nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-019-0910-6

Global estimates of mammalian viral diversity accounting for host sharing

Colin J. Carlson*, Casey M. Zipfel, Romain Garnier and Shweta Bansal

Present estimates suggest there are over 1 million virus species found in mammals alone, with about 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 non-linear relationship, produced by patterns of host sharing among virus species. To account for host sharing, we fit a power law scaling relationship for host–virus species interaction networks. We estimate that there are about 40,000 virus species in mammals (including ~10,000 viruses with zoonotic potential), a reduction of two orders of magnitude from present projections of viral diversity. We expect that the increasing availability of host–virus association data will improve the precision of these estimates and their use in the sampling and surveillance of pathogens with pandemic potential. We suggest host sharing should be more widely included in macroecological approaches to estimating biodiversity.

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 well-established. On the other hand, invertebrates and microbes, which harbour most of the global diversity of life, continue to pose a challenge. By some calculations, most 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[4–7].

To work around these challenges, several methods have been developed that estimate the richness of 'affiliate' groups like parasites or mutualists on the basis of 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 can only be appropriately used to estimate the number of host-specific species[8]. A few 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]:

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

Although 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 (Fig. 1). The reason for this non-linearity can be described intuitively: the 1,000th host sampled will on average share more parasites with the first 999 hosts than the tenth host will share with the first nine. In network terms, the total number of possible parasites each host can add is governed by the degree distribution of hosts, while the expected number of parasites is reduced on the basis of host sharing, that is the probability each potential new parasite shares a host with a parasite already sampled. This scaling pattern 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 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 working to catalogue global viral diversity, with the ultimate aim of predicting outbreaks and preventing pandemics. Recent 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 projections is the exclusion of host-sharing patterns from these methods (see Supplementary 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 available[14]. With 511 viruses catalogued from 753 mammals

Department of Biology, Georgetown University, Washington, DC, USA. *e-mail: cjc322@georgetown.edu

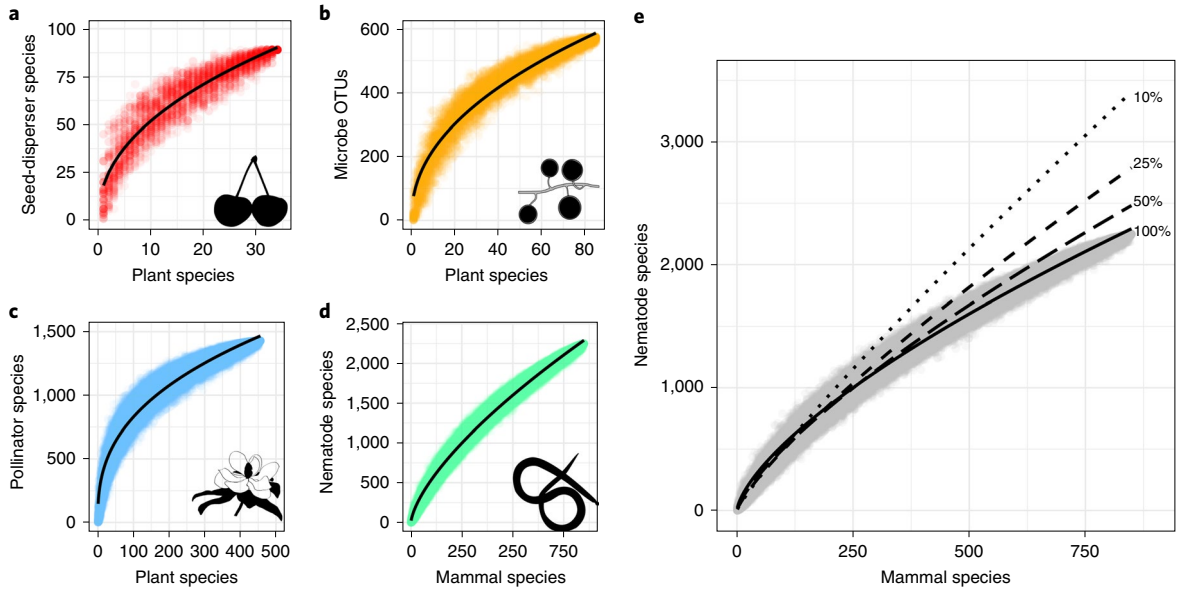

Fig. 1 | Fitting power law relationships between affiliates and host diversity, with shape $A \propto bH^z$. a–d, The power law scaling of affiliate and host richness for four networks of species interactions: plant–seed disperser ($z = 0.45$, **a**), plant–arbuscular mycorrhizae ($z = 0.46$, **b**), plant–pollinator ($z = 0.38$, **c**) and mammal–nematode ($z = 0.67$, **d**). Each point shows a network subsample used to fit the total model; $z = 1$ is linear scaling. **e**, At lower sampling levels, the same curves approach linearity, which we show in panel **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 overestimate the size of the entire network. Curves are each built with 100 iterations. OTU, operational taxonomic units. Credit: Silhouettes are from http://phylopic.org/ (Gareth Monger (nematode)) (https://creativecommons.org/licenses/by/3.0/)

(excluding humans), the network covers about 10% 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 on the basis of 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 Figs. 1 and 2 that each network produced a non-linear, concave sampling curve. To describe this pattern, we compared the fit of three simple functional forms (linear, power law and logarithmic) and compared model performance using Akaike information criterion (AIC). We reproduced the result of ref. [10] that described these curves as power law (Supplementary Figs. 2, 3; Supplementary Tables 2, 3). Examining the residuals, we noted a non-random effect, where residuals were lower (the curves overpredicted) at the lowest and highest values (Supplementary Fig. 4). One important consequence of this error structure was that power laws fit to a smaller portion of the network produced higher estimates (see Fig. 1e), with z values closer to 1. We took advantage of this property to create what we called 'upper bound' estimates, using 50% of the network to generate estimates that reflect an upper bound on the overall size of the network.

Given the pattern we observed in the residuals, we investigated whether more complex forms of power laws might better describe the data given this curvature, with six candidate models drawn from the species–area relationship literature (Supplementary Table 4). We found that these models improved predictions, with the quadratic power law most often ranking with the lowest AIC; but no one model consistently performed the best and more complex functional forms fit to these curves all produced substantially lower estimates of viral diversity (Supplementary Table 5, 6). In the absence of analytic expectations, we erred towards parsimony and limited our main results to classical power laws, noting that they were probably prone to overestimation, even when using 100% of the network.

To estimate mammalian viral diversity, we iteratively resample mammal–virus associations and use the same power law approach (Fig. 2a–c). 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% confidence interval: 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 in the association network. At the broadest level, RNA viruses generally have a higher host breadth (number of host species that can be infected) than DNA viruses, which should reduce their scaling exponent[14]; we observed this reduction in the curves we generated (Fig. 2a–c). We thus evaluate the richness for RNA and DNA viruses separately before combining them. We estimated a total 26,315 DNA viruses and 14,573 RNA viruses or a total 40,878 viruses together (Table 1). Even using the 50% upper bound method, which produces a substantially increased estimate, we only calculated 55,784 possible total viruses, still less than 10% of previous estimates (Table 2). Furthermore, diversity is distributed non-homogeneously among different pairs of host and viral families, with some host groups

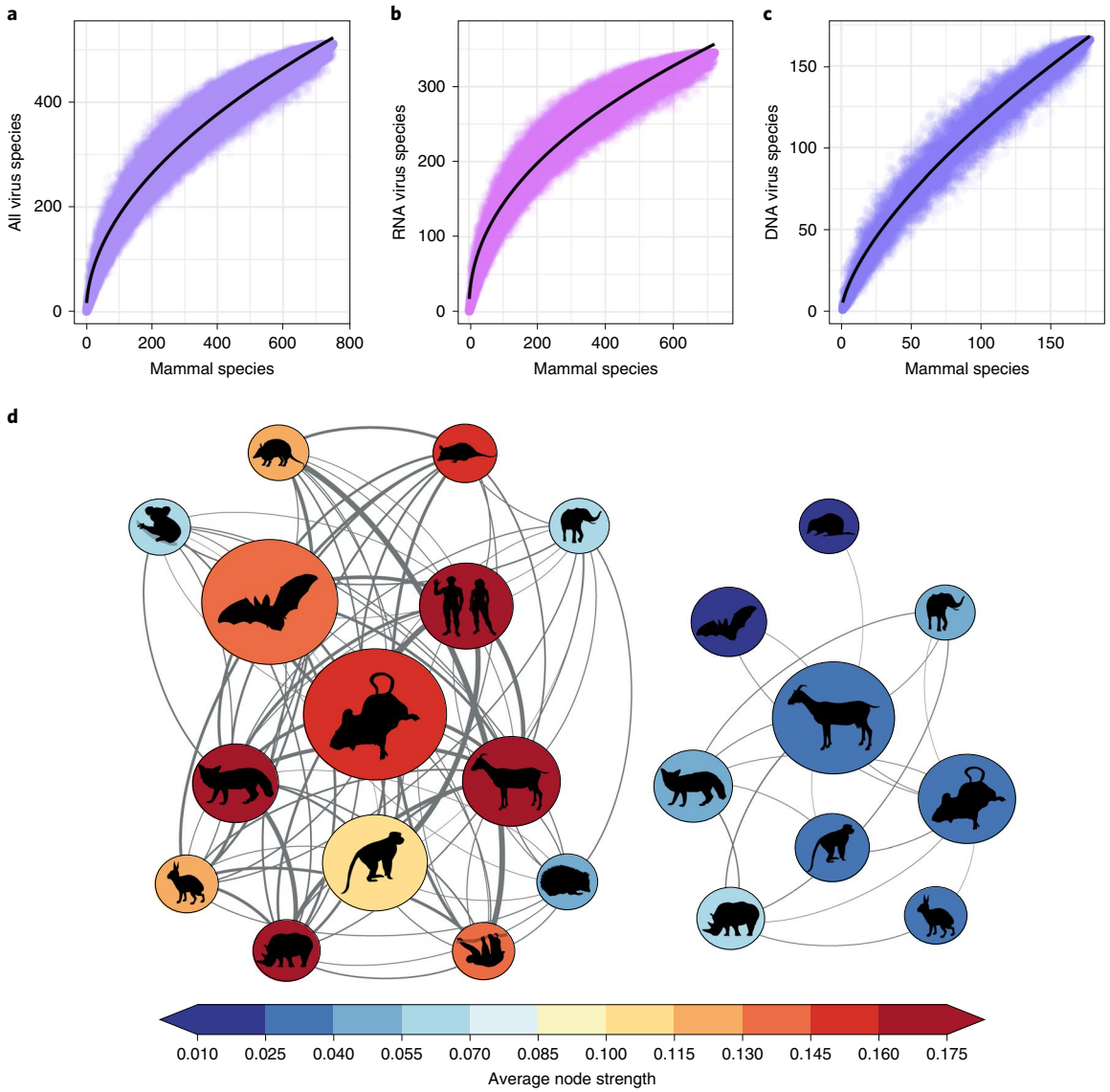

Fig. 2 | Bipartite rarefaction curves on the known viral network. a–c, Points show each subsample of the total network, and curves were fitted for all viruses ($z = 0.517$, **a**), RNA viruses ($z = 0.460$, **b**) and DNA viruses ($z = 0.667$, **c**). DNA viruses are more host specific and thus the rarefaction curve is closer to linear. **d**, Viral sharing is unevenly distributed across the network, with a few groups (bats, primates, ungulates, rodents and carnivores) accounting for most of viral sharing; 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 et al. dataset[14]. Edge weight is proportional to the Jaccard index for viruses shared between host groups. Curves were constructed from 100 samples (in-text estimates use 1,000). Node colour 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 blue is low average strength. Credit: Silhouettes are from http://phylopic.org/ (Rebecca Groom (fox), Sarah Werning (opossum, koala, sloth and elephant), Roberto Díaz Sibaja (hedgehog) and Jan A. Venter, Herbert H.T. Prins, David A. Balfour and Rob Slotow (rhinoceros)) (https://creativecommons.org/licenses/by/3.0/)

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 (Fig. 2d).

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 493,856–689,285[13] (Table 1). Even using the 50% estimation method as an upper bound, we estimate a total of 12,941 zoonoses; although higher, this is still approximately 2–3% of previous estimates (Table 2).

Discussion

Network science is useful for studying biotic interactions in modern ecology and offers powerful ways to understand data such as host–virus associations[16]. Here, we highlight how a simple scaling property of bipartite networks enables a method of estimating diversity for affiliate groups like parasites and pathogens. Using our computational approach, we found that global viral diversity in mammals has probably been overestimated by about two orders of magnitude due to the omission of host-sharing patterns.

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

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 2 | Estimation of viral diversity using the lower 50% of the subsampled curve

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)

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 (for example, a power law or exponential degree distribution), what is the expected scaling of edges in subsamples drawn randomly from a 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 (Fig. 1e, Supplementary Information). This is a promising topic for future research in mathematics and network science and 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 about 40,000 viruses in mammals, of which about 10,000 have zoonotic potential. Whereas previous estimates assumed 289.5 unique virus species per host, our study suggests there are about 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 corrects for undersampling of viruses per host, it does not account for the likelihood that host breadth is also being underestimated, which might further reduce richness estimates. Our broader finding that viral diversity has probably been overestimated is congruent with the limited 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 (ref. [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 in (and potentially between) hosts[20], the need for completeness makes the problem of cataloguing viral diversity exponentially more intensive. Targeting specific groups may make this problem more manageable: groups like bats, rodents and primates harbour disproportionate viral richness, even accounting for sampling bias due to their high zoonotic rate[16,21–23]. Moreover, zoonotic viruses in these groups may account for much of viral sharing over broad phylogenetic distances[15,24]. Focusing on describing viral sharing in and among these groups might reduce the effort needed to approximate the overall level of host sharing in the network and 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 (Fig. 2d). Even in 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. Future 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 in it.

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]. 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]. Although 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-analysing data that have been previously used to provide a linear estimate by Carroll et al.[13]. (Previous estimates are described in the Supplementary Information).

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 1,429 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 datasets were obtained from NCEAS 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; 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* is included or not in a given subsample (Supplementary Fig. 1). Consequently, we removed humans from all 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 *P. giganteus* and *M. mulatta* viral metagenomics and by constructing a rarefaction curve over the number of animals sampled (as in ref. [13]).

Bipartite richness estimators. We developed a new *R* package, 'codependent'[39], 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 non-linear 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% confidence interval.

For our viral richness estimates for mammals, we resampled a curve with every number of hosts (i) between 1 and $H = 753$, each $n = 1{,}000$ times and used the copredict function to project to 5,291 total mammal species. We repeated this 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[14] to obtain total estimates of zoonotic viral richness. The true proportion of viruses with zoonotic potential may be higher, as many viruses have yet to emerge in human populations or it may be lower, as zoonotic viruses sampled from hyper-reservoirs make up a disproportionate share of known viral diversity. The total number of zoonotic viruses is still bounded in the 0% and 100% of total viral richness estimates, which are 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 Fig. 1); this makes this confidence bound method an absolute outer bound on plausible richness. For example, using the helminth network in Fig. 1, fitting a curve with $n = 100$ iterations each gives an estimate of 2,291 nematode species (95% confidence interval: 2,271–2,311) 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[14] 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[14] (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% confidence interval of 3.4–7.4%. While these rates should be derived from a larger and representative sample of species, we note that the bat and macaque datasets are unusual 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.[13], 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' (in the sense of ref. [40]) and non-random coevolutionary diversification. It is probably 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 there are several interacting and difficult-to-quantify biases probably contained in this host–virus association dataset[41]. Ideally, at a minimum, we would 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% confidence interval: 18.1–26.0%) and an RNA virus sampling rate of 7.2% (95% confidence interval: 3.3–8.8%); both independent estimates reduce the total unsampled viral diversity. In Supplementary Table 7 we show how using these numbers would reduce overall estimates.

Network analyses. To generate a unipartite network of host sharing by viruses, we analysed associations between viruses and their hosts. We classified hosts by their orders (separating out *H. 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 in the same order (that is, self links were removed). Edges were weighted proportional to the Jaccard index[42], which is defined by

$$J = \frac{C}{A + B - C} \quad (2)$$

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. Also, there were 116 viruses with more than one order recorded and 86 zoonotic viruses with more than one order recorded. Networks were constructed and analysed using the networkx package in Python[43] and visualizations were constructed with Gephi[44] and PhyloPic (http://www.phylopic.org).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data in this study are from previous studies and available online for researchers to reproduce our results. Original sources can be found as follows: global diversity of viruses in mammal hosts can be found in Carroll et al.[13]; plant–pollinator interactions can be found in Robertson[32] and reproduced in Marlin et al.[33]; myccorhizal networks are described in Toju et al.[36]; and the host–helminth network can be obtained from the Natural History Museum London's Helminth Database, through the helminthR API (ref. [37]). All data are also available on the Github repository for the project, at github.com/cjcarlson/brevity

Code availability

All code is available on a Github repository found at github.com/cjcarlson/brevity. The codependent *R* package[39] is available at github.com/cjcarlson/codependent

Received: 15 October 2018; Accepted: 23 April 2019;
Published online: 10 June 2019

References

1. Colwell, R. K. et al. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *J. Plant Ecol.* **5**, 3–21 (2012).
2. Larsen, B. B., Miller, E. C., Rhodes, M. K. & Wiens, J. J. Inordinate fondness multiplied and redistributed: the number of species on earth and the new pie of life. *Q. Rev. Biol.* **92**, 229–265 (2017).
3. Windsor, D. A. Controversies in parasitology: most of the species on earth are parasites. *Int. J. Parasitol.* **28**, 1939–1941 (1998).
4. Bacher, S. Still not enough taxonomists: reply to Joppa et al. *Trends Ecol. Evol.* **27**, 65–66 (2012).
5. Colwell, R. K. & Coddington, J. A. Estimating terrestrial biodiversity through extrapolation. *Philos. Trans. R. Soc. Lond. B* **345**, 101–118 (1994).
6. Poulin, R. & Morand, S. *Parasite Biodiversity* (Smithsonian, 2004).
7. Quicke, D. L. We know too little about parasitoid wasp distributions to draw any conclusions about latitudinal trends in species richness, body size and biology. *PLoS One* **7**, e32101 (2012).
8. May, R. M. How many species? *Philos. Trans. R. Soc. Lond. B* **330**, 293–304 (1990).
9. Dobson, A., Lafferty, K. D., Kuris, A. M., Hechinger, R. F. & Jetz, W. Homage to Linnaeus: how many parasites? how many hosts? *Proc. Natl Acad. Sci. USA* **105**, 11482–11489 (2008).
10. Strona, G. & Fattorini, S. Parasitic worms: how many really? *Int. J. Parasitol.* **44**, 269–272 (2014).
11. Delmas, E. et al. Analysing ecological networks of species interactions. *Biol. Rev.* **94**, 16–36 (2019).
12. Pellissier, L. et al. Comparing species interaction networks along environmental gradients. *Biol. Rev.* **93**, 785–800 (2018).

13. Carroll, D. et al. The global virome project. *Science* **359**, 872–874 (2018).
14. Olival, K. J. et al. Host and viral traits predict zoonotic spillover from mammals. *Nature* **546**, 646–650 (2017).
15. Johnson, C. K. et al. Spillover and pandemic properties of zoonotic viruses with high host plasticity. *Sci. Rep.* **5**, 14830 (2015).
16. Gómez, J. M., Nunn, C. L. & Verdú, M. Centrality in primate–parasite networks reveals the potential for the transmission of emerging infectious diseases to humans. *Proc. Natl Acad. Sci. USA* **110**, 7738–7741 (2013).
17. Harte, J., Smith, A. B. & Storch, D. Biodiversity scales from plots to biomes with a universal species–area curve. *Ecol. Lett.* **12**, 789–797 (2009).
18. Wilber, M. Q., Kitzes, J. & Harte, J. Scale collapse and the emergence of the power law species–area relationship. *Glob. Ecol. Biogeogr.* **24**, 883–895 (2015).
19. Ignacio-Espinoza, J. C., Solonenko, S. A. & Sullivan, M. B. The global virome: not as big as we thought? *Curr. Opin. Virol.* **3**, 566–571 (2013).
20. Grubaugh, N. D. et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using primalseq and ivar. *Genome Biol.* **20** https://doi.org/10.1186/s13059-018-1618-7 (2019).
21. Luis, A. D. et al. A comparison of bats and rodents as reservoirs of zoonotic viruses: are bats special? *Proc. Biol. Sci.* **280**, 20122753 (2013).
22. Brook, C. E. & Dobson, A. P. Bats as 'special' reservoirs for emerging zoonotic pathogens. *Trends Microbiol.* **23**, 172–180 (2015).
23. Han, B. A., Kramer, A. M. & Drake, J. M. Global patterns of zoonotic disease in mammals. *Trends Parasitol.* **32**, 565–577 (2016).
24. Woolhouse, M. E. & Gowtage-Sequeria, S. Host range and emerging and reemerging pathogens. *Emerging Infect. Dis.* **11**, 1842 (2005).
25. Levinson, J. et al. Targeting surveillance for zoonotic virus discovery. *Emerging Infect. Dis.* **19**, 743 (2013).
26. Young, C. C. & Olival, K. J. Optimizing viral discovery in bats. *PLoS One* **11**, e0149237 (2016).
27. Restif, O. et al. Model-guided fieldwork: practical guidelines for multidisciplinary research on wildlife ecological and epidemiological dynamics. *Ecol. Lett.* **15**, 1083–1094 (2012).
28. Dallas, T., Park, A. W. & Drake, J. M. Predicting cryptic links in host–parasite networks. *PLoS Comput. Biol.* **13**, e1005557 (2017).
29. Elmasri, M., Farrell, M., & Stephens, D. A. A hierarchical Bayesian model for predicting host–parasite interactions using phylogenetic information. Preprint at https://arxiv.org/abs/1707.08354 (2017).
30. Shi, M. et al. The evolutionary history of vertebrate RNA viruses. *Nature* **556**, 197 (2018).
31. Geoghegan, J. L. et al. Hidden diversity and evolution of viruses in market fish. *Virus Evol.* **4**, vey031 (2018).
32. Robertson, C. *Flowers and Insects* (Science Press, 1929).
33. Marlin, J. C. & LaBerge, W. E. The native bee fauna of Carlinville, Illinois, revisited after 75 years: a case for persistence. *Conserv. Ecol.* **5**, 9 (2001).
34. Schleuning, M. et al. Specialization and interaction strength in a tropical plant–frugivore network differ among forest strata. *Ecology* **92**, 26–36 (2011).
35. *Interaction Web Database* (NCEAS, accessed 1 September 2018); http://www.nceas.ucsb.edu/interactionweb.
36. Toju, H., Tanabe, A. S. & Sato, H. Network hubs in root-associated fungal metacommunities. *Microbiome* **6**, 116 (2018).
37. Dallas, T. helminthr: an r interface to the London Natural History Museum's host–parasite database. *Ecography* **39**, 391–393 (2016).
38. Dallas, T. et al. Gauging support for macroecological patterns in helminth parasites. *Glob. Ecol. Biogeogr.* **27**, 1437–1447 (2018).
39. Carlson, C. J. codependent: an R package for network-based estimation of affiliate species richness. Version 1.0 https://github.com/cjcarlson/codependent (2019).
40. Jordano, P. Sampling networks of ecological interactions. *Funct. Ecol.* **30**, 1883–1893 (2016).
41. Lloyd-Smith, J. O. Infectious diseases: predictions of virus spillover across species. *Nature* **546**, 603 (2017).
42. Pilosof, S., Morand, S., Krasnov, B. R. & Nunn, C. L. Potential parasite transmission in multi-host networks based on parasite sharing. *PLoS One* **10**, e0117909 (2015).
43. Schult, D. A. Exploring network structure, dynamics, and function using NetworkX. In *Proc. 7th Python in Science Conference (SciPy2008)* (Eds Varoquaux, G. et al.) 11–15 (SciPy, 2008); https://conference.scipy.org/proceedings/scipy2008/paper_2/full_text.pdf.
44. Bastian, M., Heymann, S., & Jacomy, M. Gephi: an open source software for exploring and manipulating networks. Version 0.9.2 (International AAAI Conference on Weblogs and Social Media, 2009).

Acknowledgements

We thank T. A. Dallas, P. P. A. Staniczenko, T. Poisot, A. Barner and three anonymous reviewers for thoughtful comments and discussion about the manuscript and the methodology. We also acknowledge T. A. 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 and by a Georgetown Environment Initiative fellowship to C.J.C.

Author contributions

C.J.C., C.M.Z., R.G. and S.B. conceived of the study. C.J.C. and C.M.Z. performed all analyses. All authors contributed to the writing and approved the final draft.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-019-0910-6.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to C.J.C.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Carlson

Last updated by author(s): Mar 25, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All data in this study is from previous studies and is available online for researchers to reproduce our results. All data are also available on the Github repository for the project, at github.com/cjcarlson/brevity.
Data analysis	All code is available on a Github repository found at github.com/cjcarlson/brevity. The codependent R package is available at github.com/cjcarlson/codependent.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data in this study is from previous studies and is available online for researchers to reproduce our results. Original sources can be found as follows: global diversity of viruses in mammal hosts can be found in Carroll et al. (2018); plant-pollinator interactions can be found in Robertson (1929) and reproduced in Marlin (2001); myccorhizal networks are described in Toju et al. (2018); and the host-helminth network can be obtained from the Natural History Museum London's Helminth Database, through the helminthR API All data are also available on the Github repository for the project, at github.com/cjcarlson/brevity.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We generated curves from the size of iteratively subsampled bipartite networks to extrapolate global viral diversity.
Research sample	We reused several large, open data sources that describe ecological networks in the process, including the largest available database of host-virus associations for mammals. We generated no original primary data in the study.
Sampling strategy	We subsampled bipartite networks iteratively, selecting sample sizes based on a number far past the sample size at which the models converged. Sample size was standardized and reported for all analyses.
Data collection	Data were collected by previous studies, and can be found in the Github repository for the study.
Timing and spatial scale	N/A (ecological association data drawn from the literature)
Data exclusions	No data were excluded from analyses except for human-viral associations, which were disproportionately represented in the data and prevented curve fitting.
Reproducibility	All code is made available to ensure reproducibility.
Randomization	Not applicable to computational study.
Blinding	Not applicable to computational study.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging