Data proliferation, reconciliation, and synthesis in viral ecology

Authors: Rory Gibb[1,2*], Gregory F. Albery[3], Daniel J. Becker[4], Liam Brierley[5], Ryan Connor[6], Tad A. Dallas[7], Evan A. Eskew[8], Maxwell J. Farrell[9], Angela L. Rasmussen[10,15], Sadie J. Ryan[11,12,13], Amy Sweeny[14], Colin J. Carlson[15*], and Timothée Poisot[16,17]

1. Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene and Tropical Medicine, London, UK
2. Centre on Climate Change and Planetary Health, London School of Hygiene and Tropical Medicine, London, UK
3. Department of Biology, Georgetown University, Washington DC, USA
4. Department of Biology, University of Oklahoma, Norman OK, USA
5. Department of Health Data Science, University of Liverpool, Liverpool, UK
6. National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.
7. Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, 70806 USA
8. Department of Biology, Pacific Lutheran University, Tacoma WA, USA
9. Department of Ecology & Evolutionary Biology, University of Toronto, Toronto, Ontario, Canada
10. Vaccine Infectious Disease Organization and International Vaccine Centre, University of Saskatchewan, Saskatoon, Canada
11. Quantitative Disease Ecology and Conservation (QDEC) Lab, Department of Geography, University of Florida, Gainesville, FL 32601
12. Emerging Pathogens Institute, University of Florida, Gainesville, FL 32611
13. College of Life Sciences, University of KwaZulu Natal, Durban, 4041, South Africa
14. Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK
15. Center for Global Health Science and Security, Georgetown University Medical Center, Georgetown University, Washington, D.C., U.S.A.
16. Université de Montréal, Département de Sciences Biologiques, Montréal QC, Canada
17. Québec Centre for Biodiversity Sciences, Montréal QC, Canada

*Correspondence to: Rory Gibb (rory.j.gibb@gmail.com) and Colin J. Carlson (colin.carlson@georgetown.edu)

Acknowledgements: This work was supported by funding to the Viral Emergence Research Initiative (VERENA) consortium, including NSF BII 2021909 and a grant from Institut de Valorisation des Données (IVADO). The authors thank Noam Ross, Maya Wardeh, and many others for formative conversations about these datasets and for their tireless work making those data available to the research community.

Author biography: All the authors are members of the Viral Emergence Research Initiative (VERENA) consortium, a global scientific collaboration to predict which viruses could infect humans, which animals host them, and where they could emerge.

Abstract

The fields of viral ecology and evolution have rapidly expanded in the last two decades, driven by technological improvements, and motivated by efforts to discover potentially zoonotic wildlife viruses under the rubric of pandemic prevention. One consequence has been a massive proliferation of host-virus association data, which comprise the backbone of research in viral macroecology and zoonotic risk prediction. These data remain fragmented across numerous data portals and projects, each with their own scope, structure, and reporting standards. Here, we propose that synthesis of host-virus association data is a central challenge to improve our understanding of the global virome and develop foundational theory in viral ecology. To illustrate this, we build an open reconciled mammal-virus database from four key published datasets, applying a standardized taxonomy and metadata. We show that reconciling these datasets provides a substantially richer view of the mammal virome than that offered by any one individual database. We argue for a shift in best practice towards the incremental development and use of synthetic datasets in viral ecology research, both to improve comparability and replicability across studies, and to facilitate future efforts to use machine learning to predict the structure and dynamics of the global virome.

Introduction

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. 2020). 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. Further, 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 synthetic research (Wyborn et al. 2018).

Here, 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, to develop a community of practice that prioritizes data synthesis and reconciles semi-reproduced 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 (Carlson et al. 2020, Gibb et al. 2020, Johnson et al. 2020). Several bespoke datasets have been compiled to address these questions, and as interest in these topics has grown, so has the fragmentation of total knowledge across those datasets. To illustrate this problem (and a simple solution), we compare and reconcile four major host-virus association datasets, each of which is different enough that we anticipate the results of individual studies could be strongly shaped by choice of dataset.

Four parts of one whole

Though host-pathogen association data exist in dozens of sources and repositories, there are at least four published datasets that each capture between 0.3% and 1.5% of the estimated 50,000 species of mammal viruses (Carlson et al. 2019). Differences among these datasets, especially with regards to available metadata and frequency of data updates, make them preferable for different purposes (Table 1), but may also complicate intercomparison and synthetic inference.

*GMPD 2.0***:** The Global Mammal Parasite Database (Nunn and Altizer 2005), started in 1999 and now in its second public version (Stephens et al. 2017), emerged from continued 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 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 that were being independently maintained and updated (Stephens et al. 2017). The updated dataset encompasses 190 primate, 116 ungulate, and 158 carnivore species, and record their interactions with 2,412 unique “parasite” species, including 189 viruses, as well as bacteria, protozoa, helminths, arthropods, and fungi. Notable improvements in version 2 of the GMPD 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 version 2 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, while 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 dataset of any symbiotic interactions (Wardeh et al. 2015). EID2 is compiled from automated, dynamic scrapes of two web sources: publication titles and abstracts indexed in the PubMed database and the 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, as it includes a number of ecological interactions beyond the scope of normal host-pathogen interactions, including potentially unresolved mutualist or commensal associations. Interactions are stored as a geographic edgelist, 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 which researchers often manually bind or by personal correspondence with data curators) contains information encompassing 4,799 mammal "carrier" species and 70,614 microparasite or macroparasite "cargo" species, of which 9,605 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 here 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 dataset (HP3) was developed by EcoHealth Alliance over the better part of a decade. Published along with a landmark analysis of zoonotic spillover (Olival et al. 2017), the HP3 dataset consists of 2,805 associations between 754 mammal hosts and 586 virus species. These were compiled from literature published between 1940 and 2015, based on 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. To narrow search results for well-studied viruses, they included additional host range-related terms to identify relevant publications. 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 edgelist (see Glossary), 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 (PCR, serology including specific methods, etc.), which may be useful for researchers interested in the different levels of confidence ascribed to particular associations (Becker et al. 2020). 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., Farrell et al., 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: Recent work by Shaw *et al.* built a host-pathogen edgelist by combining a systematic literature search with cross-validation from several of the above-mentioned datasets (Shaw et al. 2020). Similar to the construction of HP3, the authors 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, the authors compiled 12,212 interactions between 2,656 vertebrate host species (including, but not limited to, mammals) and 2,595 viruses and bacteria. GMPD2, EID2, and the Global Infectious Diseases and Epidemiology Network (GIDEON) Guide to Medically Important Bacteria (Gideon Informatics, Inc. and Berger 2020) were used to validate the host-pathogen associations. The dataset 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 dataset also 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 mammal virome dataset

Though some of these datasets were validated against each other during production, they are sometimes used for cross-validation in analytical work (Albery et al. 2020), and some studies have generated a study-specific *ad hoc* reconciled dataset (Farrell et al. 2020, Gibb et al. 2020), no work has been published with the primary aim of reconciling them as correctly, comprehensively, and reproducibly as possible. Dynamic datasets like EID2, and recent datasets like Shaw, can inherently draw on a greater cumulative body of scientific work. This could mean they include most of the data captured by previous efforts, yet we found there are substantial differences among all four datasets. In isolation, we expect that these differences could impact ecological and evolutionary inference in ways that are difficult to quantify, with special relevance to significance thresholds in hypothesis-testing research (i.e., different datasets may confer different power to statistical tests). In unison, we expect that these data 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 them or implement *ad hoc* solutions that merge them prior to analysis.

To illustrate the potential for comprehensive data reconciliation, we harmonized the four major datasets described here, creating a new synthetic ‘CLOVER’ dataset out of the four “leaves” (which we have made available with this study). To do so, we first harmonized the host taxonomy of all four datasets using the R package ‘taxize’ (Chamberlain and Szöcs 2013), then manually resolved remaining discrepancies. Finally, using the Julia package 'NCBITaxonomy.jl' (Poisot 2020), we standardized host and virus taxonomy against the taxonomic hierarchy (Schoch et al. 2020) used as a reference by the National Center for Biotechnology Information's Taxonomy database (ncbi.nlm.nih.gov). With all four datasets taxonomically consistent, we were able to show that each only covered a portion of the known global mammal virome, even for the most studied hosts and viruses (Figure 1). Our taxonomic harmonization helped reconcile some discrepancies, increasing overlap among the datasets (Figure 2), but notable differences remained. This could confound inference: for example, using a simple linear model, we found that **data provenance** (see Glossary) explained 8.8% of variation in host species' viral diversity (but only 4.7% after harmonization). When studies report different findings based on slight variation around a significance threshold, readers should therefore wonder whether subtle differences in the underlying datasets might account for such variation.

Integrated datasets move us a step closer to resolving this uncertainty. The CLOVER dataset covers 1,081 mammal host species and 829 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) - perhaps a marginal improvement over the 954 mammal hosts (14.9%) and 733 viruses (1.8%) in the reconciled Shaw sub-dataset, 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 4,209 interactions among these host and virus species, while CLOVER captures 5,494. 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 datasets will significantly improve the performance of statistical models, which are usually heavily confounded by matrix sparsity (Becker et al., Dallas et al. 2017).

In addition, harmonization of metadata on virus detection methods across datasets enables a greater scrutiny of the strength of evidence in support of each host-virus association. We applied a simplified detection method classification scheme (either serology, PCR/sequencing, isolation/observation, or method unknown) based on descriptions in the source databases or, where these are not provided, adopting the most conservative definition given data source (i.e., EID2 entries derived from NCBI Nucleotide are classified under PCR/sequencing, though they might also qualify for the next strongest level of isolation/observation; whereas entries derived from PubMed are classified under method unknown). Of the 5,494 unique host-virus pairs in CLOVER, a total of 2,156 (39%) have been demonstrated using either viral isolation or direct observation and 1,895 (34%) via PCR or sequencing-based methods (with some overlap, as some associations have been reported with both of the above methods). Notably, a substantial proportion (2,257; 41%) are based solely on serological evidence which, although an indicator of past exposure, does not necessarily reflect host competence (i.e. effectiveness at transmitting a pathogen; Gilbert et al. 2013, Lachish and Murray 2018, Becker et al. 2020). These harmonized definitions 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 isolation data open new avenues for testing hypotheses about reservoir competence (Becker et al. 2020).

Data synthesis inherently relies on a scientific community that generates new, often conflicting, data. The generation of truly novel data or finding ways to resolve existing observations that are in conflict are two equally viable paths to scientific progress. However, in the current funding landscape, researchers may have a significant incentive to position themselves as creating an entirely “novel” dataset from scratch, even if it partially replicates available data sources, or to focus their limited resources on datasets that improve the depth of knowledge within a narrow scope (e.g., a focus on specific taxonomic groups). But when testing microbiological or eco-evolutionary hypotheses, rather than simply using each newly-published dataset as a benchmark for which one is “most up-to-date,” we suggest a necessary shift in scientific cultural norms towards using synthetic, reconciled data like CLOVER as an analytical best practice. To make this possible, at least a handful of researchers will need to continue the task of stepwise integration, using datasets 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 data and tools for reproducible research (Boettiger et al. 2015, Lowndes et al. 2017, Colella et al. 2020). 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

Here, we showed that a simple data synthesis effort can create a dramatically more comprehensive dataset of mammal-virus associations. However, this is a temporary solution and one that will become less sustainable if similar datasets continue to proliferate or if newer iterations of existing datasets are released, each absorbing different parts of existing efforts. Over the longer term, given global investments in viral discovery from wildlife, static datasets will quickly become out-of-date, and their relation to the most recent empirical knowledge will be left unclear. For example, the CLOVER dataset becomes significantly 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 3). This sparseness is most likely due to time lags between host-virus sampling in the field, the reporting or publication of associations, and their eventual inclusion in one of the component datasets, and suggests that CLOVER may now be missing up to a decade’s worth of known host-virus data. In the near term, microbiologists and data scientists may want to approach the task of data reconciliation with a much broader scope, and develop a more sustainable data platform.

Scaling up the aggregation of host-virus association data will not be easy, but is not an insurmountable endeavour. We suggest working backwards from the intended end product: the goals outlined here 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 tracking data provenance and ensuring proper attribution). Further, 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 (i.e., through a web API allowing for bulk download and/or other interfaces like an R package), help analysts build 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 like 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 here, though they are aimed at slightly different forms of biodiversity data. More recent contributions dedicated to ecological network data include Global Biotic Interactions (Poelen et al. 2014) (GLOBI, which consumes flat files and formats them), helminthR (Dallas 2016), and mangal (Poisot et al. 2016) (which stores a metadata-rich representation of species interaction networks), 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, but they can leverage existing products, expertise, and capacity in neighbouring 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 post-pandemic. We aim, in our current efforts, to lay that groundwork.

An integrated platform for the deposition, curation, archival, and sharing of host-virus associations in a *prêt-à-manger*, metadata-rich format has inherent value for the entire scientific community. When the format of a dataset 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 (Farrell et al. 2020, Mollentze et al. 2020), generate priority targets for wildlife reservoir sampling (Becker et al., 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.

Data and code availability

The four raw datasets and harmonized CLOVER dataset can be obtained from the archived project repository: https://dx.doi.org/10.5281/zenodo.4435128. Code used to generate the analyses and figures in this study can be found at https://github.com/viralemergence/reconciliation.

References.

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, Albery GF, Sjodin AR, Poisot T, Dallas TA, Eskew EA, Farrell MJ, Guth S, Han BA, Simmons NB, Stock M, Teeling EC, Carlson CJ. Predicting wildlife hosts of betacoronaviruses for SARS-CoV-2 sampling prioritization: a modeling study.

Becker DJ, Seifert SN, Carlson CJ. 2020. Beyond Infection: Integrating Competence into Reservoir Host Prediction. Trends in Ecology & 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.

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

Carlson CJ, Albery GF, Merow C, Trisos CH, Zipfel CM. 2020. Climate change will drive novel cross-species viral transmission. bioRxiv.

Carlson CJ, Zipfel CM, Garnier R, Bansal S. 2019. Global estimates of mammalian viral diversity accounting for host sharing. Nature ecology & 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.

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.

Farrell MJ, Elmasri M, Stephens D, Jonathan Davies T. 2020. Predicting missing links in global host-parasite networks. bioRxiv preprint https://doi.org/10.1101/2020.02.25.965046

Fernández N, Guralnick R, Daniel Kissling W. 2019. A minimum set of Information Standards for Essential Biodiversity Variables. Biodiversity Information Science and Standards 3.

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

Gilbert AT, Fooks AR, Hayman DTS, Horton DL, Müller T, Plowright R, Peel AJ, Bowen R, Wood JLN, Mills J, Cunningham AA, Rupprecht CE. 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 of London. Series B, Biological sciences 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 of the United States of America 112: 7039–7044.

Jetz W, McGeoch MA, Guralnick R, Ferrier S, Beck J, Costello MJ, Fernandez M, Geller GN, Keil P, Merow C, Meyer C, Muller-Karger FE, Pereira HM, Regan EC, Schmeller DS, Turak E. 2019. Essential biodiversity variables for mapping and monitoring species populations. Nature ecology & 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. Biological sciences / The Royal Society 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 & evolution 1: 160.

Mollentze N, Babayan SA, Streicker DG. 2020. Identifying and prioritizing potential human-infecting viruses from their genome sequences. bioRxiv preprint https://www.biorxiv.org/content/10.1101/2020.11.12.379917v1.full

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 of the United States of America 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: Issues, News, and Reviews 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.

Olival KJ, Hosseini PR, Zambrana-Torrelio C, Ross N, Bogich TL, Daszak P. 2017. Data from: Host and viral traits predict zoonotic spillover from mammals. https://zenodo.org/record/807517#.YABU4RanxPZ

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

Pereira HM, Ferrier S, Walters M, Geller GN, Jongman RHG, Scholes RJ, Bruford MW, Brummitt N, Butchart SHM, Cardoso AC, Coops NC, Dulloo E, Faith DP, Freyhof J, Gregory RD, Heip C, Höft R, Hurtt G, Jetz W, Karp DS, McGeoch MA, Obura D, Onoda Y, Pettorelli N, Reyers B, Sayre R, Scharlemann JPW, Stuart SN, Turak E, Walpole M, Wegmann M. 2013. Ecology. 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 datasets. 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.

Poisot T. 2020. NCBITaxonomy.jl: Interact with the NCBI Taxonomy backbone from Julia. https://doi.org/10.5281/zenodo.4282820

Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, Mcveigh R, O'Neill K, Robbertse B, Sharma S, Soussov V, Sullivan JP, Sun L, Turner S, Karsch-Mizrachi I. 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database: the journal of biological databases and curation 2020.

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.

Shaw LP, Wang AD, Dylus D, Meier M, Pogacnik G, Dessimoz C, Balloux F. 2020. Data from: The phylogenetic range of bacterial and viral pathogens of vertebrates. https://figshare.com/articles/dataset/The_phylogenetic_range_of_bacterial_and_viral_pathogens_of_vertebrates_dataset_and_supplementary_material/8262779

Stephens PR, Pappalardo P, Huang S, Byers JE, Farrell MJ, Gehman A, Ghai RR, Haas SE, Han B, Park AW, Schmidt JP, Altizer S, Ezenwa VO, Nunn CL. 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. Biological sciences / The Royal Society 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. 2020. How accurately can we assess zoonotic risk? bioRxiv preprint https://doi.org/10.1101/2020.08.17.254961

Wyborn C, Louder E, Harrison J, Montambault J, Montana J, Ryan M, Bednarek A, Nesshöver C, Pullin A, Reed M, Dellecker E, Kramer J, Boyd J, Dellecker A, Hutton J. 2018. Understanding the Impacts of Research Synthesis. Environmental Science & Policy 86: 72–84.

Figures and Tables

Table 1. Available “big data” on host-virus associations, and major features of each dataset. 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). *Number of associations and taxa accurate as of 2015 static release in *Scientific Data* paper.

Dataset	**GMPD2**	**EID2***	**HP3**	**Shaw**
Source	U. Georgia	U. Liverpool	EcoHealth Alliance	Shaw LP, *et al*. *Molecular Ecology* (2020).
Nature of dataset	Static	Dynamic	Static	Static
Association records	893	1,360	2,783	4,207
Host species	225	415	750	954
Virus species	154	395	561	733
Original taxonomic scope of pathogens	All parasites and pathogens (incl. viruses, bacteria, macroparasites, protozoans, prions)	All symbionts (incl. viruses, bacteria, macroparasites, protozoans, prions, green algae, molluscs, and cnidarians)	Viruses	Viruses and bacteria
Original taxonomic scope of hosts	Mammals (subset: only ungulates, carnivores, and primates)	Vertebrates and invertebrates	Mammals	Vertebrates
Diagnostic method identified (PCR, serology, etc.)?	Yes	No	Yes	Yes
URL of current version	http://onlinelibrary.wiley.com/doi/10.1002/ecy.1799/suppinfo	https://eid2.liverpool.ac.uk/	https://github.com/ecohealthalliance/HP3	https://doi.org/10.6084/m9.figshare.8262779

Box 1. Glossary.

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

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

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

Edgelist: 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 queryable 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.

Figure 1. Network representation of the CLOVER dataset. The nodes of the entire CLOVER network have been projected to a two-dimensional space using t-SNE; in each panel, only the nodes found in the dataset are shown in colour. In each dataset, a non-trivial proportion of associations are completely unique and unrecorded elsewhere, even after taxonomic reconciliation. This was the case for 203 of 1360 associations in EID2 (14.9%); 614/2783 in HP3 (22.1%); 269/893 in GMPD2 (30.1%); and 1705/4207 in Shaw (40.5%).

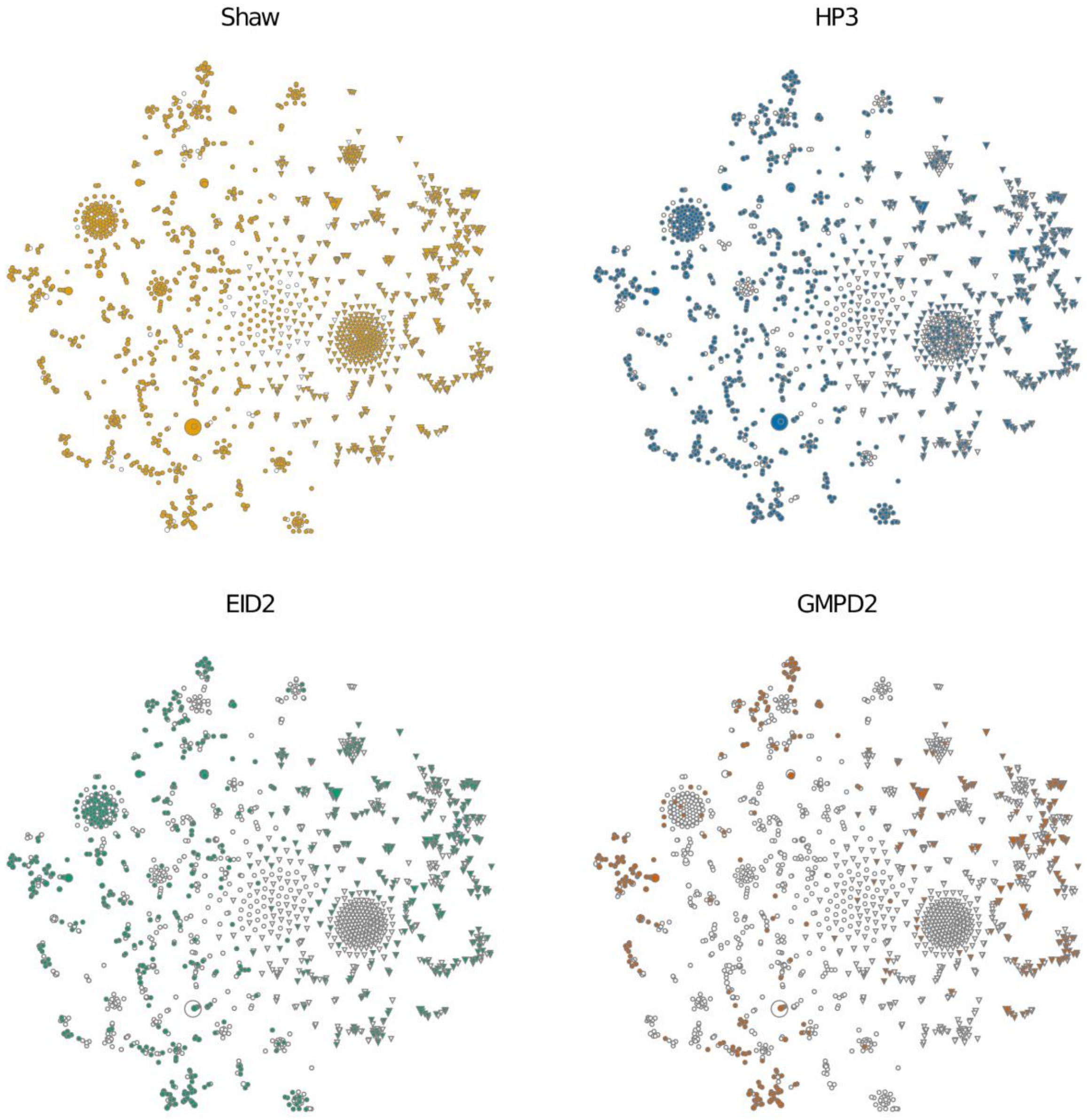


Figure 2. Proportional overlap before and after host taxonomic updating. The percentages and fill colours in these tiles can be interpreted as "% y axis was contained in x axis"; for example, 32% of uncleaned EID2 hosts were also represented in GMPD2, while 47% of cleaned Shaw associations were also contained in HP3. Darker colours represent greater overlap.

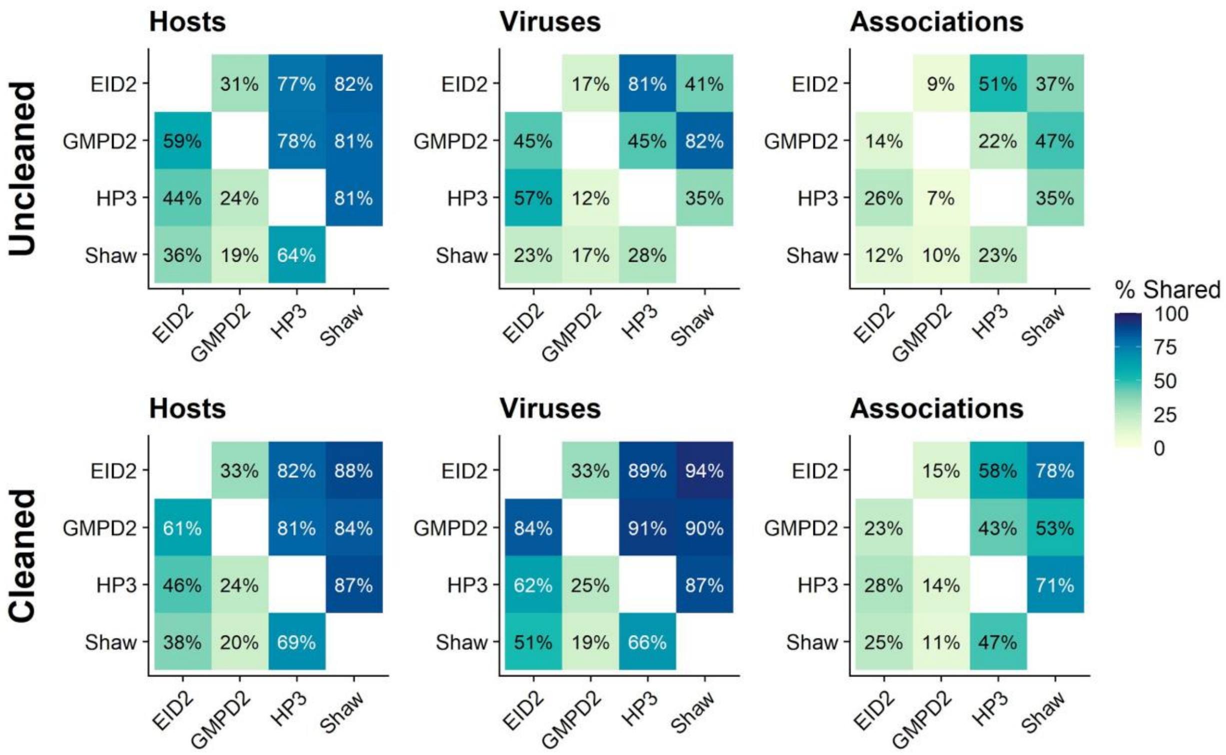
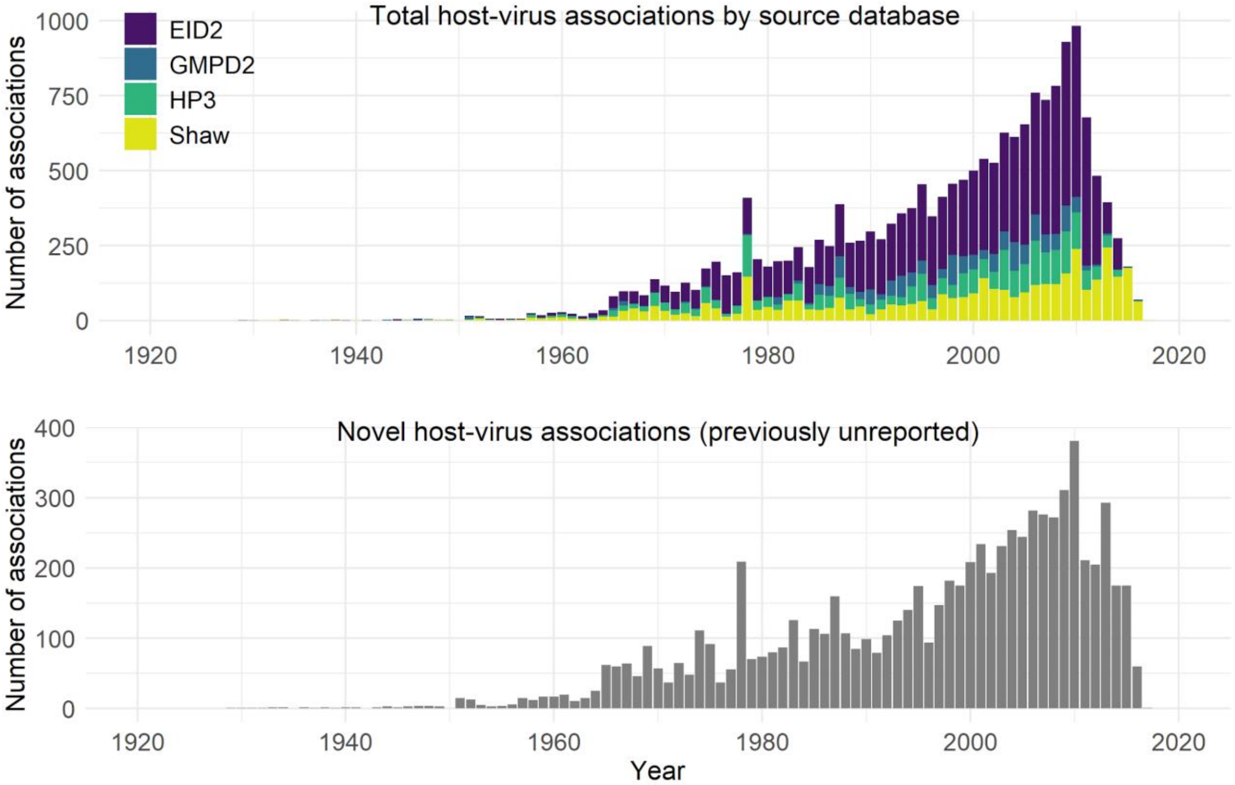


Figure 3. Temporal trends in reporting of host-virus associations in the CLOVER dataset. Bar graphs show, for each year, the total number of reported associations coloured by source database (which can include duplicates of the same association reported over multiple years; top graph) and the number of novel unique associations (i.e. previously unreported; bottom graph). 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).


Total host-virus associations by source database
Number of associations
EID2
GMPD2
HP3
Shaw
Novel host-virus associations (previously unreported)
Number of associations
Year