Global viral biodiversity estimates accounting for host sharing

Colin J. Carlson[1,2,†], Casey M. Zipfel[1], Romain Garnier[1] and Shweta Bansal[1]

[1]*Department of Biology, Georgetown University, Washington, D.C. 20057, USA.*
[2]*National Socio-Environmental Synthesis Center, Annapolis, Maryland 21401, USA.*
[†]*Correspondence should be directed to cjc322@georgetown.edu.*

Submitted to *Nature Ecology and Evolution* on January 28, 2019

Abstract

Present estimates suggest there are over one million virus species found in mammals alone, with roughly half a million posing a possible threat to human health. Although previous estimates assume linear scaling between host and virus diversity, we show that ecological network theory predicts a nonlinear relationship, produced by broad patterns of plasticity in host-virus associations. To account for host sharing, we fit a power law scaling relationship for host-virus species interaction networks. We estimate that there are approximately 40,000 virus species in mammals (including $\sim 10,000$ viruses with zoonotic potential), a reduction of two orders of magnitude from current projections of viral diversity. We expect that the increasing availability of host-virus association data will improve the precision of these estimates, and their utility in the sampling and surveillance of pathogens with pandemic potential. More broadly, we suggest host sharing should be more widely included in macroecological approaches to estimating biodiversity.

Introduction

Measuring global biodiversity is one of the longest-standing problems in ecology. Over several decades, methods have been proposed that range from back-of-the-envelope calculations to sophisticated mechanistic macroecological models. In most cases, these methods estimate diversity from the asymptote of sampling curves over time, effort, or space [1]. Although new species are described every year, the diversity of several major groups of life, like vertebrates and vascular plants, is now fairly well-resolved. On the other hand, invertebrates and microbes, which harbor most of the global diversity of life, continue to pose a challenge. By some calculations, the majority of life on the planet is made up by groups like viruses, helminths, and parasitoid wasps that have become hyperdiverse through coevolution [2, 3]. However, chronic data deficiencies prevent the use of most macroecological methods to estimate the diversity of microbes and parasites, and the number of many taxa is still growing exponentially [4, 5, 6, 7].

To work around these challenges, several methods have been developed that estimate the richness of "affiliate" groups like parasites or mutualists based on the richness of their hosts. The simplest way to estimate the diversity of affiliate species is to multiply host richness by an independent estimate of the mean per-host affiliate richness for a particular pair of host and affiliate groups. For example, a recent study suggested that if every arthropod species has at least one host-specific parasite, there could be at least 81.6 million species of nematode parasites of arthropods [2]. However, this approach deliberately excludes generalist parasites, and thus can only be appropriately used to estimate the number of host-specific species [8]. A handful of other studies focused on parasite diversity have acknowledged the existence of host sharing, adapting the "linear estimation" method by dividing estimates of macroparasite diversity by the average degree of host specificity [6, 9]:

$$\texttt{Total affiliates } A = \frac{\texttt{Mean affiliates per host}}{\texttt{Mean hosts per affiliate}} \texttt{ * Total hosts } H \quad (1)$$

Though this method seems intuitive, one recent study resampled host-helminth association networks to show that diversity actually scales non-linearly. The authors describe this pattern as a power law scaling of host and parasite richness ($A \propto H^{\sim 0.3-0.7}$). In most cases, using this method led to substantially reduced estimates of helminth diversity in vertebrates [10]. This non-linear scaling between host and affiliate richness can be reproduced for several types of species interactions (Figure 1). The reason for this nonlinearity can be described intuitively: the $1{,}000^{th}$ host sampled will on average share more parasites with the first 999 hosts than the 10^{th} host will share with the first nine. This has only recently been proposed as a power law [10]; 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 [11, 12].

In this study, we build on recent advances to estimate the global diversity of mammalian viruses, a problem with applied significance far beyond macroecology. The emergence of viral threats such as Ebola and Zika demands an improved understanding of the landscape of viral emergence, and several ambitious projects are currently working to catalog global viral diversity, with the ultimate aim of predicting outbreaks and preventing pandemics. Recent such work has estimated a global richness of over 1.6 million virus species in mammals and birds, extrapolating total diversity from viral sampling data from the Indian flying fox (*Pteropus giganteus*) and the rhesus macaque (*Macaca mulatta*) [13]. A fundamental gap in such projections is the exclusion of host sharing patterns from these methods (see Supporting Information).

Here, we show that non-linear scaling in bipartite networks implies viral diversity has been overestimated by approximately two orders of magnitude, and discuss how a network perspective can help optimize viral sampling. In the absence of theoretical expectations, we introduce a new simulation method that performs iterative resampling, curve-fitting, and extrapolation on bipartite networks, and we apply it to the most detailed list of mammal-virus associations currently published [14]. With 511 viruses catalogued from 753 mammals (excluding humans), the network covers roughly ten percent of mammal diversity. Our approach estimates viral diversity in two steps: first, we use a power law to extrapolate from sampled hosts to all hosts (100% host sampling but incomplete viral sampling). Second, we extrapolate to the unsampled portion of viral diversity, by using an estimate of sampling completeness derived from the bat and macaque datasets used in previous studies. We repeat this analysis separately for DNA and RNA virus networks based on approximate zoonotic rates in both groups, and ultimately estimate the number of zoonotic viruses in all mammals.

Results

To understand the general relationship between affiliate and host diversity, we resampled the association networks of plants and seed dispersers, plants and mycorrhizal fungi, plants and pollinators and mammals and nematodes. We show in Figures 1 and 2 that each network produced a nonlinear sampling curve. In every case, the power law fitted to these curves overpredicted at higher values. In the Supplementary Methods, we show that more complex functional forms fit to these curves all produced substantially lower estimates of viral diversity, and in the absence of analytic expectations, we err towards parsimony and limit our main results to classical power laws.

To estimate mammalian viral diversity, we iteratively resample mammal-virus

associations and use the same power law approach. We estimate 1,434 viral species would be affiliated to 5,291 mammals (100% host sampling but incomplete viral sampling per host). Using the viral profiles of the bat and the macaque, we then estimated that our host-virus association dataset covered roughly 6% of viral diversity for sampled hosts. This coverage allows us to extrapolate our overall estimate of viral diversity to 23,419 virus species (Table 1). Iteratively fitting models to 50% of the network and projecting out for an upper confidence bound gives an estimate for all mammals of 1,860 virus species, or an extrapolated total of 30,368 species (Table 2). The same method estimates a total of 603 viruses (95% CI: 590—617) for the total network compared to a true value of 511 species, highlighting the small overestimation.

Next, we address the issue of heterogeneity and structure within the association network. At the broadest level, RNA viruses generally have a higher host plasticity than DNA viruses, which should reduce their scaling exponent [14]. We thus evaluate the richness for RNA and DNA viruses separately before combining them. We estimated a total of 26,315 DNA viruses and 14,573 RNA viruses, or a total of 40,878 viruses together (Table 1); even using the 50% upper bound method, we only estimate 55,784 possible total viruses (Table 2). Furthermore, diversity is distributed inhomogeneously among different pairs of host and viral families, with some host groups forming disproportionate reservoirs of viruses with high host-sharing [15]. Curves could be constructed for each of these sub-groups, but would be sensitive to existing taxonomic biases, and simply adding these estimates together would ignore the high degree of sharing among host groups (Figure 2).

Finally, we estimate the total number of potentially zoonotic mammal viruses. Given that RNA viruses have an apparently higher rate of zoonotic infection, we use the separate RNA and DNA virus richness estimates to estimate the number of total potential zoonoses. Using the zoonotic rate in DNA and RNA viruses in the dataset (14.1% and 41.7% respectively), this would suggest a total of 3,710 zoonotic DNA viruses and 6,077 zoonotic RNA viruses—a total of 9,787 compared to the previous estimate of 493,856 to 689,285 [13]. (Table 1) Even using the 50% estimation method as an upper bound, we only estimate a total of 12,941 zoonoses; though higher, this is still only approximately 2-3% of previous estimates. (Table 2)

Discussion

Network science is a useful tool for studying biotic interactions in modern ecology, and offers powerful new ways to understand data such as host-virus associations [16]. Here, we highlight how a simple scaling property of bipartite networks enables a new method of estimating diversity for affiliate groups like parasites and pathogens. Using our computational approach, we found that global viral diversity

in mammals has likely been overestimated by roughly two orders of magnitude, due to the omission of host-sharing patterns. The possible power law we identified here has fundamental implications for biodiversity research, but we have found no analytical solution to the underlying mathematical problem: under a certain set of expectations (e.g., a power law or exponential degree distribution), what is the expected scaling of edges in subsamples drawn randomly from bipartite network? Fitting a power law to these data appears to be adequate for our purposes, but the possibility remains that like the species-area relationship, the scaling of affiliate richness is scale-dependent and described by a more complex pattern [17, 18]; our evidence suggests this scale-dependence may exist and lead to overestimation, as the slope may collapse at broader scales (Figure 1E, Supporting Information). We believe that this is a promising topic for future research in mathematics and network science, and expect that a solution might have broader implications beyond the biological sciences. An analytical expectation for this scaling pattern will also improve the precision—and confidence—of future species richness estimates.

Our study suggests that there are roughly 40,000 viruses in mammals, of which roughly 10,000 have zoonotic potential. Whereas previous estimates assumed 289.5 unique virus species per host, our study suggests there are roughly five to ten times as many virus species as mammal species, with most viruses shared by a few hosts (mean = 4.79, median = 2). While our estimate roughly corrects for undersampling of viruses per host, it does not account for the fact that host plasticity is also likely being significantly underestimated, which might further reduce richness estimates. 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 similarly overestimated in the early 2000s. [19]

Our results highlight the need for completeness not just in viral inventories but in host-virus association data. Even with the development of viral sequencing techniques allowing easier access to diversity estimates within – and potentially between – hosts [20], the need for completeness makes the problem of cataloging viral diversity exponentially more intensive. Targeting specific groups may help make this problem more manageable: groups like bats, rodents, and primates harbor disproportionate viral richness, even accounting for sampling bias due to their high zoonotic rate [21, 22, 16, 23]. Moreover, zoonotic viruses in these groups may ultimately account for the majority of viral sharing over broad phylogenetic distances [15, 24]. Focusing on describing viral plasticity within and among these groups might reduce the effort needed to approximate the overall level of host-sharing in the network, and therefore, the effort needed to update viral richness

estimates.

On the other hand, it is difficult to assess how much the dominance of zoonoses in sharing networks is a feature, not an artefact, of current sampling schemes; separating the zoonotic and non-zoonotic viruses in our association data shows a tight coupling between sampling, sharing, and existing priorities for zoonotic virus description (Figure 2D). Even within well-sampled groups like bats, sampling priorities may poorly reflect underlying patterns of viral richness [25, 26]; for groups that are less common reservoirs of zoonoses, there is almost certainly a disproportionate level of undersampling in host-virus associations, and a disproportionately high observed zoonotic rate. The methods we use here can help standardize estimates of viral richness for sampling effort and, in conjunction with real-time data collection, dynamically target hotspots of undiscovered viral richness for sampling [27]. Advances in machine learning that predict possible host-virus links [28, 29] may help further target sampling in this regard. In coming years, new evidence may change conventional knowledge about the structure of the mammalian viral sharing network, and decouple the tight correlation between zoonotic sampling and the centrality of groups like bats and carnivores within it.

Finally, we note that mammal viruses are only a subset of the hyperdiverse affiliate taxa on earth, and many groups remain unassessed using methods that account for host sharing. Bird viral diversity is a logical next target, as the existing estimate was calculated using the same estimates derived from one monkey and one bat species [13]. But the viral diversity of all vertebrates is an important end goal, given recent work showing that RNA viruses are widely distributed across all five classes of vertebrates—even viral families, including the Filoviridae or Flaviviridae, that pose some of the greatest emerging threats to human health [30]. Though viruses like Wenzhou shark flavivirus or Wenling triplecross lizardfish picornavirus may never pose a threat to human health, they remain an important part of understanding, defining, and measuring the global virome [30, 31].

Methods

In this study we estimate the global diversity of viruses in mammal hosts, by re-analyzing data that has been previously used to provide a linear estimate by Carroll *et al.* [13]. (Previous estimates are described in the Supplementary Methods.)

Biotic interaction data

To illustrate the scaling properties of bipartite species association networks, we provide four examples, using published association datasets. For plant-pollinator interactions, we used Robertson's classic 1929 study in southwest Illinois, with 456 plant and 1429 pollinator species [32, 33]. 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 [34]. Both of these datasets were obtained from NCEAS's Interaction Web Database [35]. For mycorrhizal interaction networks, we used a dataset on fungal associations in 150 Japanese plant species/taxa (not all resolved to species level), including 8,080 total operational taxonomic units (OTUs); we only used data on arbuscular mycorrhizae, for convenience [36]. Finally, for helminth-vertebrate interactions, we used the `helminthR` package to compile a global interaction web of nematode-mammal interactions, with 849 mammal species and 2,248 nematode species [37, 38].

To develop our estimates of mammalian viral diversity, we constructed a viral interaction network using the raw data made available by Olival *et al.* [14]. Humans are disproportionately represented in this dataset, so much so that constructing resampled curves produces two distinct curves depending on whether *Homo sapiens* are included or not in a given subsample (Figure S1). Consequently, we removed humans from all of our network analyses. The remaining network includes 511 viruses hosted by 753 mammal species. Several features in the database, such as host classification and virus classification, were used in subsequent analyses; for analyses involving zoonotic proportions, the non-stringent classifications of zoonotic risk were used. The proportion of viruses described was derived using the proportion of estimated viral diversity known from *Pteropus giganteus* and *Macaca mulatta* viral metagenomics and by constructing a rarefaction curve over the number of individual animals sampled (as in [13]).

Bipartite richness estimators

We developed a new *R* package, `codependent`, to streamline bipartite richness estimation. The method subsamples a network with **H** host species and **A** affiliate species, and $\forall i \in [1, H]$, subsamples i host species n times, and counts the number of affiliate species $\hat{a}$. (This assumes every host has at least one affiliate species, and in some cases overestimates affiliate richness for this reason.) A power law function

is then fit of the form $a \propto bi^z$ using nonlinear least squares regression (`nls`), with initial parameters $\hat{b} = 1, \hat{z} = 0.5$. The `copredict` function in `codependent` returns the point estimates for curve parameters with a 95% confidence interval using the `confint2` function in the `nlstools` package, and then extrapolates the curve to the total number of host species (in this case, an estimate of 5,291 mammal species), including a 95% CI.

For our viral richness estimates for mammals, we resampled a curve with every number of hosts (i) between 1 and $H = 753$, each $n = 1000$ times, and used the `copredict` function to project to 5,291 total mammal species. We repeated this process separately for DNA and RNA viruses, which have different overall patterns of diversity and host specificity. We multiply these by the proportion reported as zoonotic in the Olival *et al.* dataset to obtain total estimates of zoonotic viral richness. The true proportion of viruses with zoonotic potential may be higher, as many viruses simply have yet to emerge in human populations, or it may be lower, as zoonotic viruses sampled from hyperreservoirs make up a disproportionate share of known viral diversity. But the total number of zoonotic viruses is still bounded within the 0% and 100% of total viral richness estimates, which are ultimately still much smaller than previous estimates of zoonoses alone.

As a final method for bounding uncertainty, we use the `codependent.ci` function, which iterates the same rarefaction method on 50% of the network (half the total number of hosts), and projects it out to a given proportion of hosts. Fitting the curve on smaller portions of the network leads to z values closer to 1, and therefore the method overestimates (see Figure 1); this makes this confidence bound method an absolute outer bound on plausible richness. For example, using the helminth network in Figure 1, fitting a curve with $n = 100$ iterations each gives an estimate of 2,291 nematode species (95% CI: 2,271 to 2,311) compared to a true richness of 2,248 species. We apply this methodology to the virus network with 200 iterations again, and project over the total network (753 mammal species) and out to total mammal richness (5,291 species).

Correcting for sampling

To estimate how comprehensive the Olival *et al.* dataset is, we compare the number of recorded viruses in those data versus the viral metagenomics dataset. For both the bat and macaque, we first count the number of virus species recorded in the Olival dataset (host-virus associations). Next, we estimate the "true richness" by adding the number of known virus species (from the metagenomic data) to the number of undescribed species estimated using the Chao-1 method (also included in the previous metagenomic estimates). The estimates of undescribed diversity come with their own lower and upper 95% confidence bounds, which we used to create upper and lower 95% bounds respectively on the proposed sampling rate.

We average these two rates between the bats and macaques to estimate a sampling rate of 6.1%, with a 95% CI of 3.4% to 7.4%. While these rates should ideally be derived from a larger and representative sample of species, we note that the bat and macaque datasets are fairly unique in their completeness.

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 [39]) and non-random coevolutionary diversification. It is likely also a liberal estimate of undersampling, given that bats (especially *Pteropus*, a major zoonotic reservoir) have a disproportionately high underlying viral richness [21].

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 [40]. Ideally, at a minimum, we would ideally be able to derive separate sampling rate estimates for DNA and RNA viruses. However, our ability to do so is increasingly limited by sample size: for example, no DNA viruses are recorded for *Pteropus* in the main dataset. If we used these methods, we could derive a DNA virus sampling rate of 25.5% (95% CI: 18.1%—26.0%) and an RNA virus sampling rate of 7.2% (95% CI: 3.3%—8.8%); both independent estimates reduce the total unsampled viral diversity. In Table S7 we show how using these numbers would reduce overall estimates.

Network analyses

To generate a unipartite network of host sharing by viruses, we analyzed associations between viruses and their hosts [8]. We classified hosts by their orders (separating out *Homo sapiens* from primates) and represented these orders as the nodes in the network. Links between these nodes represent instances of shared viruses between host species belonging to different orders. We ignored viral sharing between host species within the same order (i.e., self links were removed). Edges were weighted proportional to the Jaccard index [41], which is defined by

$$J = \frac{C}{A + B - C} \tag{2}$$

where A and B are the number of viruses in two orders, respectively, and C are the number shared between orders.

This network was created separately for zoonotic and non-zoonotic viruses. There were 296 viruses with more than one non-human host recorded and 149 zoonotic viruses with more than one non-human host recorded. Additionally, there were 116 viruses with more than one order recorded, and 86 zoonotic viruses with

more than one order recorded. Networks were constructed and analyzed using the `networkx` package in *Python* [42].

Data and code availability

All data in this study is from previous studies and is available online for researchers to reproduce our results. All code from this study is also available. All code and data can be found at github.com/cjcarlson/brevity. The `codependent` *R* package is available at github.com/cjcarlson/codependent

Author Contributions

CJC, CMZ, RG, and SB conceived of the study. CJC and CMZ performed all analyses. All authors contributed to the writing and approved the final draft.

Acknowledgements

We thank Tad A. Dallas, Phillip P.A. Staniczenko, and three anonymous reviewers for thoughtful comments on the manuscript and the methodology. We also acknowledge Dr. Dallas for assistance with the `codependent` package. This work was supported by the National Socio-Environmental Synthesis Center (SESYNC) under funding received from the National Science Foundation DBI-1639145.

References

[1] Colwell, R. K., Chao, A., Gotelli, N. J., Lin, S.-Y., Mao, C. X., Chazdon, R. L., and Longino, J. T. *Journal of Plant Ecology* **5**(1), 3–21 (2012).

[2] Larsen, B. B., Miller, E. C., Rhodes, M. K., and Wiens, J. J. *The Quarterly Review of Biology* **92**(3), 229–265 (2017).

[3] Windsor, D. A. *International Journal for Parasitology* **28**(12), 1939–1941 (1998).

[4] Bacher, S. *Trends in Ecology & Evolution* **27**(2), 65–66 (2012).

[5] Colwell, R. K. and Coddington, J. A. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **345**(1311), 101–118 (1994).

[6] Poulin, R. and Morand, S. *Parasite biodiversity.* Smithsonian Books, (2004).

[7] Quicke, D. L. *PLoS One* **7**(2), e32101 (2012).

[8] May, R. M. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **330**(1257), 293–304 (1990).

[9] Dobson, A., Lafferty, K. D., Kuris, A. M., Hechinger, R. F., and Jetz, W. *Proceedings of the National Academy of Sciences* **105**, 11482–11489 (2008).

[10] Strona, G. and Fattorini, S. *International Journal for Parasitology* **44**(5), 269–272 (2014).

[11] Delmas, E., Besson, M., Brice, M.-H., Burkle, L. A., Dalla Riva, G. V., Fortin, M.-J., Gravel, D., Guimarães Jr, P. R., Hembry, D. H., Newman, E. A., et al. *Biological Reviews* (2017).

[12] Pellissier, L., Albouy, C., Bascompte, J., Farwig, N., Graham, C., Loreau, M., Maglianesi, M. A., Melián, C. J., Pitteloud, C., Roslin, T., et al. *Biological Reviews* **93**(2), 785–800 (2018).

[13] Carroll, D., Daszak, P., Wolfe, N. D., Gao, G. F., Morel, C. M., Morzaria, S., Pablos-Méndez, A., Tomori, O., and Mazet, J. A. *Science* **359**(6378), 872–874 (2018).

[14] Olival, K. J., Hosseini, P. R., Zambrana-Torrelio, C., Ross, N., Bogich, T. L., and Daszak, P. *Nature* **546**(7660), 646–650 (2017).

[15] Johnson, C. K., Hitchens, P. L., Evans, T. S., Goldstein, T., Thomas, K., Clements, A., Joly, D. O., Wolfe, N. D., Daszak, P., Karesh, W. B., et al. *Scientific Reports* **5**, 14830 (2015).

[16] Gómez, J. M., Nunn, C. L., and Verdú, M. *Proceedings of the National Academy of Sciences* , 201220716 (2013).

[17] Harte, J., Smith, A. B., and Storch, D. *Ecology Letters* **12**(8), 789–797 (2009).

[18] Wilber, M. Q., Kitzes, J., and Harte, J. *Global Ecology and Biogeography* **24**(8), 883–895 (2015).

[19] Ignacio-Espinoza, J. C., Solonenko, S. A., and Sullivan, M. B. *Current Opinion in Virology* **3**(5), 566–571 (2013).

[20] Grubaugh, N. D., Gangavarapu, K., Quick, J., Matteson, N. L., De Jesus, J. G., Main, B. J., Tan, A. L., Paul, L. M., Brackney, D. E., Grewal, S., et al. *bioRxiv* , 383513 (2018).

[21] Luis, A. D., Hayman, D. T., O'Shea, T. J., Cryan, P. M., Gilbert, A. T., Pulliam, J. R., Mills, J. N., Timonin, M. E., Willis, C. K., Cunningham, A. A., et al. *Proceedings of the Royal Society of London B: Biological Sciences* **280**(1756), 20122753 (2013).

[22] Brook, C. E. and Dobson, A. P. *Trends in Microbiology* **23**(3), 172–180 (2015).

[23] Han, B. A., Kramer, A. M., and Drake, J. M. *Trends in Parasitology* **32**(7), 565–577 (2016).

[24] Woolhouse, M. E. and Gowtage-Sequeria, S. *Emerging Infectious Diseases* **11**(12), 1842 (2005).

[25] Levinson, J., Bogich, T. L., Olival, K. J., Epstein, J. H., Johnson, C. K., Karesh, W., and Daszak, P. *Emerging Infectious Diseases* **19**(5), 743 (2013).

[26] Young, C. C. and Olival, K. J. *PLoS One* **11**(2), e0149237 (2016).

[27] Restif, O., Hayman, D. T., Pulliam, J. R., Plowright, R. K., George, D. B., Luis, A. D., Cunningham, A. A., Bowen, R. A., Fooks, A. R., O'Shea, T. J., et al. *Ecology Letters* **15**(10), 1083–1094 (2012).

[28] Dallas, T., Park, A. W., and Drake, J. M. *PLoS Computational Biology* **13**(5), e1005557 (2017).

[29] Elmasri, M., Farrell, M., and Stephens, D. A. *arXiv preprint arXiv:1707.08354* (2017).

[30] Shi, M., Lin, X.-D., Chen, X., Tian, J.-H., Chen, L.-J., Li, K., Wang, W., Eden, J.-S., Shen, J.-J., Liu, L., et al. *Nature* **556**(7700), 197 (2018).

[31] Geoghegan, J. L., Di Giallonardo, F., Cousins, K., Shi, M., Williamson, J. E., and Holmes, E. C. *Virus Evolution* **4**(2), vey031 (2018).

[32] Robertson, C. *Flowers and insects lists of visitors of four hundred and fifty three flowers*. The Science Press Printing Company, (1929).

[33] Marlin, J. C. and LaBerge, W. E. *Conservation Ecology* **5**(1), 9 (2001).

[34] Schleuning, M., Blüthgen, N., Flörchinger, M., Braun, J., Schaefer, H. M., and Böhning-Gaese, K. *Ecology* **92**(1), 26–36 (2011).

[35] NCEAS Interaction Web Database. www.nceas.ucsb.edu/interactionweb. Accessed: 2018-09-01.

[36] Toju, H., Tanabe, A. S., and Sato, H. *Microbiome* **6**(1), 116 (2018).

[37] Dallas, T. *Ecography* **39**(4), 391–393 (2016).

[38] Dallas, T. A., Aguirre, A. A., Budischak, S., Carlson, C., Ezenwa, V., Han, B., Huang, S., and Stephens, P. R. *Global Ecology and Biogeography* **27**(12), 1437–1447.

[39] Jordano, P. *Functional Ecology* **30**(12), 1883–1893 (2016).

[40] Lloyd-Smith, J. O. *Nature* **546**(7660), 603 (2017).

[41] Pilosof, S., Morand, S., Krasnov, B. R., and Nunn, C. L. *PloS One* **10**(3), e0117909 (2015).

[42] Schult, D. A. In *In Proceedings of the 7th Python in Science Conference (SciPy)*, 11–15, (2008).

Figures

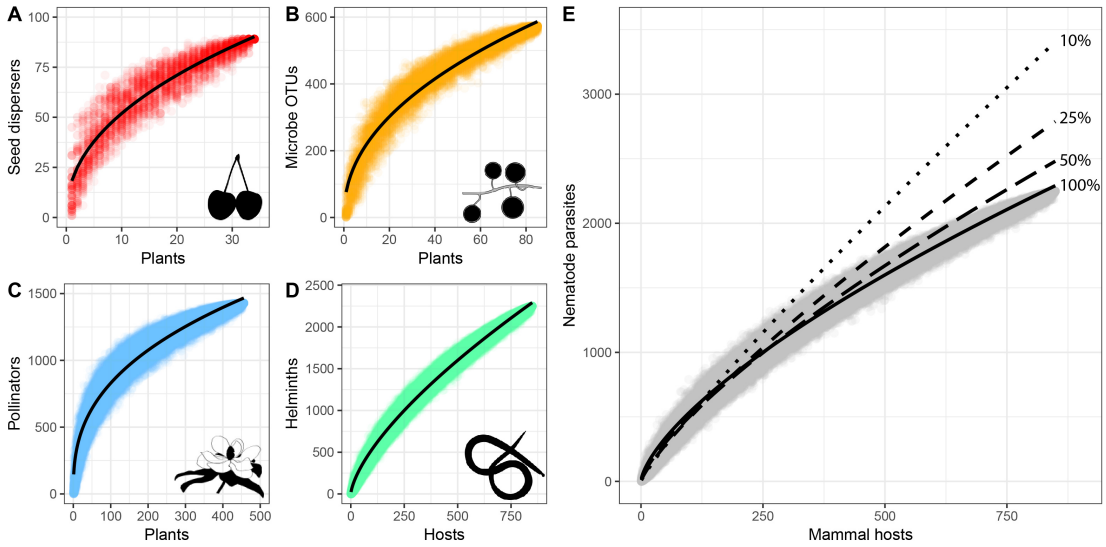

Figure 1: Fitting power law relationships between affiliates and host diversity, with shape $A \propto bH^z$ (where $z = 1$ is linear scaling). The power law scaling of affiliate and host richness for four networks of species interactions: plant-seed disperser (A; $z = 0.45$), plant-arbuscular mycorrhizae (B; $z = 0.46$), plant-pollinator (C; $z = 0.38$), and mammal-nematode (D; $z = 0.67$). Each point shows a network subsample used to fit the total model. At lower sampling levels, the same curves approach linearity, which we show in (E) by resampling the mammal-nematode network for only 10% of hosts ($z = 0.89$), 25% ($z = 0.81$), 50% ($z = 0.75$), and 100% ($z = 0.68$), and refitting curves. Linear approximations may seem appropriate at low sampling levels, but significantly overestimate the size of the entire network. Curves are each built with 100 iterations.

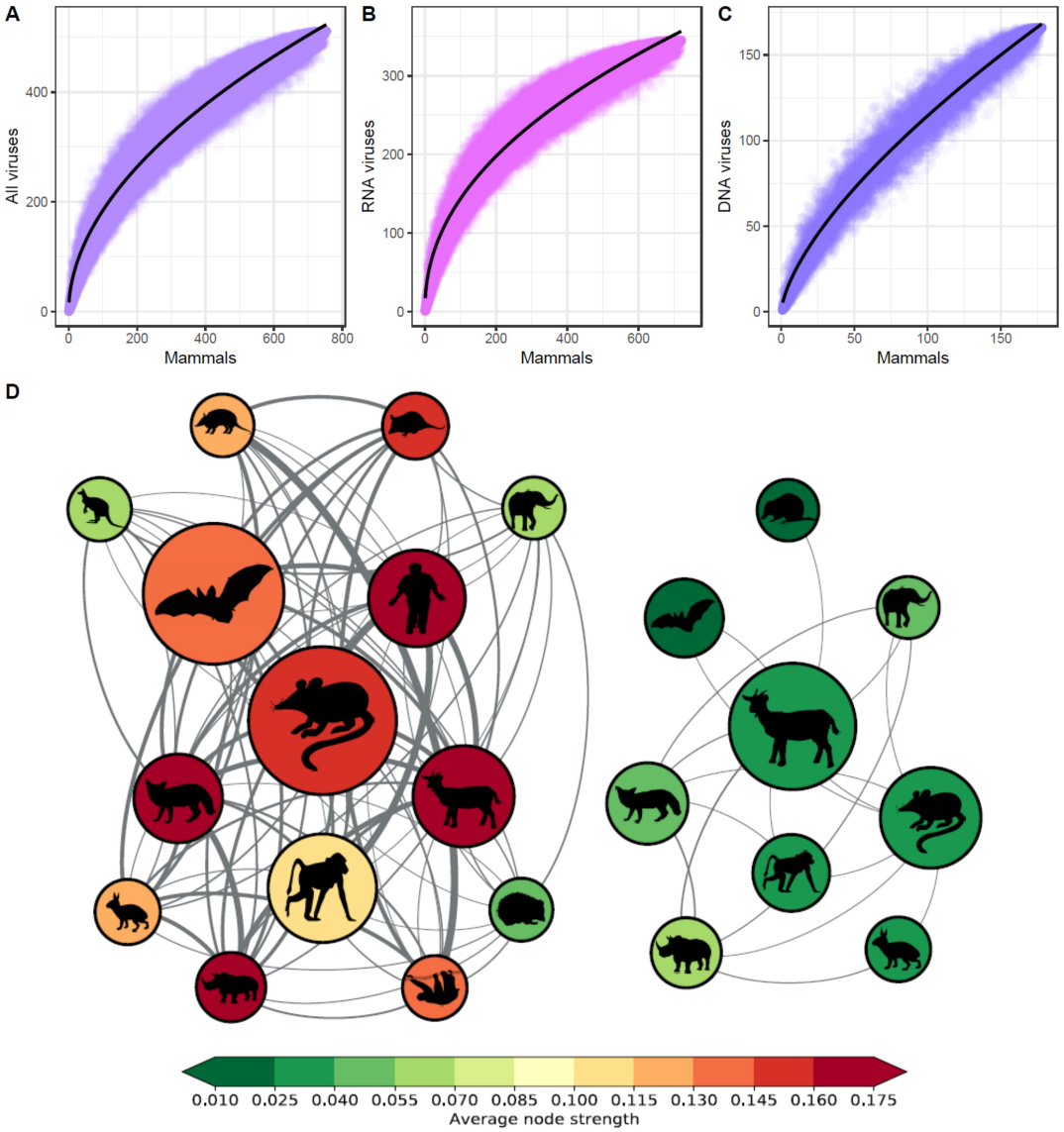

Figure 2: Bipartite rarefaction curves on the known viral network constructed from 100 samples (in-text estimates use 1000). Points show each sub-sample of the total network, and curves were fitted for all viruses (A; $z = 0.517$), RNA viruses (B; $z = 0.460$), and DNA viruses (C; $z = 0.667$). DNA viruses are more host specific, and thus the rarefaction curve is closer to linear. Viral sharing is unevenly distributed across the network, with a handful of groups—bats, primates, ungulates, rodents, and carnivores—accounting for the majority of viral sharing (D); zoonotic viruses are shown on the left and non-zoonotic are shown on the right. Node size is proportional to number of viruses sampled in the Olival dataset. Edge weight is proportional to the Jaccard index for viruses shared between host groups. Node color relates to average node strength (calculated for each node as the sum of the edge weights divided by the number of edges), where red is high average strength and green is low average strength.

Entire network		**Estimate**	**95% CI**
100% method	Raw estimate	1,434	(1,428—1,441)
	Sampling correction	23,419	(19,191—42,397)

DNA and RNA separate		**Estimate**	**95% CI**
DNA viruses	Raw estimate	1,612	(1,593—1,631)
	Sampling correction	26,315	(21,413—47,977)
	Zoonoses	3,710	(3,109—6,765)
RNA viruses	Raw estimate	893	(889—897)
	Sampling correction	14,573	(11,944—26,377)
	Zoonoses	6,077	(4,981—10,999)
DNA + RNA	Sampling corrected	40,878	(33,357—74,354)
	Zoonoses	9,787	(8,000—17,764)

Table 1: Estimation of viral diversity using 100% of the viral network, with the entire network, and then separation of DNA and RNA viruses.

All viruses		Estimate	95% CI
50.0% method	Raw estimate	1,860	(1,811—1,910)
	Sampling correction	30,368	(24,350—56,183)

DNA and RNA separate		Estimate	95% CI
DNA viruses	Raw estimate	2,290	(2,243—2,339)
	Sampling correction	37,394	(30,147—68,807)
	Zoonoses	5,273	(4,251—9,702)
RNA viruses	Raw estimate	1,126	(1,118—1,135)
	Sampling correction	18,390	(15,025—33,390)
	Zoonoses	7,668	(6,265—13,924)
DNA + RNA	Sampling corrected	55,784	(45,173—102,196)
	Zoonoses	12,941	(10,516—23,626)

Table 2: Estimation of viral diversity using the lower 50% of the subsampled curve.