Data Proliferation, Reconciliation, and Synthesis in Viral Ecology

RORY GIBB, GREGORY F. ALBERY, DANIEL J. BECKER, LIAM BRIERLEY, RYAN CONNOR, TAD A. DALLAS, EVAN A. ESKEW, MAXWELL J. FARRELL, ANGELA L. RASMUSSEN, SADIE J. RYAN, AMY SWEENY, COLIN J. CARLSON, AND TIMOTHÉE POISOT

The fields of viral ecology and evolution are rapidly expanding, motivated in part by concerns around emerging zoonoses. One consequence is the proliferation of host–virus association data, which underpin viral macroecology and zoonotic risk prediction but remain fragmented across numerous data portals. In the present article, we propose that synthesis of host–virus data is a central challenge to characterize the global virome and develop foundational theory in viral ecology. To illustrate this, we build an open database of mammal host–virus associations that reconciles four published data sets. We show that this offers a substantially richer view of the known virome than any individual source data set but also that databases such as these risk becoming out of date as viral discovery accelerates. We argue for a shift in practice toward the development, incremental updating, and use of synthetic data sets in viral ecology, to improve replicability and facilitate work to predict the structure and dynamics of the global virome.

Keywords: viral ecology, disease ecology, virus, zoonotic risk, data synthesis

The emergence of SARS-CoV-2 was a harsh reminder that uncharacterized wildlife viruses can suddenly become globally relevant. Efforts to identify wildlife viruses with the potential to infect humans and to predict spillover and emergence trajectories are becoming more popular than ever (including with major scientific funders). However, the value of these efforts is limited by an incomplete understanding of the global virome (Wille et al. 2021). Significant knowledge gaps exist regarding the mechanisms of viral transmission and replication, host–pathogen associations and interactions, spillover pathways, and several other dimensions of viral emergence. Furthermore, although billions of dollars have been invested in these scientific challenges over the last decade alone, much of the data relevant to these problems remains unsynthesized. Fragmented data access and a lack of standardization preclude an easy reconciliation process across data sources, making the whole less than the sum of its parts and hindering viral research (Wyborn et al. 2018).

In the present article, we propose that data synthesis is a seminal challenge for translational work in viral ecology. This requires researchers to go beyond the usual steps of data collection and publication and to develop a community of practice that prioritizes data synthesis and reconciles semireproduced work across different teams and disciplines. As an illustrative example, we describe the analytical hurdles of working with host–virus association data, a format that characterizes the global virome as a bipartite network of hosts and viruses, with pairs connected by observed potential for infection. Recent studies highlight the central role for these data in efforts to understand viral macroecology and evolution (Carlson et al. 2019, Dallas et al. 2019, Albery et al. 2020), to predict zoonotic emergence risk (Han et al. 2015, 2016, Olival et al. 2017, Wardeh et al. 2020), and to anticipate the impacts of global environmental change on infectious disease (https://doi.org/10.1101/2020.01.24.918755 [preprint: not peer reviewed], Gibb et al. 2020, Johnson et al. 2020). Several bespoke data sets have been compiled to address these questions, each of which differs in sources and scope. Scientific knowledge of the global host–virus network is continually evolving as a consequence of novel discoveries, changing research priorities and taxonomic revision, and as interest in this field has grown, so has the fragmentation of total knowledge across these data sets. To illustrate this problem (and a simple solution), we compare and reconcile four major host–virus association data sets, each of which is different enough that we anticipate the results of individual studies could be strongly shaped by choice of data set.

Four snapshots of one host–virus network

Although host–pathogen association data exist in dozens of sources and repositories, there are four particularly large and widely used published data sets, which each capture between 0.3% and 1.5% of the estimated 50,000 species of mammal

BioScience XX: 1–9.
https://doi.org/10.1093/biosci/biab080

Table 1. Available "big data" on host–virus associations, and major features of each data set.

Data set	Source	Nature of data set	Association records	Host species	Virus species	Original taxonomic scope of pathogens	Original taxonomic scope of hosts	Diagnostic method identified (PCR, serology, etc.)?	URL of current version
GMPD2	University of Georgia	Static	895	226	154	All parasites and pathogens (including viruses, bacteria, macroparasites, protozoans, prions)	Mammals (subset: only ungulates, carnivores, and primates)	Yes	http://onlinelibrary.wiley.com/doi/10.1002/ecy.1799/suppinfo
EID2[a]	University of Liverpool	Dynamic	1,342	418	398	All symbionts (including viruses, bacteria, macroparasites, protozoans, prions, green algae, molluscs, and cnidarians)	Vertebrates and invertebrates	No	https://eid2.liverpool.ac.uk/
HP3	EcoHealth Alliance	Static	2,784	751	561	Viruses	Mammals	Yes	https://github.com/ecohealthalliance/HP3
Shaw	Shaw LP and colleagues (2020).	Static	4,210	957	733	Viruses and bacteria	Vertebrates	Yes	https://doi.org/10.6084/m9.figureshare.8262779

Note: Numbers of unique association records and host, virus, and pathogen species are all derived from the reconciled version presented in the CLOVER database, and therefore these numbers may differ from those presented in the main text (which are taken from the source data, or from self-reporting by the data curators). [a]Number of associations and taxa accurate as of 2015 static release in *Scientific Data* paper.

viruses (Carlson et al. 2019). Individually, each of these data sets forms the basis for numerous studies in host–pathogen ecology and macroecology, and the differences between them—especially with regards to taxonomic scope, available metadata, and frequency of data updates—make them preferable for different purposes (table 1). However, these differences may also complicate cross-comparison and synthetic inference.

GMPD 2.0. The Global Mammal Parasite Database (GMPD; Nunn and Altizer 2005), started in 1999 and now in its second public version (Stephens et al. 2017), emerged from efforts to compile mammal–parasite association data from published literature sources. Construction of the GMPD used a variety of similar strategies that combined host Latin names with a string of parasite-related terms to search online literature databases. Pertinent literature was then manually identified and relevant association and metadata were compiled. The initial database was focused on primate hosts (Nunn and Altizer 2005) and expanded to include separate sections for ungulates (Ezenwa et al. 2006) and carnivores (Lindenfors et al. 2007).

In 2017, GMPD 2.0 was released, which merged these three previously independent databases (Stephens et al. 2017). The updated data set encompasses 190 primate, 116 ungulate, and 158 carnivore species, and records their interactions with 2412 unique "parasite" species, including 189 viruses, as well as bacteria, protozoa, helminths, arthropods, and fungi. Notable improvements GMPD 2.0 are the construction of a unified parasite taxonomy that bridges occurrence records across host taxa, the expansion of host–parasite association data along with georeferencing, and enhanced parasite trait data (e.g., transmission mode).

The original data are available as a web resource (www.mammalparasites.org), and the data from GMPD 2.0 can also be downloaded as static files from a data paper (Stephens et al. 2017). In addition, one subsection of the GMPD, named the Global Primate Parasite Database, has been independently maintained and regularly updated by Charles Nunn (data available at https://parasites.nunn-lab.org). Consequently, the primate subsection of GMPD 2.0 includes papers published up to 2015, whereas the ungulate and carnivore subsections stop after 2010 (Stephens et al. 2017).

EID2. The ENHanCEd Infectious Diseases Database (EID2), curated by the University of Liverpool, may be the largest dynamic data set of any symbiotic interactions (Wardeh et al. 2015). EID2 is regularly compiled from automated scrapes of two web sources: publication titles and abstracts indexed in the PubMed database and the National Center for Biotechnology Information (NCBI) Nucleotide Sequence database (along with its associated taxonomic metadata). The EID2 data is structured using the concepts of carrier and cargo rather than host and pathogen, because it includes a number of ecological interactions beyond the

Box 1. Glossary.

Association data: a format that records ecological interactions between a host and symbiont (an *association*) in the form of an edge list.

Data provenance: The primary literature origin of a particular record or set of records in a synthetic data set.

Data reconciliation: the task of harmonizing the language of a given data set's fields and metadata to allow a researcher to merge data of different provenance, and generate a new synthetic product.

Edge list: a table, spreadsheet, or matrix of "links" in a host–symbiont network, where each row records the known association of a different host–symbiont pair.

Flat file: a static document in Excel or similar spreadsheet or data format, with no dynamic component (no updating) and all data available from a single file rather than a query interface.

Metadata: additional data describing focal data of interest and that is relevant to interpretation and analysis. Important examples for host–virus associations include sampling method (for example, serological assay, PCR or pathology), date and geographical location of sampling, and standardized information on host and virus taxonomy.

Open data: data that is directly and freely accessible for reuse and exploration without impediment, gatekeeping, or cost restriction.

scope of normal host–pathogen interactions, including potentially unresolved mutualist or commensal associations. Interactions are stored as a geographic edge list, where each carrier and cargo can also have locality information; additional metadata include the number of sequences in GenBank and related publications.

EID2's dynamic web interface (currently available through download on a limited, query-by-query basis that researchers often manually bind or by personal correspondence with data curators) to date contains information encompassing 1560 mammal carrier species and 3986 microparasite or macroparasite cargo species, of which 1446 are viruses (Wardeh et al. 2020). However, many researchers continue to use the static, open release of EID2 from a 2015 data paper (Wardeh et al. 2015), which we focus on in the present article for comparative purposes as a stable version of the database available to the community of practice. The EID2 data were originally validated for completeness against GMPD 1.0.

HP3. The Host–Parasite Phylogeny Project data set (HP3) was developed by EcoHealth Alliance over the better part of a decade. Published along with a landmark analysis of the correlates of zoonotic potential (data from Olival et al. 2017), the HP3 data set consists of 2805 associations between 754 mammal hosts and 586 virus species. These were compiled from literature published between 1940 and 2015, on the basis of targeted searches of online reference databases. Complementary with the search strategy used for the GMPD, rather than starting with a list of host names, HP3 started with names of known mammal viruses listed in the International Committee on Taxonomy of Viruses (ICTV) database. These virus names along with their synonyms were then used as search terms to identify literature containing host–virus association data.

Data collection and cleaning for HP3 began in 2010, and the database has been static since 2017; it can be obtained as a flat file in the published study's data repository (Olival et al. 2017). HP3 includes a host–virus edge list (see box 1), separate files for host and virus taxonomy, and separate files for host and virus traits. Host–virus association records are provided with a note about method of identification (polymerase chain reaction [PCR], serological methods, etc.), which may be useful for researchers interested in the different levels of confidence ascribed to particular associations (https://doi.org/10.1101/2020.05.22.111344 [preprint: not peer reviewed]). HP3's internal taxonomy is also harmonized with two mammal trees (Bininda-Emonds et al. 2007, Fritz et al. 2009), facilitating analyses that seek to account for host phylogenetic structure while testing hypotheses about viral ecology and evolution (e.g., Becker et al. 2020, https://doi.org/10.1101/2020.02.25.965046 [preprint: not peer reviewed], Olival et al. 2017, Washburne et al. 2018, Guth et al. 2019, Park 2019, Albery et al. 2020, Mollentze and Streicker 2020). HP3 was also validated against GMPD 1.0.

Shaw. In recent work, Shaw and colleagues (2020) built a host–pathogen edge list by combining a systematic literature search with cross-validation from several of the above-mentioned data sets. Similar to the construction of HP3, Shaw and colleagues (2020) started with lists of known pathogenic bacteria and viruses found in humans and animals. They then conducted Google Scholar searches pairing pathogen names with disease-related keywords, followed by manual review of search results. For well-studied pathogens, they limited their manual review to a subset of the top 200 most relevant publications as determined by Google. From the resulting literature searches, Shaw's team compiled 12,212 interactions between 2656 vertebrate host species (including, but not limited to, mammals) and 2595 viruses and bacteria. GMPD2, EID2, and the Global Infectious Diseases and Epidemiology Network Guide to Medically Important Bacteria (Gideon Informatics and Berger 2020) were used to validate the host–pathogen associations.

The data set is available as a static flat file through figshare and the project GitHub repository (Shaw et al. 2020). Host–pathogen associations are provided alongside pathogen

metadata (e.g., genome size, bacterial traits, transmission mode, zoonotic status) and diagnostic method (i.e., PCR, pathogen isolation, pathology). The data set al.o includes a comprehensive host phylogeny, developed specifically for the study using nine mitochondrial genes for downstream analyses of host phylogenetic similarity and host breadth.

A reconciled mammalian virome data set

Some of these data sets were validated against each other during production, and others have been used for cross-validation in analytical work (Albery et al. 2020), and certain studies have generated a study-specific ad hoc reconciled data set (https://doi.org/10.1101/2020.02.25.965046 [preprint: not peer reviewed], Gibb et al. 2020). However, no work has been published with the primary aim of reconciling them as correctly, comprehensively, and reproducibly as possible. More recently developed data sets such as Shaw's can inherently draw on a greater cumulative body of scientific work. This could mean they include most of the data captured by previous efforts, but we found there are substantial differences among all four data sets. In isolation, we expect that these differences could affect ecological and evolutionary inference in ways that are difficult to quantify, with special relevance to significance thresholds in hypothesis-testing research (i.e., different data sets may confer different power to statistical tests). We expected that separate host–virus data sources could be standardized into one shared format, allowing them to cover a greater percentage of the global virome, a greater diversity of host species, and obviating the need for researchers to either choose between individual data sets or implement ad hoc solutions that merge them prior to analysis.

To illustrate the potential for comprehensive data reconciliation, we harmonized the four major data sets described in the present article, creating a new synthetic CLOVER data set out of the four leaves (which we have made available with this study). Doing this required harmonizing and standardizing both host and virus taxonomy, as well as metadata describing the strength of evidence for interactions. This process involved several steps applied to each source data set. First, we manually harmonized virus names across all four data sets to revolve subtle formatting differences. Second, we applied a standardized scheme of virus detection methods using information provided in each source data set (described further below). Finally, using the R package taxize (Chamberlain and Szöcs 2013), we accessed the most current binomial for each host species and applied a standardized host and virus taxonomy (species, genus, family, order, and class) using the same taxonomic hierarchy (Schoch et al. 2020) as the NCBI's Taxonomy database (ncbi.nlm.nih.gov). Host (n = 34) and virus (n = 24) species that did not return an exact automated match (i.e., fuzzy matches) were manually checked and resolved where possible against the NCBI Taxonomy database (or against the International Union for Conservation of Nature Red List database, https://iucnredlist.org, for 14 mammal species without a match in the NCBI Taxonomy database). All virus names are given at the species level even if finer classifications exist, and viruses that could not be resolved to species are resolved to the next-lowest taxonomic level (genus or family, although all original reported names are retained and accessible from the column "VirusOriginal"). Host and virus names, metadata, NCBI unique taxonomic identifiers, virus ICTV ratification status, and primary data sources as originally described were included in the combined data set, to ensure traceability.

With all four data sets taxonomically consistent, we were able to show that each only covered a portion of the known global mammalian virome, even for the most studied hosts and viruses (figure 1). Our taxonomic harmonization helped reconcile some discrepancies, increasing overlap among the data sets (figure 2), but notable differences remained. This could confound inference: For example, using a simple linear model, we found that data provenance (see box 1) explained 8.8% of variation in host species' viral diversity (but only 4.7% after harmonization). When viral ecology studies report different findings based on slight variation around a significance threshold, readers should therefore consider whether subtle differences in the underlying data sets might account for such variation.

Integrated data sets move us a step closer to resolving this uncertainty. The CLOVER data set covers 1085 mammal host species and 831 associated viruses. This only represents 16.9% of extant mammals (Burgin et al. 2018) and, at most, 2.1% of their viruses (Carlson et al. 2019)—a marginal improvement over the 957 mammal hosts (14.9%) and 733 viruses (1.8%) in the reconciled Shaw data subset but an improvement nonetheless. The biggest functional gain is not in the *breadth* of the reconciled data but in its *depth*: the Shaw database records 4209 interactions among these host and virus species, whereas CLOVER captures 5477. Given that previous studies have estimated that 20%–40% of host–parasite links are unknown (in GMPD2 (Dallas et al. 2017)), this 30% improvement is notable and shows the value of data synthesis: Both building out and filling in synthetic data sets will significantly improve the performance of statistical models, which are usually heavily confounded by matrix sparseness (https://doi.org/10.1101/2020.05.22.111344 [preprint: not peer reviewed], Dallas et al. 2017).

In addition, harmonization of metadata on virus detection methods across data sets enables a greater scrutiny of the strength of evidence in support of each host–virus association. We applied a simplified detection method classification scheme (i.e., either serology, PCR or sequencing, isolation or observation, or method unknown) based on descriptions in the source databases or, where these are not provided, adopted the most conservative definition given the data source in question (i.e., EID2 entries derived from the NCBI Nucleotide database are classified under PCR or sequencing, although they might also qualify for the next strongest level of isolation or observation,

Figure 1. Network representation of the CLOVER data set. The nodes of the entire CLOVER network have been projected to a two-dimensional space using t-SNE, and disaggregated to each of the four data sources. In each panel, only the nodes found in the given data set are shown with filled symbols (the unfilled symbols indicate associations recorded in the other data sets); the triangles represent mammal hosts, whereas the circles represent viruses. In each data set, a nontrivial proportion of associations is completely unique and unrecorded elsewhere, even after taxonomic reconciliation. This was the case for 186 of 1342 associations in EID2 (13.8%), 611 of 2783 in HP3 (22%), 271 of 895 in GMPD2 (30.3%), and 1707 of 4210 in Shaw (40.5%).

whereas entries derived from PubMed are classified under method unknown). Of the 5477 unique host–virus pairs in CLOVER, a total of 2160 (39%) have been demonstrated using either viral isolation or direct observation and 1871 (34%) via PCR or sequencing-based methods (with some overlap, because some associations have been reported with both of the above methods). Notably, a substantial proportion (2256, 41%) are based solely on serological evidence, which, although it is an indicator of past exposure, does not reflect host competence (i.e., effectiveness at transmitting a pathogen; Gilbert et al. 2013, Lachish and Murray 2018, Becker et al. 2020). Such harmonized metadata facilitate investigation of inferential stability using various types of evidence, as well as enabling a best practice of subsetting data for a particular research purpose. For example, serological assays are a much weaker form of evidence if the aim of a study is zoonotic reservoir host prediction, whereas virus isolation data open new avenues for testing hypotheses about reservoir competence (https://www.biorxiv.org/content/10.1101/2021.01.01.425052v1).

Data synthesis inherently relies on a scientific community that generates new, often conflicting, data. The generation of truly novel data and finding ways to resolve existing observations that are in conflict are two equally viable paths to scientific knowledge production. However, in the current funding landscape, researchers may have a significant incentive to position themselves as creating an entirely "novel" data set from scratch, even if it partially replicates available data sources, or to focus their limited resources on data sets that improve the depth of knowledge within a narrow scope (e.g., a focus on specific taxonomic groups). But when testing microbiological or ecoevolutionary hypotheses, rather than simply using the newest published data set as a benchmark for which one is most up to date, we suggest a necessary shift in scientific cultural norms toward using synthetic, reconciled data as an analytical best practice. As an example, two studies have already used CLOVER to advance the science of viral ecology: One showed that the apparently higher diversity of zoonotic pathogens in urban-adapted mammals is likely a consequence of sampling bias (https://www.biorxiv.org/content/10.1101/2021.01.02.425084v1 [preprint: not peer reviewed]), whereas another showed that a two-step process of network imputation and graph embedding can be used to substantially improve a model that identifies zoonotic viruses on the basis of their genome composition (https://arxiv.org/abs/2105.14973 [preprint: not peer reviewed]).

To make this kind of work possible, at least a handful of researchers will need to continue the task of stepwise integration, using data sets that synthesize existing knowledge across teams, institutions, and funding programs to fill in critical data with even more detail. The required tasks (e.g., identifying relevant source data, cleaning taxonomic information, harmonizing metadata on diagnostic information or spatiotemporal structure) can be time consuming but are relatively straightforward to conduct and can increasingly be automated thanks to the rapid growth of new tools for reproducible research (Boettiger et al. 2015, Lowndes et al. 2017, Colella et al. 2020).

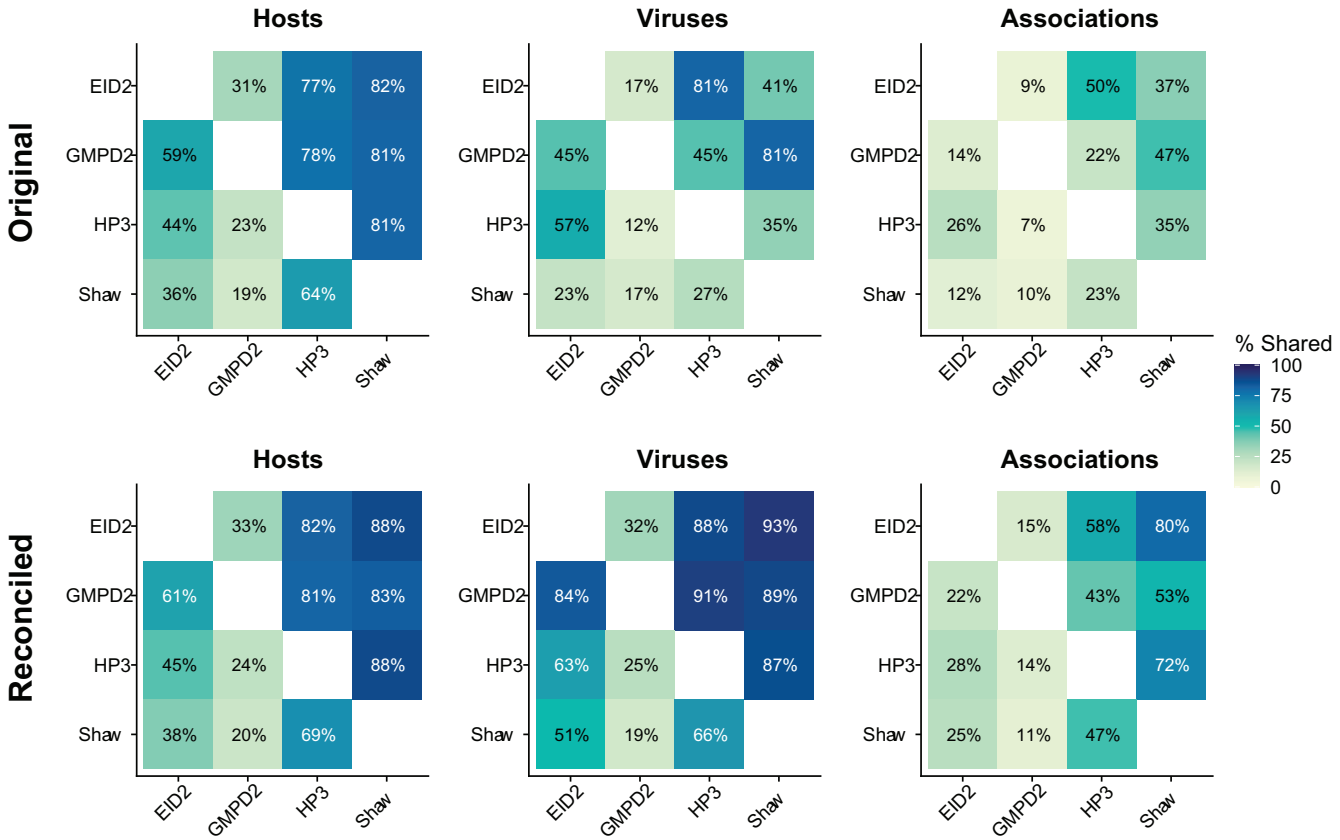

Figure 2. Proportional overlap between data sets before and after host and virus taxonomic reconciliation. The percentages and fill colors in these tiles can be interpreted as the percentage of the y-axis that was contained in the x-axis; for example, 31% of originally reported EID2 hosts were also represented in GMPD2, whereas 47% of reconciled Shaw associations were also contained in HP3. The darker colors represent higher proportions of shared data.

There is a clear need and no obvious technical barrier to invest more effort in data harmonization: Engaging in this process as a form of open science will accelerate progress for the entire research community.

Relevance to future efforts

In the present article, we showed that a simple data synthesis effort can create a dramatically more comprehensive data set of mammal–virus associations. However, this is a temporary solution and one that is becoming less sustainable given global investments in accelerating the rates of viral discovery in wildlife (Wille et al. 2021). Even if similar data sets continue to proliferate or if newer iterations of existing data sets are periodically released, static data sets will quickly become out of date, and their relation to the most recent empirical knowledge will be left unclear. This is already a significant issue with the CLOVER data set, which becomes much sparser after 2010, both in terms of the overall number of reported host–virus associations, and the reporting of novel (i.e., previously undetected) associations (figure 3a, 3b). This sparseness is most likely because of time lags between host–virus sampling in the field, the reporting or publication of associations and their eventual inclusion in one of the component data sets and suggests that CLOVER may now be missing up to a decade's worth of complete host–virus data. This gap is concerning, given that the last decade has seen unprecedented and exponential growth in viral discovery and research effort in wildlife (figure 3c).

In the near term, microbiologists and data scientists may therefore need to approach the task of data reconciliation with a much broader scope and develop a more sustainable data platform—one that is dynamic, and minimizes the time between scientific discoveries and their documentation in an aggregate data source. The reconciliation process we describe in the present article will need to evolve in order to power these kinds of databases; to integrate data sources that update every day (e.g., NCBI's GenBank database or the Global Biotic Interactions database), the taxonomic reconciliation process cannot rely on manual curation steps such as those undertaken to generate CLOVER. The development of automated taxonomic pipelines is not an unfamiliar challenge in ecological data synthesis, but it poses a particular problem with respect to viral taxonomy, which is in a constant state of flux. Often, a substantial lag between virus discovery and official ratification by the International Committee on the Taxonomy of Viruses (ICTV) exacerbates

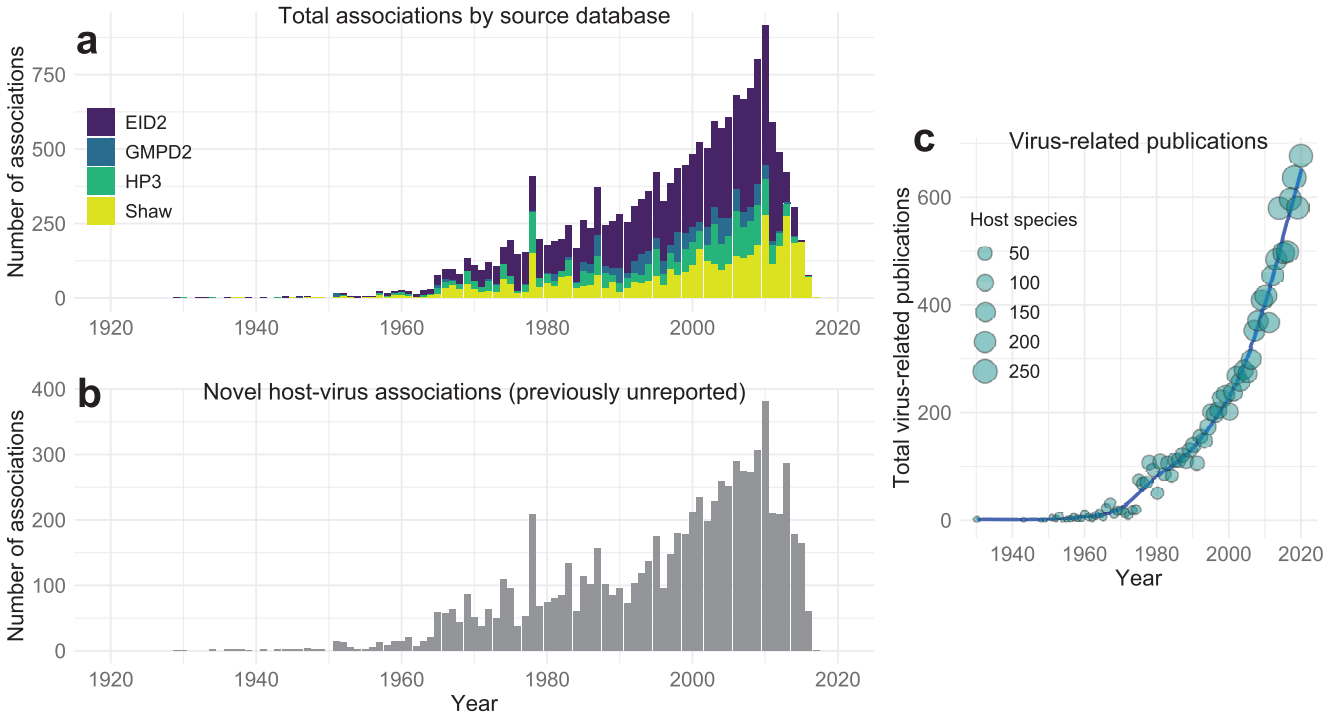

Figure 3. Temporal trends in host–virus association reports and virus-related research effort. The bar graphs show, for each year, the annual number of reported associations color-coded by source database, which can include duplicates of the same association reported over multiple years (a), and the number of novel unique associations (i.e., unreported before that year) (b). The years reflect the date when an association was reported, either in a published paper or report (for literature-based records) or to the NCBI Nucleotide database (EID2 only). The trend plot (c) shows the trend in virus-related publications across all hosts in the CLOVER data set up to 2020 (PubMed search term: "host binomial and virus or viral"). The points represent the annual total publications summed across all host species, and point size denotes the number of host species with virus-related publications in a given year.

the gulf between scientific knowledge and available data. Furthermore, the global virome is not simply one static, incompletely characterized entity; viruses evolve more rapidly than most targets of biodiversity databases, and the continual emergence of new lineages through reassortment and recombination unfortunately implies that host–virus associations are not a static property that can be captured through snapshots of the system (Shi et al. 2018).

Given these problems, databases might even be forced in the long term to move away from the familiar format of species concepts and toward data structures based on operational taxonomic units (OTUs). Although an OTU-based host–virus network would be better tailored to the underlying virology, it will require the incorporation of genetic sequence data, which comes with additional logistical challenges in terms of both data curation and the logistics and governance of data sharing. In the coming decade, these kinds of radical solutions may be unavoidable.

Steps toward an atlas of the global virome

Scaling up the aggregation of host–virus association data will not be easy, but is not an insurmountable endeavor. We suggest working backward from the intended end product: The goals outlined in the present article are best served by a central system (with an online access point to the consumable data), spanning the information available from multiple data sources (which demands backend engines drawing from existing databases while data provenance is tracked and proper attribution is ensured). Furthermore, the most valuable data resource would be easily updatable by practitioners (which demands a portal for manual user input or an integrated publishing toolkit to work from flat files). For users, these data should be accessible in a programmatic way (through a web API allowing for bulk download or other interfaces such as an R package), encourage reproducibility (through versioning of the entire database, or of a specific user query), and offer predictable formats (through a data specification standard devised by a multidisciplinary group).

Fortunately, the field of ecoinformatics has the capacity to help inform this design and development process. Massive bioinformatic data portals such as the Global Biodiversity Informatics Facility (gbif.org), the Encyclopedia of Life (eol.org), and the Ocean Biodiversity Information System

(obis.org) all offer most of the functionalities we outline in the present article, although they are aimed at slightly different forms of biodiversity data. More recent contributions dedicated to ecological network data include GloBI (for *global biotic interactions*; Poelen et al. 2014), helminthR (Dallas 2016), and mangal (Poisot et al. 2016), all of which reconcile their taxonomy with other databases through the use of unique taxon keys. In short, researchers interested in the global virome need not divert their attention, resources, and effort away from the pressing tasks related to monitoring viral pathogens. Rather, they can leverage existing products, expertise, and capacity in neighboring fields to bolster their ability to do so. Given the eagerness ecologists have shown to participate in SARS-CoV-2 research, we anticipate that our field may be especially well poised to jump into this task after the pandemic. We aim, in our current efforts, to lay that groundwork: The CLOVER database is the first step toward a project called the Virome in One Network, a prototype of the next-generation database described in the present article.

An atlas of the global virome would have inherent value for the entire scientific community. When the format of a data set is well established, it allows for the development of tools that mine the data in real time. For example, the field of biodiversity studies has adopted the concept of essential biodiversity variables, which can be updated when the underlying data change (Pereira et al. 2013, Fernández et al. 2019, Jetz et al. 2019). Having the ability to revisit predictions about the host–virus network could improve models that assess zoonotic potential of wildlife viruses (https://doi.org/10.1101/2020.02.25.965046 [preprint: not peer reviewed], https://doi.org/10.1101/2020.11.12.379917 [preprint: not peer reviewed]), generate priority targets for wildlife reservoir sampling (Becker et al. 2020, Babayan et al. 2018, Plowright et al. 2019), and help benchmark model performance related to these tasks. Beyond training and validation, link prediction models built on these reconciled databases may be used to target future literature searches, shifting from systematic literature searches to a model-based approach to database updating. Increased collaboration between data collectors, data managers, and data scientists that leads to better data standardization and reconciliation is the only way to productively synthesize our knowledge of the global virome.

Acknowledgments

This work was supported by funding to the Viral Emergence Research Initiative consortium, including National Science Foundation grant no. BII 2021909 and a grant from the Institut de Valorisation des Données. The authors thank Noam Ross, Maya Wardeh, and many others for formative conversations about these data sets and for their tireless work making those data available to the research community.

Supplemental material

The four raw data sets and harmonized CLOVER data set can be obtained from the archived link https://zenodo.org/record/4945274. Code used to generate the analyses and figures in this study can be found at https://github.com/viralemergence/reconciliation.

References cited

Albery GF, Eskew EA, Ross N, Olival KJ. 2020. Predicting the global mammalian viral sharing network using phylogeography. Nature communications 11: 2260.

Babayan SA, Orton RJ, Streicker DG. 2018. Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes. Science 362: 577–580.

Becker DJ, Seifert SN, Carlson CJ. 2020. Beyond infection: Integrating competence into reservoir host prediction. Trends in Ecology and Evolution 35: 1062–1065.

Bininda-Emonds ORP, Cardillo M, Jones KE, MacPhee RDE, Beck RMD, Grenyer R, Price SA, Vos RA, Gittleman JL, Purvis A. 2007. The delayed rise of present-day mammals. Nature 446: 507–512.

Boettiger C, Chamberlain S, Hart E, Ram K. 2015. Building software, building community: Lessons from the rOpenSci Project. Journal of Open Research Software 3: e8.

Burgin CJ, Colella JP, Kahn PL, Upham NS. 2018. How many species of mammals are there? Journal of Mammalogy 99: 1–14.

Carlson CJ, Zipfel CM, Garnier R, Bansal S. 2019. Global estimates of mammalian viral diversity accounting for host sharing. Nature Ecology and Evolution 3: 1070–1075.

Chamberlain SA, Szöcs E. 2013. taxize: Taxonomic search and retrieval in R. F1000Research 2: 191.

Colella JP, Stephens RB, Campbell ML, Kohli BA, Parsons DJ, Mclean BS. 2020. The Open-Specimen Movement. BioScience 71: 405–414.

Dallas T. 2016. helminthR: an R interface to the London natural history museum's host–parasite database. Ecography 39: 391–393.

Dallas TA, Han BA, Nunn CL, Park AW, Stephens PR, Drake JM. 2019. Host traits associated with species roles in parasite sharing networks. Oikos 128: 23–32.

Dallas T, Park AW, Drake JM. 2017. Predicting cryptic links in host–parasite networks. PLOS Computational Biology 13: e1005557.

Ezenwa VO, Price SA, Altizer S, Vitone ND, Cook KC. 2006. Host traits and parasite species richness in even and odd-toed hoofed mammals, Artiodactyla and Perissodactyla. Oikos 115: 526–536.

Fernández N, Guralnick R, Daniel Kissling W. 2019. A minimum set of information standards for essential biodiversity variables. Biodiversity Information Science and Standards 3: e35212.

Fritz SA, Bininda-Emonds ORP, Purvis A. 2009. Geographical variation in predictors of mammalian extinction risk: Big is bad, but only in the tropics. Ecology letters 12: 538–549.

Gibb R, Redding DW, Chin KQ, Donnelly CA, Blackburn TM, Newbold T, Jones KE. 2020. Zoonotic host diversity increases in human-dominated ecosystems. Nature 584: 398–402.

Gideon Informatics, Berger S. 2020. GIDEON Guide to Medically Important Bacteria. GIDEON Informatics.

Gilbert AT, et al. 2013. Deciphering serology to understand the ecology of infectious diseases in wildlife. EcoHealth 10: 298–313.

Guth S, Visher E, Boots M, Brook CE. 2019. Host phylogenetic distance drives trends in virus virulence and transmissibility across the animal–human interface. Philosophical Transactions of the Royal Society B 374: 20190296.

Han BA, Kramer AM, Drake JM. 2016. Global patterns of zoonotic disease in mammals. Trends in Parasitology 32: 565–577.

Han BA, Schmidt JP, Bowden SE, Drake JM. 2015. Rodent reservoirs of future zoonotic diseases. Proceedings of the National Academy of Sciences 112: 7039–7044.

Jetz W, et al. 2019. Essential biodiversity variables for mapping and monitoring species populations. Nature Ecology and Evolution 3: 539–551.

Johnson CK, Hitchens PL, Pandit PS, Rushmore J, Evans TS, Young CCW, Doyle MM. 2020. Global shifts in mammalian population trends reveal

key predictors of virus spillover risk. Proceedings of the Royal Society B 287: 20192736.

Lachish S, Murray KA. 2018. The certainty of uncertainty: Potential sources of bias and imprecision in disease ecology studies. Frontiers in Veterinary Science 5: 90.

Lindenfors P, Nunn CL, Jones KE, Cunningham AA, Sechrest W, Gittleman JL. 2007. Parasite species richness in carnivores: Effects of host body mass, latitude, geographical range and population density. Global Ecology and Biogeography 16: 496–509.

Lowndes JSS, Best BD, Scarborough C, Afflerbach JC, Frazier MR, O'Hara CC, Jiang N, Halpern BS. 2017. Our path to better science in less time using open data science tools. Nature Ecology and Evolution 1: 160.

Mollentze N, Streicker DG. 2020. Viral zoonotic risk is homogenous among taxonomic orders of mammalian and avian reservoir hosts. Proceedings of the National Academy of Sciences 117: 9423–9430.

Nunn CL, Altizer SM. 2005. The global mammal parasite database: An online resource for infectious disease records in wild primates. Evolutionary Anthropology 14: 1–2.

Olival KJ, Hosseini PR, Zambrana-Torrelio C, Ross N, Bogich TL, Daszak P. 2017. Host and viral traits predict zoonotic spillover from mammals. Nature 546: 646–650.

Park AW. 2019. Phylogenetic aggregation increases zoonotic potential of mammalian viruses. Biology Letters 15: 20190668.

Pereira HM, et al. 2013. Essential biodiversity variables. Science 339: 277–278.

Plowright RK, Becker DJ, Crowley DE, Washburne AD, Huang T, Nameer PO, Gurley ES, Han BA. 2019. Prioritizing surveillance of Nipah virus in India. PLOS Neglected Tropical Diseases 13: e0007393.

Poelen JH, Simons JD, Mungall CJ. 2014. Global biotic interactions: An open infrastructure to share and analyze species-interaction data sets. Ecological Informatics 24: 148–159.

Poisot T, Baiser B, Dunne JA, Kéfi S, Massol F, Mouquet N, Romanuk TN, Stouffer DB, Wood SA, Gravel D. 2016. Mangal: Making ecological network analysis simple. Ecography 39: 384–390.

Schoch CL, et al. 2020. NCBI Taxonomy: A comprehensive update on curation, resources and tools. Database 2020: baaa062.

Shaw LP, Wang AD, Dylus D, Meier M, Pogacnik G, Dessimoz C, Balloux F. 2020. The phylogenetic range of bacterial and viral pathogens of vertebrates. Molecular Ecology 29: 3361–3379.

Shi M, et al. 2018. The evolutionary history of vertebrate RNA viruses. Nature 556: 197–202.

Stephens PR, et al. 2017. Global mammal parasite database, version 2.0. Ecology 98: 1476.

Wardeh M, Risley C, McIntyre MK, Setzkorn C, Baylis M 2015. Database of host–pathogen and related species interactions, and their global distribution. Scientific Data 2: 150049.

Wardeh M, Sharkey KJ, Baylis M. 2020. Integration of shared-pathogen networks and machine learning reveals the key aspects of zoonoses and predicts mammalian reservoirs. Proceedings of the Royal Society B 287: 20192882.

Washburne AD, Crowley DE, Becker DJ, Olival KJ, Taylor M, Munster VJ, Plowright RK. 2018. Taxonomic patterns in the zoonotic potential of mammalian viruses. PeerJ 6: e5979.

Wille M, Geoghegan JL, Holmes EC. 2021. How accurately can we assess zoonotic risk? PLOS Biology 19: e3001135.

Wyborn C, et al. 2018. Understanding the impacts of research synthesis. Environmental Science and Policy 86: 72–84.

Rory Gibb (rory.j.gibb@gmail.com) is affiliated with the Centre for Mathematical Modelling of Infectious Diseases and with the Centre on Climate Change and Planetary Health, at the London School of Hygiene and Tropical Medicine, in London, England, in the United Kingdom. Gregory F. Albery is affiliated with the Department of Biology at Georgetown University, in Washington, DC, in the United States. Daniel J. Becker is affiliated with the Department of Biology at the University of Oklahoma, in Norman Oklahoma, in the United States. Liam Brierley is affiliated with the Department of Health Data Science at the University of Liverpool, in Liverpool, England, in the United Kingdom. Ryan Connor is affiliated with the National Center for Biotechnology Information, at the National Library of Medicine, in the National Institutes of Health, in Bethesda, Maryland, in the United States. Tad A. Dallas is affiliated with the Department of Biological Sciences at Louisiana State University, in Baton Rouge, Louisiana, in the United States. Evan A. Eskew is affiliated with the Department of Biology at Pacific Lutheran University, in Tacoma, Washington, in the United States. Maxwell J. Farrell is affiliated with the Department of Ecology and Evolutionary Biology at the University of Toronto, in Toronto, Ontario, Canada. Angela L. Rasmussen is affiliated with the Vaccine Infectious Disease Organization and International Vaccine Centre, at the University of Saskatchewan, in Saskatchewan, Saskatoon, Canada, and she and Colin J. Carlson (colin.carlson@georgetown.edu) are affiliated with the Center for Global Health Science and Security, at the Georgetown University Medical Center, at Georgetown University, in Washington, DC, in the United States. Sadie J. Ryan is affiliated with the Quantitative Disease Ecology and Conservation Lab, in the Department of Geography and with the Emerging Pathogens Institute at the University of Florida, in Gainesville, Florida, in the United States, and with the College of Life Sciences at the University of KwaZulu Natal, in Durban, South Africa. Amy Sweeny is affiliated with the Institute of Evolutionary Biology at the University of Edinburgh, in Edinburgh, Scotland, in the United Kingdom. Timothée Poisot is affiliated with the Département de Sciences Biologiques at the Université de Montréal, and with the Québec Centre for Biodiversity Sciences, both in Montréal, Québec, Canada. All of the authors are members of the Viral Emergence Research Initiative consortium, a global scientific collaboration to predict which viruses could infect humans, which animals host them, and where they could emerge.