans as hosts (i.e. zoonoses). This is highly topical, given the recent launch of the Global Virome Project and similar schemes, which has involved much fanfare about the huge number of undiscovered viruses and the threats they might pose. The manuscript work points out a flawed assumption underpinning those estimates, and then develops and applies a rational approach rooted in ecological principles to make better estimates. The new numbers turn out to be drastically lower, and the authors make arguments for how their approach might be used to target on-going sampling efforts.   
  
I was startled to learn from this manuscript that the existing, well-publicized estimates of global viral diversity were based on the simplistic (and obviously wrong) assumption that host sharing could be ignored. The authors have done a major service to the field in uncovering this error and proposing a reasonable fix. They have also put the virus-host literature in an appropriate ecological context, by comparing the ‘affiliate-host’ relationships for viruses to those observed for four other classes of close species interactions (Fig 1 A-D). For these reasons I am enthusiastic about this study, and believe it merits placement in a high-ranked journal where it will get the attention it deserves.

We appreciate the positive feedback!  
  
With that said, I do believe there are aspects that could be improved, either in substance or presentation. I detail these below.  
  
- The scaling exponents z are reported without uncertainties. This seems odd given that the power-law relationships are at the heart of the study.

We have updated the procedure by which we fit curves to derive 95% CI’s for the point estimates of these parameters, and now we include a table of the exponents for the six main curves we fit, in the supplement. The codependent package has also been updated to directly export point estimates and confidence intervals on ‘b’ and ‘z’ as part of the ‘copredict’ procedure, thanks to this helpful suggestion.

I recognize that the authors used a sub-sampling approach to estimate the uncertainty in total viral richness directly, but still the underlying claims about power-law behaviour should be supported statistically.

- In most or all of the power-law fits to data, the fitted line seems to run high at the upper limit of host diversity (i.e. the right-most part of the curve fit). Is this simply a manifestation of the overestimation problem described in Fig 1E? Or is there something else going on here? Did the authors try any other functional forms to fit the data?

As part of our expanded supplement, we include a figure showing the residuals; address the possibility this pattern is better represented by other functional forms; fit several other examples; and discuss their unsuitability. We still conclude the power law is the most appropriate model but now include several pages in all explaining our decision and justifying it with statistical analysis.

- Uncertainties are reported (as 95% CIs) for many key quantities throughout the study, but they don’t seem to capture all the uncertainty at hand. In particular, there is a major leap in assuming that the viral database they are using (from the Olival et al paper) captures 6.7% of the viral richness for all host groups and all virus groups considered in the paper. This number is reached by comparing the relevant numbers from the viral database to numbers reached by intensive metagenomic screening of two host species. To their credit, the authors acknowledge in the Methods that this number is a ‘back-of-the-envelope estimate’ and is ‘tenuous’. However they also go on to apply this factor to re-scale every richness estimate reported in the manuscript, and it seems to be treated as having no uncertainty at all. While I am sympathetic to the predicament (some number is needed, and this seems to be the best one, and there is no obvious way to accurately capture the uncertainty on it), this approach undermines all the 95% CIs presented for richness estimates. Why bother propagating the small uncertainties from other steps in the process, while ignoring uncertainty in this factor? This only serves to make the uncertainty look smaller than it really is. It would be much better to take a swing at estimating CIs around each estimate that encompass all sources of uncertainty. Or at minimum, there needs to be explicit language about how the stated uncertainty ranges should be interpreted.

We appreciate this point and have focused on this especially in our revision. Using the confidence limits on the rarefaction that Carroll et al. did , we have added confidence bounds to the re-calculated sampling rate (which is now 6.1%, due to a slight difference in how we estimated the total number of viruses: previously we used the total of the rarefaction estimates, which omitted singletons not used in the rarefaction; we have fixed this now). That 95% CI is now propagated through with the other confidence bounds, to accurately represent this uncertainty. As we state below, we also attempted dividing up this rate for DNA and RNA viruses, though we felt this result was unsuitable for the main text. All of these results are included in a new section of the methods called “Correcting for sampling.”

- Also, when the estimates of total zoonotic virus richness are presented on lines 88-89, the authors abandon the CIs and just state the numbers with a preceding ~ (‘approximately’). To me, this read as conceding the point that the uncertainty estimates wouldn’t be reliable anymore, since more factors with unknown uncertainties were used to reach these values. This is not very satisfying, though I am open to arguments about why it’s reasonable. However this approach is unevenly applied, since it seems it should also apply to the estimates relying on the 6.7% number, which also lacks an uncertainty.

We have now added a confidence interval to the 6.7% figure, and carry forward confidence intervals to every value, including our zoonotic estimates (see Tables 1 and 2).   
  
- Why is the 6.7% rate applied to both DNA and RNA viruses? Shouldn’t there be different scaling factors for the two groups?

We attempted a new analysis doing so. However, the sample sizes are so limited in the Olival dataset already that this presented some problems; for example, one of the two species had no DNA viruses in the Olival data. We include this as a supplementary analysis (Table S7) and show that it reduces our estimates, but caution against overinterpretation of this result. As with the rest of the paper, we restricted our main text analysis to one that is cautious and largely an upper bound.

And while I like the idea of performing the analysis separately for DNA and RNA viruses, why stop at DNA vs RNA? Why not break the viruses down further, e.g. to ss vs ds, or viral families?

As stated above, sampling is limited in the Olival data and patchy across different viral groups; splitting our estimates any further would only increase the uncertainty of our results. We hope that future work will continue this further down to these finer taxonomic levels and state this in the Results.  
  
- Some details in the results confuse me. The authors use the ‘zoonotic rates’ from Olival et al (14.1% and 41.7%) to go from estimated total DNA and RNA virus richness to estimated numbers of zoonoses in each class. And they report them with ~ as discussed above. But then when using the 50% estimation method, they give 95% CIs for the zoonotic estimates, and the zoonotic rate for DNA viruses is clearly higher than 14%. At first these two differences made me think they had somehow extrapolated the zoonotic numbers directly. But then I calculated the zoonotic rates from the reported numbers, and found that both the DNA and RNA groups used the 41.7% rate. This appears to be an error. If so, then the later statement on lines 103-109 (about more zoonotic DNA viruses than zoonotic RNA viruses) does not hold. This would be fine, especially since the rationale provided in lines 107-109 seems tenuous.

We appreciate the reviewer’s attention to detail! This was indeed an error, and the numbers have been corrected throughout as part of our new analysis, with the sentences about DNA zoonoses being the majority removed.  
  
- I didn’t get much out of Figure 2D-E. It’s a nice illustration, but there seems to be little information in the degree distribution (all the important nodes are roughly the same size, surprisingly). Nothing about the specific edges jumped out. So really it’s only the ‘node strength’ which differs significantly, and it’s not obvious to me why I should care about that. The text in lines 123-131 seems to imply that this mapping of diversity patterns will somehow help to guide future sampling, but I couldn’t follow the logic (or at least, I couldn’t see how the network in Fig 2 informed it). Can the authors please make their thinking more explicit here?

We have expanded this figure by comparing the zoonotic and non-zoonotic networks and including humans per Reviewer 2’s suggestion, and have also made some graphical tweaks (edges now represent the Jaccard index, and node size is now proportional to the number of viruses sampled in the Olival data for each group), which we hope makes the figure more informative.  
  
- Similarly I could not understand how the authors’ methods can standardize viral richness efforts for sampling effort (line 116), especially since current sampling is likely to be biased in unknown ways, and probably in different ways for different virus and host groups. (See the News & Views commentary about the Olival et al paper for more discussion of these possible biases.)

As mentioned above, we have added an entire section clarifying why we used 6.1% (with new confidence intervals) as our standardization. We also address this point head-on in our sampling section, and include a citation of the very helpful Lloyd-Smith perspective referenced: “Using one sampling rate for all host-virus group pairs is a simplifying assumption, and in reality, there are several interacting and difficult-to-quantify biases likely contained within this host-virus association dataset (Lloyd-Smith 2017)”

Minor points  
  
Line 34: I was initially confused about the lower bound argument. It later became clear that this is contingent on assuming that every host species has at least one host-specific affiliate species, which is not obviously true. So I would argue that this isn’t really a lower bound, since the estimate could easily be lower if this assumption doesn’t hold.

While the Carroll method assumes every host species has one host-specific affiliate, this is not necessarily true of the linear methods; for example, one could assume every host has 0.5 host specific parasites, on average, as a way of dealing with this issue (and some studies have done so).

However, we have clarified this sentence: “However, this approach deliberately excludes generalist parasites, and thus can only be appropriately used to estimate the number of host-specific species (May 1990).”

Lines 63-73 – This passage was a bit hard to understand, before reading the methods to learn about the different fitting approaches.

The methods have been heavily rewritten with most of the results separated out into two tables for clarity, and these sentences are also now preceded by several sentences in the introduction with an overview of the approach.  
  
It is sub-optimal to go through all the main results on viral species without alluding to the results shown in Fig 2A-C.

We now reference this in the first sentence of the results.

Lines 167-179 – Please give the numbers of host and affiiliate species for each example.

We have added this for the seed-dispersal network, which was lacking this information: “For seed dispersal interactions, we used data from a 2007-2008 study of Kenyan rainforest, with 34 plants and 89 dispersers aggregated across all sampling sites…”   
  
Line 249 – Is this statement restricted to non-human hosts? Must be, or else the point about zoonotic viruses makes no sense. Please make this explicit.

This is correct; we have made this point explicit: “There were 296 viruses with more than one non-human host recorded and 149 zoonotic viruses with more than one non-human host recorded.”  
  
Caption to Fig 2 – first sentence of this caption is a bit technical, as it describes quantities that aren’t explained except in the Methods. Also how does node size scale with degree? (by area? Radius?)

In our updated Figure 2, node size is now proportional to the number of viruses sampled in the Olival data for each group; we have added to the methods section to explain Figure 2D clearer.