

1 Applying ecological network theory to re-estimate  
2 global viral diversity: host sharing matters

3 Colin J. Carlson<sup>1,2,†</sup>, Casey M. Zipfel<sup>1</sup>, Romain Garnier<sup>1</sup> and Shweta Bansal<sup>1</sup>

4 <sup>1</sup>*Department of Biology, Georgetown University, Washington, D.C. 20057, USA.*

5 <sup>2</sup>*National Socio-Environmental Synthesis Center, Annapolis, Maryland 21401, USA.*

6 <sup>†</sup>*Correspondence should be directed to ccarlson@sesync.org.*

7 Submitted to *Nature Ecology and Evolution* on October 15, 2018

8 **Abstract**

9 Current estimates of viral richness in vertebrates assume linear scaling be-  
10 tween host and virus diversity, but ecological network theory predicts a non-  
11 linear relationship. Taking host sharing into account, we fit a power law  
12 scaling relationship for host-virus species interaction networks, and estimate  
13 that there are approximately 37,000 virus species in mammals (including  
14  $\sim 8,980$  viruses with zoonotic potential), a reduction of two orders of mag-  
15 nitude from current projections of viral diversity. The structure of host-virus  
16 networks has fundamental implications for optimal sampling of global viral  
17 diversity and surveillance of pathogens with pandemic potential.

18 The emergence of global viral threats such as Ebola and Zika demands an  
19 improved understanding of the landscape of viral emergence. Several ambitious  
20 projects are currently working to catalog global viral diversity, with the ultimate  
21 aim of predicting outbreaks and preventing pandemics. Recent such work has es-  
22 timated a global richness of over 1.6 million virus species in mammals and birds,  
23 extrapolating total diversity from viral sampling data from the flying fox (*Ptero-*  
24 *pus giganteus*) and a macaque (*Macaca mulatta*) [1]. A fundamental gap in such  
25 projections is the lack of attention to host sharing patterns. In this study, we  
26 examine how species richness scales in bipartite ecological networks, and find that  
27 host-sharing produces a nonlinear relationship between host and affiliate diversity.  
28 We show that this scaling pattern implies viral diversity has been severely overes-  
29 timated, and discuss how a network perspective can help optimize viral sampling.

30 Estimating the richness of hyperdiverse taxa is a long-standing problem in  
31 ecology [2, 3]. The simplest way to estimate the diversity of affiliate species (like  
32 parasites or mutualists) is to multiply host richness by an independent estimate of  
33 per-host affiliate richness. Because this approach ignores host sharing, it overesti-  
34 mates total diversity and can only provide a lower bound on the number of host-  
35 specific species [4]. For example, a recent study suggested that if every arthropod  
36 species has at least one host-specific parasite, there should be at least 81.6 million  
37 species of nematode parasites of arthropods [5]. Other studies acknowledge the  
38 existence of host sharing, but correct by dividing their estimates by the average  
39 degree of host specificity [6, 4]. However, one recent study showed that resampling  
40 host-helminth association networks actually produces a power law scaling between  
41 host and parasite richness ( $A \propto H^{\sim 0.4-0.7}$ ), and in most cases, this substantially  
42 reduced helminth richness estimates [7].

43 This scaling between host and affiliate richness seems to be a fairly consistent  
44 property of ecological association networks, and as we show here, can be repro-  
45 duced for several types of species interactions (**Figure 1A-D**). The reason for this  
46 nonlinearity can be described intuitively: the 1,000<sup>th</sup> host sampled will on average  
47 share more parasites with the first 999 hosts than the 10<sup>th</sup> host will share with the  
48 first nine. This has only recently been previously described as a power law [7]; the  
49 pattern may be subtle enough at smaller scales that the pattern would not have  
50 been evident without the kind of large network data, and ease of computational  
51 resampling, that has only become available recently. Fitting a power law to these  
52 data appears to be adequate for describing the shape of these sampling curves.  
53 But the possibility remains that like the species-area relationship, the slope of this  
54 scaling law is scale-dependent and described by a more complex pattern [8, 9]; our  
55 evidence suggests this scale-dependence may exist and lead to overestimation, as  
56 the slope may collapse at broader scales (**Figure 1E**).

57 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 [10]. With 511 viruses catalogued from 753 mammals (excluding humans), the network covers roughly ten percent of mammal diversity. (Based on the results shown in **Figure 1E**, this seems to suggest our analysis will probably be predisposed to overestimation.) Using the power law method, we estimate 1,435 (95% CI: 1,431-1,491) viruses would be described from 5,291 mammals (100% host sampling but incomplete viral sampling per host). Using the viral profiles of the bat and the macaque, we can estimate that roughly 6.7% of viral diversity is catalogued in that association database, and correct our overall estimate to 21,433 virus species (95% CI: 21,374-21,493). Iteratively fitting models to 50% of the network and projecting out for an upper confidence bound (see **Methods**) gives an estimate for all mammals of 1,768 (95% CI: 1,743-1,794) virus species, or an extrapolated total of 26,388 species (95% CI: 26,014-26,776). The same method estimates a total of 568 viruses (95% CI: 562-575) for the total network compared to a true value of 511 species, highlighting the small overestimation.

A further problem is heterogeneity and structure within the association network, both in terms of host and virus taxonomy. Some hosts are also more connected than others, meaning they should accumulate richness faster [11]. Curves could be constructed for each of these sub-groups, but simply adding these estimates together would ignore the high degree of sharing among host groups. On the other hand, RNA viruses generally have a higher host plasticity than DNA viruses, which should reduce their scaling exponent [10], and these curves can be separately constructed and summed. Using the same methods, we estimate a total of 894 RNA viruses (95% CI: 892-897) and 1,613 DNA viruses (1,592-1,634) before correction for undersampling, or an extrapolated total of 13,353 RNA viruses (13,323-13,398) and 24,087 DNA viruses (23,772-24,402).

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  $\sim 3,407$  zoonotic DNA viruses and  $\sim 5,573$  zoonotic RNA viruses—a total of 8,980 compared to the previous estimate of 493,856 to 689,285 [1]. Even using the 50% estimation method as an upper bound, we estimate a total of 31,239 (30,373-32,119) DNA viruses, 13,037 (12,675-13,404) zoonotic DNA viruses, 17,269 (17,045-17,478) total RNA viruses, and 7,207 (7,113-7,294) zoonotic RNA viruses; this sets an upper bound of  $\sim 20,244$  (19,788-20,698) zoonotic viruses. Though higher, this is still only approximately 2-3% of previous estimates.

Our study indicates that global viral diversity in mammals has likely been overestimated by roughly two orders of magnitude, due to the omission of host-sharing

98 patterns. Whereas the previous estimate assumes 289.5 unique virus species per  
 99 host, our study suggests there are roughly five to ten times as many virus species  
 100 as mammal species, with most viruses shared by a few hosts (mean = 4.79, median  
 101 = 2). While our estimate roughly corrects for undersampling of viruses per host, it  
 102 does not account for viral connectedness (host plasticity), which is also likely being  
 103 significantly underestimated, and might further reduce estimates. These results  
 104 have basic implications for how we target viral sampling strategies; while many  
 105 recent emerging diseases have been single-stranded RNA viruses [10, 11, 12], our  
 106 results indicate these may be outnumbered by potentially-zoonotic DNA viruses.  
 107 Previous work may suggest that DNA viruses emerge at a slower rate, in no small  
 108 part because of their lower host plasticity [12], and this might account for the  
 109 smaller number of currently-known zoonotic DNA viruses.

110 As we show here, estimating global viral diversity requires attention to hotspots  
 111 of both viral richness and host-sharing. Current sampling priorities reflect conven-  
 112 tional knowledge that groups like bats, rodents, and primates harbor dispropor-  
 113 tionate viral richness, even accounting for different rates of zoonoses [13, 14, 15].  
 114 But even within well-sampled groups like bats, sampling priorities may poorly re-  
 115 flect underlying patterns of viral richness [16, 17]. The methods we use here can  
 116 help standardize estimates of viral richness for sampling effort and, in conjunction  
 117 with real-time data collection, dynamically target hotspots of undiscovered viral  
 118 richness for sampling [18]. But our results also highlight the need for complete-  
 119 ness in host-virus association data, not just within groups but at broader scales.  
 120 Even with the development of viral sequencing techniques allowing easier access  
 121 to diversity estimates within – and potentially between – hosts [19], the need for  
 122 completeness makes the problem of cataloging viral diversity exponentially more  
 123 intensive. Luckily, a handful of mammal groups account for the majority of vi-  
 124 ral sharing across groups (**Figure 2, Table S1**), not just for zoonoses but for all  
 125 viruses. The fact that host plasticity predicts zoonotic risk has been observed be-  
 126 fore [11, 12], and is already used to target many of these high-risk host groups for  
 127 higher-priority viral discovery efforts. Focusing on describing viral plasticity within  
 128 and among these groups reduces the effort needed to approximate the overall level  
 129 of host-sharing in the network, and therefore, the effort needed to update viral  
 130 richness estimates. Advances in machine learning that predict possible host-virus  
 131 links [20, 21] may help further target sampling in this regard.

132 Network theory is a useful tool for studying biotic interactions in modern ecol-  
 133 ogy, and offers powerful new ways to understand data such as host-virus associ-  
 134 ations. In this study, we used network methods to quantify and explore global  
 135 mammal viral diversity, but this framework could readily be extended to the rest  
 136 of the vertebrate tree of life. Bird viral diversity is an important next target, as the  
 137 existing estimate was calculated using the same mammal viral richness estimates

138 derived from one monkey and one bat species [1]. But the viral diversity of all  
 139 vertebrates is an important target for future estimation, given recent work showing  
 140 that RNA viruses are widely distributed across all five classes of vertebrates—even  
 141 viral families, like the Filoviridae or Flaviviridae, that include some of the great-  
 142 est emerging threats to human health [22]. Though viruses like Wenzhou shark  
 143 flavivirus or Wenling triplecross lizardfish picornavirus may never pose a threat  
 144 to human health, they remain an important part of understanding, defining, and  
 145 measuring the global virome.

## 146 Methods

147 In this study we re-estimate the global diversity of viruses in mammal hosts. We  
 148 follow up on previously proposed estimates, which used a methodology that treats  
 149 all viruses as “100% host specific” and all hosts as equivalent in diversity [23, 1, 10].  
 150 In this case, rarefaction curves were constructed for one or two individual species  
 151 as a function of sampling effort, and then per-viral family per-host extrapolation  
 152 was extended over the total diversity of mammals. In the work of Anthony *et al.*,  
 153 for example, *Pteropus giganteus* was sampled [23]; in the work of Carroll *et al.*,  
 154 both *Pteropus giganteus* and *Macaca mulatta* were sampled [1]. Their estimate of  
 155 mammal diversity is broken down as

$$\begin{aligned} 1,531,745 \text{ viruses} &= 11.58 \text{ viruses per family} \\ &\times 25 \text{ viral families} \\ &\times 5,291 \text{ estimated mammal species} \end{aligned}$$

156 Given the importance of influenza viruses from birds in human (and mammal)  
 157 health, the authors similarly extend their estimate for birds using the same 11.58  
 158 viruses per family per host and one viral family (Orthomyxoviridae, the family of  
 159 RNA viruses that include influenza), and add the estimate for birds (137,362.96)  
 160 to their total to estimate there are 1,669,106 viruses total. In their dataset, 32.2%  
 161 of viruses are zoonotic and 45.0% are human viruses, which they use as confidence  
 162 bounds and multiply by 1.6 million to obtain an estimate of 631,218 to 826,647  
 163 zoonotic viruses. Here, we use a combination of new network methods and similar  
 164 information on the proportion of zoonotic viruses and undiscovered viruses, to  
 165 reproject viral diversity among mammals.

## 166 Network data

167 To illustrate the scaling properties of bipartite species association networks, we  
 168 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 [24, 25]. For seed dispersal interactions, we used data from a 2007-2008 study of Kenyan rainforest, aggregated across all sampling sites [26]. Both of these datasets were obtained from NCEAS’s Interaction Web Database [27]. 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 [28]. 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 [29, 30].

A viral interaction network was constructed using the raw data made available by Olival *et al.* [10]. Humans are disproportionately represented in this dataset, so much so that running the rarefaction process with *Homo sapiens* included produces two distinct curves depending on whether they are included or not in a given subsample (**Figure S1**). Consequently, we removed humans from all 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 or undescribed was derived in the same method as the Carroll *et al.* study, using the proportion of estimated viral diversity known from *Pteropus giganteus* and *Macaca mulatta* viral metagenomics and a rarefaction curve over number of individual animals sampled [1]. 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, to arrive at the back-of-the-envelope estimate that  $\sim 6.7\%$  of all viruses have been described for the hosts present in the association dataset. This estimate is the most tenuous in our analysis, but uses much the same logic as the linear extrapolation used by Carroll *et al.* 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 [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 for  $i \in (1, \dots, H)$  subsamples  $i$  host species  $n$  times, and counts the number of affiliate species  $\hat{a}_i$ . (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 regres-

sion (`nls`), with initial parameters  $\hat{b} = 1, \hat{z} = 0.5$ . The `copredict` function in `codependent` runs this process for a set number of iterations, and in each iteration extrapolates the curve to the total number of host species (in this case, an estimate of 5,291 mammal species). The average estimate is returned with a 95% confidence interval based on  $\pm 1.96 * SE$ .

For our viral richness estimates for mammals, we resampled a curve with every number of hosts between 1 and 753 each once ( $n = 1$ ), 200 times, and used the `copredict` function to project out 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 both a set endpoint for extrapolation, and to 100% of the network. Estimates are log-normally distributed, and we use the `elnorm` function in the `EnvStats` package [31] to derive an appropriate minimum variance unbiased 95% confidence interval on the estimates. 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 100 curves with  $n = 1$  iteration each gives an estimate of 2,492 nematode species (95% CI: 2,472 to 2,512) 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).

## 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 (excluding *Homo sapiens*) 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 number of viruses shared between orders. This network was

created for all viruses in the dataset and for just zoonotic viruses in the dataset. There were 296 viruses with more than one host recorded and 149 zoonotic viruses with more than one 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 generated using the NetworkX package in Python [32].

## Data and code availability

All data in this study is taken from previous studies and is available online for researchers to reproduce our results. All code from this study is available on Github at [github.com/cjcarlson/brevity](https://github.com/cjcarlson/brevity), which also includes copies of all raw data. The codependent R package is available at [github.com/cjcarlson/codependent](https://github.com/cjcarlson/codependent)

## Author Contributions

CJC, RG, and CMZ 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 and Phillip P.A. Staniczenko for thoughtful comments on the manuscript and the methodology. This work was supported by the National Socio-Environmental Synthesis Center (SESYNC) under funding received from the National Science Foundation DBI-1639145.

## References

- [1] 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).
- [2] Colwell, R. K. and Coddington, J. A. *Phil. Trans. R. Soc. Lond. B* **345**(1311), 101–118 (1994).
- [3] Quicke, D. L. *PLoS One* **7**(2), e32101 (2012).
- [4] May, R. M. *Phil. Trans. R. Soc. Lond. B* **330**(1257), 293–304 (1990).
- [5] Larsen, B. B., Miller, E. C., Rhodes, M. K., and Wiens, J. J. *The Quarterly Review of Biology* **92**(3), 229–265 (2017).
- [6] 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).



- 277 [7] Strona, G. and Fattorini, S. *International Journal for Parasitology* **44**(5),  
278 269–272 (2014).
- 279 [8] Harte, J., Smith, A. B., and Storch, D. *Ecology Letters* **12**(8), 789–797 (2009).
- 280 [9] Wilber, M. Q., Kitzes, J., and Harte, J. *Global Ecology and Biogeography*  
281 **24**(8), 883–895 (2015).
- 282 [10] Olival, K. J., Hosseini, P. R., Zambrana-Torrel, C., Ross, N., Bogich, T. L.,  
283 and Daszak, P. *Nature* **546**(7660), 646–650 (2017).
- 284 [11] Johnson, C. K., Hitchens, P. L., Evans, T. S., Goldstein, T., Thomas, K.,  
285 Clements, A., Joly, D. O., Wolfe, N. D., Daszak, P., Karesh, W. B., et al.  
286 *Scientific Reports* **5**, 14830 (2015).
- 287 [12] Woolhouse, M. E. and Gowtage-Sequeria, S. *Emerging Infectious Diseases*  
288 **11**(12), 1842 (2005).
- 289 [13] Luis, A. D., Hayman, D. T., O’Shea, T. J., Cryan, P. M., Gilbert, A. T.,  
290 Pulliam, J. R., Mills, J. N., Timonin, M. E., Willis, C. K., Cunningham,  
291 A. A., et al. *Proc. R. Soc. B* **280**(1756), 20122753 (2013).
- 292 [14] Brook, C. E. and Dobson, A. P. *Trends in microbiology* **23**(3), 172–180 (2015).
- 293 [15] Han, B. A., Kramer, A. M., and Drake, J. M. *Trends in parasitology* **32**(7),  
294 565–577 (2016).
- 295 [16] Levinson, J., Bogich, T. L., Olival, K. J., Epstein, J. H., Johnson, C. K.,  
296 Karesh, W., and Daszak, P. *Emerging Infectious Diseases* **19**(5), 743 (2013).
- 297 [17] Young, C. C. and Olival, K. J. *PLoS One* **11**(2), e0149237 (2016).
- 298 [18] Restif, O., Hayman, D. T., Pulliam, J. R., Plowright, R. K., George, D. B.,  
299 Luis, A. D., Cunningham, A. A., Bowen, R. A., Fooks, A. R., O’Shea, T. J.,  
300 et al. *Ecology Letters* **15**(10), 1083–1094 (2012).
- 301 [19] Grubaugh, N. D., Gangavarapu, K., Quick, J., Matteson, N. L., De Jesus,  
302 J. G., Main, B. J., Tan, A. L., Paul, L. M., Brackney, D. E., Grewal, S., et al.  
303 *bioRxiv* , 383513 (2018).
- 304 [20] Dallas, T., Park, A. W., and Drake, J. M. *PLoS computational biology* **13**(5),  
305 e1005557 (2017).
- 306 [21] Elmasri, M., Farrell, M., and Stephens, D. A. *arXiv preprint*  
307 *arXiv:1707.08354* (2017).

- 308 [22] Shi, M., Lin, X.-D., Chen, X., Tian, J.-H., Chen, L.-J., Li, K., Wang, W.,  
309 Eden, J.-S., Shen, J.-J., Liu, L., et al. *Nature* **556**(7700), 197 (2018).
- 310 [23] Anthony, S. J., Epstein, J. H., Murray, K. A., Navarrete-Macias, I., Zambrana-  
311 Torrelío, C. M., Solovyov, A., Ojeda-Flores, R., Arrigo, N. C., Islam, A.,  
312 Khan, S. A., et al. *MBio* **4**(5), e00598–13 (2013).
- 313 [24] Robertson, C. *Flowers and insects lists of visitors of four hundred and fifty*  
314 *three flowers*. The Science Press Printing Company, (1929).
- 315 [25] Marlin, J. C. and LaBerge, W. E. *Conservation Ecology* **5**(1), 9 (2001).
- 316 [26] Schleuning, M., Blüthgen, N., Flörchinger, M., Braun, J., Schaefer, H. M.,  
317 and Böhning-Gaese, K. *Ecology* **92**(1), 26–36 (2011).
- 318 [27] NCEAS Interaction Web Database. <https://www.nceas.ucsb.edu/interactionweb/index.html>.  
319 Accessed: 2018-09-01.
- 320 [28] Toju, H., Tanabe, A. S., and Sato, H. *Microbiome* **6**(1), 116 (2018).
- 321 [29] Dallas, T. *Ecography* **39**(4), 391–393 (2016).
- 322 [30] Dallas, T. A., Aguirre, A. A., Budischak, S., Carlson, C., Ezenwa, V., Han,  
323 B., Huang, S., and Stephens, P. R. *Global Ecology and Biogeography* **0**(0).
- 324 [31] Millard, S. P. *EnvStats: An R Package for Environmental Statistics*. Springer,  
325 New York, (2013).
- 326 [32] Schult, D. A. In *In Proceedings of the 7th Python in Science Conference*  
327 *(SciPy)*, 11–15, (2008).

# Figures

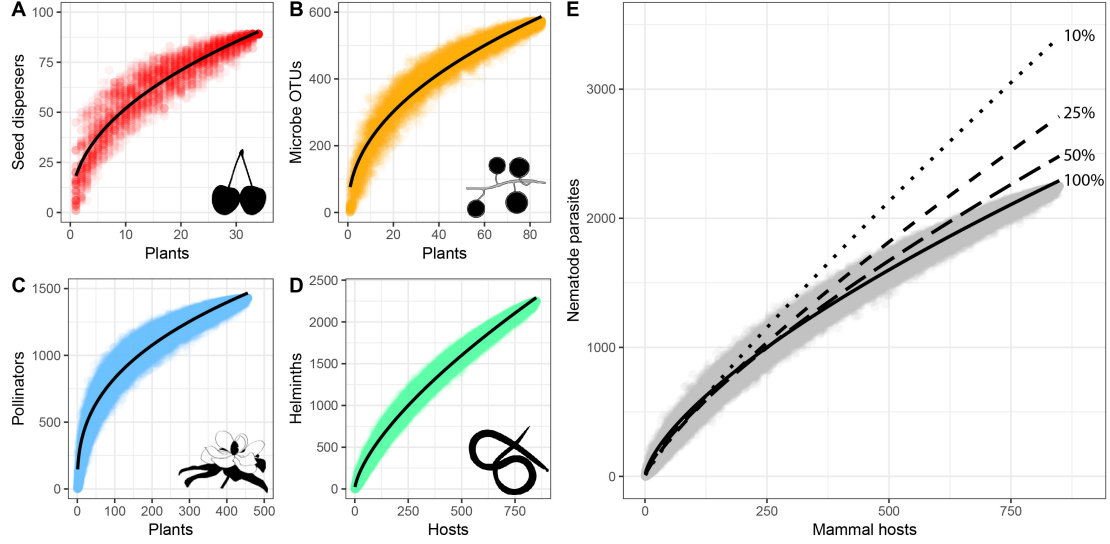


Figure 1: Fitting power law relationships between affiliates and host diversity, with shape  $A \propto bH^z$  (where  $z = 1$  is linear scaling). The power law scaling of affiliate and host richness for four networks of species interactions: plant-seed disperser (A;  $z = 0.45$ ), plant-arbuscular 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.

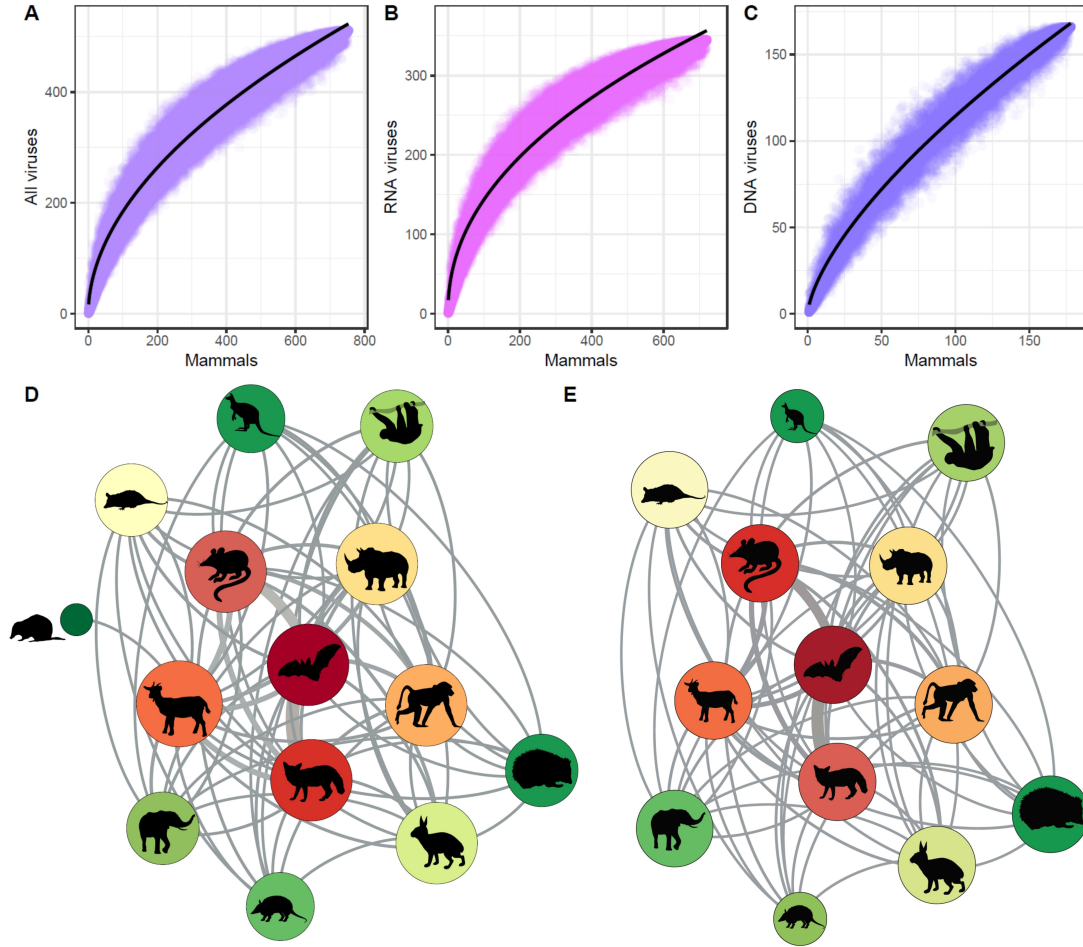


Figure 2: Bipartite rarefaction curves on the known viral network constructed from a single iteration with 100 samples. Points show each sub-sample of the total network, and curves were fitted for all viruses (A;  $z = 0.517$ ), RNA viruses (B;  $z = 0.460$ ), and DNA viruses (C;  $z = 0.667$ ). DNA 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. This pattern is consistent for the entire network (D) and for a sub-network of only zoonotic viruses (E). Node size is proportional to degree. Edge weight is proportional to the number of viruses shared between two orders. 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.

## 329 Supplementary Material

330

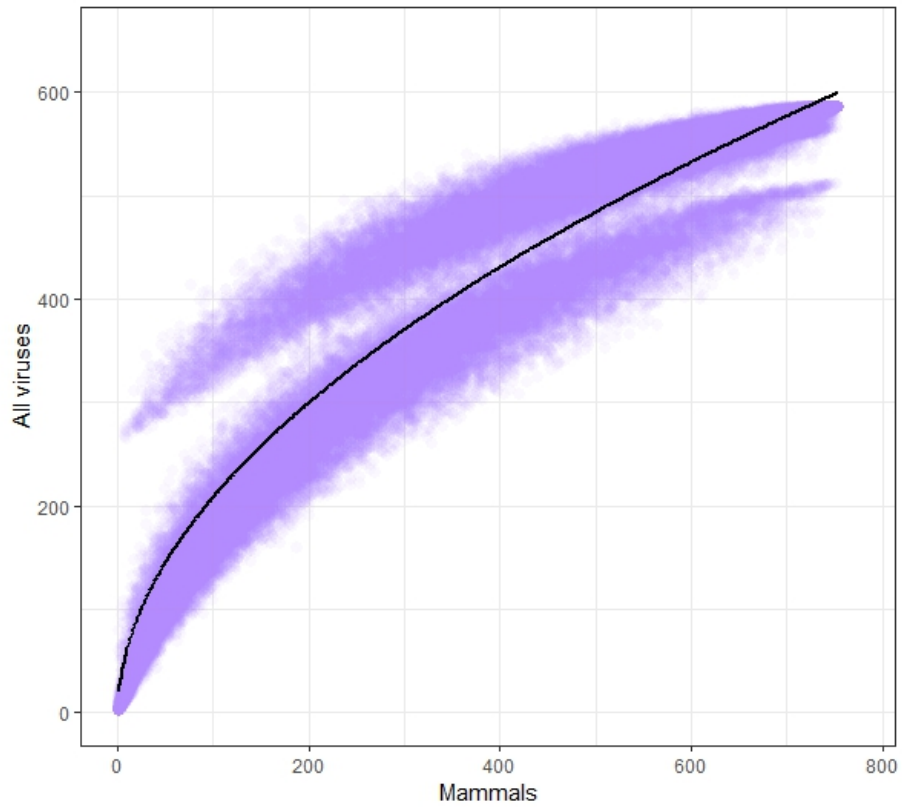

















Figure S1: Bipartite rarefaction curves on the known viral network, if humans are included. Curves bifurcate due to the atypically high degree of *Homo sapiens*, leading to a poor fit in the overall model.

Table S1. Network properties of orders and key to images from Figure 2.

Order icon	Order name	All virus network			Zoonotic virus network		
		Degree	Eigenvector centrality	Node strength	Degree	Eigenvector centrality	Node strength
	Carnivora	12	0.539	475.9	12	0.540	429.4
	Cetartiodactyla	13	0.282	280.2	12	0.230	228.2
	Chiroptera	12	0.620	636.6	12	0.638	633.8
	Cingulata	9	0.025	35.3	9	0.026	35.3
	Didelphimorphia	10	0.047	60.8	10	0.049	60.8
	Diprotodontia	9	0.007	12.4	9	0.007	12.4
	Eulipotyphla	9	0.007	10.8	9	0.006	10.8
	Lagomorpha	12	0.053	57.2	12	0.054	56.8
	Peramelemorphia	1	1.67x10 <sup>-4</sup>	3.0	0	N/A	0
	Perissodactyla	12	0.104	106.75	12	0.088	81.0
	Pilosa	10	0.037	50.0	10	0.038	50.0
	Primates	12	0.127	122.1	12	0.131	119.9
	Proboscidea	10	0.034	45.5	10	0.026	31.4
	Rodentia	12	0.459	469.7	12	0.464	451.2
	Scandentia	0	N/A	0	0	N/A	0