

## 1 2 Data proliferation, reconciliation, and synthesis in viral ecology

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

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

33  
34 \*Correspondence to: Rory Gibb ([rory.j.gibb@gmail.com](mailto:rory.j.gibb@gmail.com)) and Colin J. Carlson  
35 ([colin.carlson@georgetown.edu](mailto:colin.carlson@georgetown.edu))

36  
37 **Acknowledgements:** This work was supported by funding to the Viral Emergence Research  
38 Initiative (VERENA) consortium, including NSF BII 2021909 and a grant from Institut de  
39 Valorisation des Données (IVADO). The authors thank Noam Ross, Maya Wardeh, and many  
40 others for formative conversations about these datasets and for their tireless work making those  
41 data available to the research community.

42  
43 **Author biography:** All the authors are members of the Viral Emergence Research Initiative  
44 (VERENA) consortium, a global scientific collaboration to predict which viruses could infect  
45 humans, which animals host them, and where they could emerge.

46 **Abstract**

47

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

64

65

66 **Introduction**

67

68 The emergence of SARS-CoV-2 was a harsh reminder that uncharacterized wildlife  
69 viruses can suddenly become globally relevant. Efforts to identify wildlife viruses with the  
70 potential to infect humans, and to predict spillover and emergence trajectories, are  
71 becoming more popular than ever (including with major scientific funders). However, the  
72 value of these efforts is limited by an incomplete understanding of the global virome (Wille  
73 et al. 2020). Significant knowledge gaps exist regarding the mechanisms of viral  
74 transmission and replication, host-pathogen associations and interactions, spillover  
75 pathways, and several other dimensions of viral emergence. Further, although billions of  
76 dollars have been invested in these scientific challenges over the last decade alone, much  
77 of the data relevant to these problems remains unsynthesized. Fragmented data access  
78 and a lack of standardization preclude an easy reconciliation process across data  
79 sources, making the whole less than the sum of its parts, and hindering synthetic research  
80 (Wyborn et al. 2018).

81

82 Here, we propose that data synthesis is a seminal challenge for translational work in viral  
83 ecology. This requires researchers to go beyond the usual steps of data collection and  
84 publication, to develop a community of practice that prioritizes data synthesis and  
85 reconciles semi-reproduced work across different teams and disciplines. As an illustrative  
86 example, we describe the analytical hurdles of working with **host-virus association data**,  
87 a format that characterizes the global virome as a bipartite network of hosts and viruses,  
88 with pairs connected by observed potential for infection. Recent studies highlight the  
89 central role for these data in efforts to understand viral macroecology and evolution  
90 (Carlson et al. 2019, Dallas et al. 2019, Albery et al. 2020), to predict zoonotic emergence  
91 risk (Han et al. 2015, 2016, Olival et al. 2017, Wardeh et al. 2020), and to anticipate the  
92 impacts of global environmental change on infectious disease (Carlson et al. 2020, Gibb  
93 et al. 2020, Johnson et al. 2020). Several bespoke datasets have been compiled to  
94 address these questions, and as interest in these topics has grown, so has the  
95 fragmentation of total knowledge across those datasets. To illustrate this problem (and a  
96 simple solution), we compare and reconcile four major host-virus association datasets,

97 each of which is different enough that we anticipate the results of individual studies could  
98 be strongly shaped by choice of dataset.

99

100 **Four parts of one whole**

101

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

108

109 **GMPD 2.0:** The Global Mammal Parasite Database (Nunn and Altizer 2005), started in  
110 1999 and now in its second public version (Stephens et al. 2017), emerged from  
111 continued efforts to compile mammal-parasite association data from published literature  
112 sources. Construction of the GMPD used a variety of similar strategies that combined  
113 host Latin names with a string of parasite-related terms to search online literature  
114 databases. Pertinent literature was then manually identified and relevant association and  
115 metadata compiled. The initial database was focused on primate hosts (Nunn and Altizer  
116 2005), and expanded to include separate sections for ungulates (Ezenwa et al. 2006) and  
117 carnivores (Lindenfors et al. 2007). In 2017, GMPD 2.0 was released, which merged  
118 these three previously independent databases that were being independently maintained  
119 and updated (Stephens et al. 2017). The updated dataset encompasses 190 primate, 116  
120 ungulate, and 158 carnivore species, and record their interactions with 2,412 unique  
121 “parasite” species, including 189 viruses, as well as bacteria, protozoa, helminths,  
122 arthropods, and fungi. Notable improvements in version 2 of the GMPD are the  
123 construction of a unified parasite taxonomy that bridges occurrence records across host  
124 taxa, the expansion of host-parasite association data along with georeferencing, and  
125 enhanced parasite trait data (e.g., transmission mode). The original data are available as  
126 a web resource ([www.mammalparasites.org](http://www.mammalparasites.org)), and the data from GMPD version 2 can  
127 also be downloaded as static files from a data paper (Stephens et al. 2017). In addition,

128 one subsection of the GMPD, named the “Global Primate Parasite Database,” has been  
129 independently maintained and regularly updated by Charles Nunn (data available at  
130 <https://parasites.nunn-lab.org/>). Consequently, the primate subsection of GMPD 2.0  
131 includes papers published up to 2015, while the ungulate and carnivore subsections stop  
132 after 2010 (Stephens et al. 2017).

133

134 *EID2*: The ENHanCEd Infectious Diseases Database (EID2), curated by the University of  
135 Liverpool, may be the largest dynamic dataset of any symbiotic interactions (Wardeh et  
136 al. 2015). EID2 is compiled from automated, dynamic scrapes of two web sources:  
137 publication titles and abstracts indexed in the PubMed database and the NCBI Nucleotide  
138 Sequence database (along with its associated taxonomic metadata). The EID2 data is  
139 structured using the concepts of “carrier” and “cargo” rather than host and pathogen, as  
140 it includes a number of ecological interactions beyond the scope of normal host-pathogen  
141 interactions, including potentially unresolved mutualist or commensal associations.  
142 Interactions are stored as a geographic edgelist, where each carrier and cargo can also  
143 have locality information; additional metadata include the number of sequences in  
144 GenBank and related publications. EID2’s dynamic web interface (currently available  
145 through download on a limited query-by-query basis which researchers often manually  
146 bind or by personal correspondence with data curators) contains information  
147 encompassing 4,799 mammal “carrier” species and 70,614 microparasite or  
148 macroparasite “cargo” species, of which 9,605 are viruses (Wardeh et al. 2020). However,  
149 many researchers continue to use the static, open release of EID2 from a 2015 data paper  
150 (Wardeh et al. 2015), which we focus on here for comparative purposes as a stable  
151 version of the database available to the community of practice. The EID2 data were  
152 originally validated for completeness against GMPD 1.0.

153

154 *HP3*: The Host-Parasite Phylogeny Project dataset (HP3) was developed by EcoHealth  
155 Alliance over the better part of a decade. Published along with a landmark analysis of  
156 zoonotic spillover (Olival et al. 2017), the HP3 dataset consists of 2,805 associations  
157 between 754 mammal hosts and 586 virus species. These were compiled from literature  
158 published between 1940 and 2015, based on targeted searches of online reference

159 databases. Complementary with the search strategy used for the GMPD, rather than  
160 starting with a list of host names, HP3 started with names of known mammal viruses listed  
161 in the International Committee on Taxonomy of Viruses (ICTV) database. These virus  
162 names along with their synonyms were then used as search terms to identify literature  
163 containing host-virus association data. To narrow search results for well-studied viruses,  
164 they included additional host range-related terms to identify relevant publications. Data  
165 collection and cleaning for HP3 began in 2010 and the database has been static since  
166 2017; it can be obtained as a flat file in the published study's data repository (Olival et al.  
167 2017). HP3 includes a host-virus edgelist (see Glossary), separate files for host and virus  
168 taxonomy, and separate files for host and virus traits. Host-virus association records are  
169 provided with a note about method of identification (PCR, serology including specific  
170 methods, etc.), which may be useful for researchers interested in the different levels of  
171 confidence ascribed to particular associations (Becker et al. 2020). HP3's internal  
172 taxonomy is also harmonized with two mammal trees (Bininda-Emonds et al. 2007, Fritz  
173 et al. 2009), facilitating analyses that seek to account for host phylogenetic structure while  
174 testing hypotheses about viral ecology and evolution (e.g. Becker et al., Farrell et al.,  
175 Olival et al. 2017, Washburne et al. 2018, Guth et al. 2019, Park 2019, Albery et al. 2020,  
176 Mollentze and Streicker 2020). HP3 was also validated against GMPD 1.0.

177  
178 *Shaw*: Recent work by Shaw *et al.* built a host-pathogen edgelist by combining a  
179 systematic literature search with cross-validation from several of the above-mentioned  
180 datasets (Shaw et al. 2020). Similar to the construction of HP3, the authors started with  
181 lists of known pathogenic bacteria and viruses found in humans and animals. They then  
182 conducted Google Scholar searches pairing pathogen names with disease-related  
183 keywords, followed by manual review of search results. For well-studied pathogens they  
184 limited their manual review to a subset of the top 200 most “relevant” publications as  
185 determined by Google. From the resulting literature searches, the authors compiled  
186 12,212 interactions between 2,656 vertebrate host species (including, but not limited to,  
187 mammals) and 2,595 viruses and bacteria. GMPD2, EID2, and the Global Infectious  
188 Diseases and Epidemiology Network (GIDEON) Guide to Medically Important Bacteria  
189 (Gideon Informatics, Inc. and Berger 2020) were used to validate the host-pathogen

190 associations. The dataset is available as a static flat file through figshare and the project  
191 GitHub repository (Shaw et al. 2020). Host-pathogen associations are provided alongside  
192 pathogen metadata (e.g., genome size, bacterial traits, transmission mode, zoonotic  
193 status) and diagnostic method (i.e., PCR, pathogen isolation, pathology). The dataset  
194 also includes a comprehensive host phylogeny, developed specifically for the study using  
195 nine mitochondrial genes for downstream analyses of host phylogenetic similarity and  
196 host breadth.

197

## 198 **A reconciled mammal virome dataset**

199

200 Though some of these datasets were validated against each other during production, they  
201 are sometimes used for cross-validation in analytical work (Albery et al. 2020), and some  
202 studies have generated a study-specific *ad hoc* reconciled dataset (Farrell et al. 2020,  
203 Gibb et al. 2020), no work has been published with the primary aim of reconciling them  
204 as correctly, comprehensively, and reproducibly as possible. Dynamic datasets like EID2,  
205 and recent datasets like Shaw, can inherently draw on a greater cumulative body of  
206 scientific work. This could mean they include most of the data captured by previous  
207 efforts, yet we found there are substantial differences among all four datasets. In isolation,  
208 we expect that these differences could impact ecological and evolutionary inference in  
209 ways that are difficult to quantify, with special relevance to significance thresholds in  
210 hypothesis-testing research (i.e., different datasets may confer different power to  
211 statistical tests). In unison, we expect that these data could be standardized into one  
212 shared format, allowing them to cover a greater percentage of the global virome, a greater  
213 diversity of host species, and obviating the need for researchers to either choose between  
214 them or implement *ad hoc* solutions that merge them prior to analysis.

215

216 To illustrate the potential for comprehensive data reconciliation, we harmonized the four  
217 major datasets described here, creating a new synthetic 'CLOVER' dataset out of the four  
218 "leaves" (which we have made available with this study). To do so, we first harmonized  
219 the host taxonomy of all four datasets using the R package 'taxize' (Chamberlain and  
220 Szöcs 2013), then manually resolved remaining discrepancies. Finally, using the Julia

221 package 'NCBITaxonomy.jl' (Poisot 2020), we standardized host and virus taxonomy  
222 against the taxonomic hierarchy (Schoch et al. 2020) used as a reference by the National  
223 Center for Biotechnology Information's Taxonomy database (ncbi.nlm.nih.gov). With all  
224 four datasets taxonomically consistent, we were able to show that each only covered a  
225 portion of the known global mammal virome, even for the most studied hosts and viruses  
226 (Figure 1). Our taxonomic harmonization helped reconcile some discrepancies,  
227 increasing overlap among the datasets (Figure 2), but notable differences remained. This  
228 could confound inference: for example, using a simple linear model, we found that **data**  
229 **provenance** (see Glossary) explained 8.8% of variation in host species' viral diversity  
230 (but only 4.7% after harmonization). When studies report different findings based on slight  
231 variation around a significance threshold, readers should therefore wonder whether subtle  
232 differences in the underlying datasets might account for such variation.

233

234 Integrated datasets move us a step closer to resolving this uncertainty. The CLOVER  
235 dataset covers 1,081 mammal host species and 829 associated viruses. This only  
236 represents 16.9% of extant mammals (Burgin et al. 2018) and at most 2.1% of their  
237 viruses (Carlson et al. 2019) - perhaps a marginal improvement over the 954 mammal  
238 hosts (14.9%) and 733 viruses (1.8%) in the reconciled Shaw sub-dataset, but an  
239 improvement nonetheless. The biggest functional gain is not in the *breadth* of the  
240 reconciled data, but in its *depth*: the Shaw database records 4,209 interactions among  
241 these host and virus species, while CLOVER captures 5,494. Given that previous studies  
242 have estimated that 20-40% of host-parasite links are unknown (in GMPD2 (Dallas et al.  
243 2017)), this 30% improvement is notable and shows the value of data synthesis: both  
244 building out *and* filling in synthetic datasets will significantly improve the performance of  
245 statistical models, which are usually heavily confounded by matrix sparsity (Becker et al.,  
246 Dallas et al. 2017).

247

248 In addition, harmonization of metadata on virus detection methods across datasets  
249 enables a greater scrutiny of the strength of evidence in support of each host-virus  
250 association. We applied a simplified detection method classification scheme (either  
251 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).

269  
270 Data synthesis inherently relies on a scientific community that generates new, often  
271 conflicting, data. The generation of truly novel data or finding ways to resolve existing  
272 observations that are in conflict are two equally viable paths to scientific progress.  
273 However, in the current funding landscape, researchers may have a significant incentive  
274 to position themselves as creating an entirely “novel” dataset from scratch, even if it  
275 partially replicates available data sources, or to focus their limited resources on datasets  
276 that improve the depth of knowledge within a narrow scope (e.g., a focus on specific  
277 taxonomic groups). But when testing microbiological or eco-evolutionary hypotheses,  
278 rather than simply using each newly-published dataset as a benchmark for which one is  
279 “most up-to-date,” we suggest a necessary shift in scientific cultural norms towards using  
280 synthetic, reconciled data like CLOVER as an analytical best practice. To make this  
281 possible, at least a handful of researchers will need to continue the task of stepwise  
282 integration, using datasets that synthesize existing knowledge across teams, institutions,

283 and funding programs to fill in critical data with even more detail. The required tasks (e.g.,  
284 identifying relevant source data, cleaning taxonomic information, harmonizing metadata  
285 on diagnostic information or spatiotemporal structure) can be time-consuming but are  
286 relatively straightforward to conduct, and can increasingly be automated thanks to the  
287 rapid growth of new data and tools for reproducible research (Boettiger et al. 2015,  
288 Lowndes et al. 2017, Colella et al. 2020). There is a clear need, and no obvious technical  
289 barrier, to invest more effort in data harmonization: engaging in this process as a form of  
290 open science will accelerate progress for the entire research community.

291

## 292 **Relevance to future efforts**

293

294 Here, we showed that a simple data synthesis effort can create a dramatically more  
295 comprehensive dataset of mammal-virus associations. However, this is a temporary  
296 solution and one that will become less sustainable if similar datasets continue to  
297 proliferate or if newer iterations of existing datasets are released, each absorbing different  
298 parts of existing efforts. Over the longer term, given global investments in viral discovery  
299 from wildlife, static datasets will quickly become out-of-date, and their relation to the most  
300 recent empirical knowledge will be left unclear. For example, the CLOVER dataset  
301 becomes significantly sparser after 2010, both in terms of the overall number of reported  
302 host-virus associations, and the reporting of novel (i.e. previously undetected)  
303 associations (Figure 3). This sparseness is most likely due to time lags between host-  
304 virus sampling in the field, the reporting or publication of associations, and their eventual  
305 inclusion in one of the component datasets, and suggests that CLOVER may now be  
306 missing up to a decade's worth of known host-virus data. In the near term, microbiologists  
307 and data scientists may want to approach the task of data reconciliation with a much  
308 broader scope, and develop a more sustainable data platform.

309

310 Scaling up the aggregation of host-virus association data will not be easy, but is not an  
311 insurmountable endeavour. We suggest working backwards from the intended end  
312 product: the goals outlined here are best served by a central system (with an online  
313 access point to the consumable data), spanning the information available from multiple

314 data sources (which demands backend engines drawing from existing databases, while  
315 tracking data provenance and ensuring proper attribution). Further, the most valuable  
316 data resource would be easily updatable by practitioners (which demands a portal for  
317 manual user input or an Integrated Publishing Toolkit to work from flat files). For users,  
318 these data should be accessible in a programmatic way (i.e., through a web API allowing  
319 for bulk download and/or other interfaces like an R package), help analysts build  
320 reproducibility (through versioning of the entire database, or of a specific user query), and  
321 offer predictable formats (through a data specification standard devised by a  
322 multidisciplinary group).

323

324 Fortunately, the field of ecoinformatics has the capacity to help inform this design and  
325 development process. Massive bioinformatic data portals like the Global Biodiversity  
326 Informatics Facility (gbif.org), the Encyclopedia of Life (eol.org), and the Ocean  
327 Biodiversity Information System (obis.org) all offer most of the functionalities we outline  
328 here, though they are aimed at slightly different forms of biodiversity data. More recent  
329 contributions dedicated to ecological network data include Global Biotic Interactions  
330 (Poelen et al. 2014) (GLOBI, which consumes flat files and formats them), helminthR  
331 (Dallas 2016), and mangal (Poisot et al. 2016) (which stores a metadata-rich  
332 representation of species interaction networks), all of which reconcile their taxonomy with  
333 other databases through the use of unique taxon keys. In short, researchers interested in  
334 the global virome need not divert their attention, resources, and effort away from the  
335 pressing tasks related to monitoring viral pathogens, but they can leverage existing  
336 products, expertise, and capacity in neighbouring fields to bolster their ability to do so.  
337 Given the eagerness ecologists have shown to participate in SARS-CoV-2 research, we  
338 anticipate that our field may be especially well-poised to jump into this task post-  
339 pandemic. We aim, in our current efforts, to lay that groundwork.

340

341 An integrated platform for the deposition, curation, archival, and sharing of host-virus  
342 associations in a *prêt-à-manger*, metadata-rich format has inherent value for the entire  
343 scientific community. When the format of a dataset is well established, it allows for the  
344 development of tools that mine the data in real-time. For example, the field of biodiversity

345 studies has adopted the concept of Essential Biodiversity Variables, which can be  
346 updated when the underlying data change (Pereira et al. 2013, Fernández et al. 2019,  
347 Jetz et al. 2019). Having the ability to revisit predictions about the host-virus network could  
348 improve models that assess zoonotic potential of wildlife viruses (Farrell et al. 2020,  
349 Mollentze et al. 2020), generate priority targets for wildlife reservoir sampling (Becker et  
350 al., Babayan et al. 2018, Plowright et al. 2019), and help benchmark model performance  
351 related to these tasks. Beyond training and validation, link prediction models built on these  
352 reconciled databases may be used to target future literature searches, shifting from  
353 systematic literature searches to a model based approach to database updating.  
354 Increased collaboration between data collectors, data managers, and data scientists that  
355 leads to better data standardization and reconciliation is the only way to productively  
356 synthesize our knowledge of the global virome.

357

### 358 **Data and code availability**

359

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

364

365 **References.**

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

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

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

373 Becker DJ, Seifert SN, Carlson CJ. 2020. Beyond Infection: Integrating Competence into  
374 Reservoir Host Prediction. *Trends in Ecology & Evolution* 35: 1062–1065.

375 Bininda-Emonds ORP, Cardillo M, Jones KE, MacPhee RDE, Beck RMD, Grenyer R, Price SA,  
376 Vos RA, Gittleman JL, Purvis A. 2007. The delayed rise of present-day mammals. *Nature*  
377 446: 507–512.

378 Boettiger C, Chamberlain S, Hart E, Ram K. 2015. Building Software, Building Community:  
379 Lessons from the rOpenSci Project. *Journal of Open Research Software* 3.

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

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

384 Carlson CJ, Zipfel CM, Garnier R, Bansal S. 2019. Global estimates of mammalian viral  
385 diversity accounting for host sharing. *Nature ecology & evolution* 3: 1070–1075.

386 Chamberlain SA, Szöcs E. 2013. taxize: taxonomic search and retrieval in R. *F1000Research* 2:  
387 191.

388 Colella JP, Stephens RB, Campbell ML, Kohli BA, Parsons DJ, Mclean BS. 2020. The Open-  
389 Specimen Movement. *BioScience*.

390 Dallas T. 2016. helminthR: an R interface to the London Natural History Museum's Host-  
391 Parasite Database. *Ecography* 39: 391–393.

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

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

396 Ezenwa VO, Price SA, Altizer S, Vitone ND, Cook KC. 2006. Host traits and parasite species  
397 richness in even and odd-toed hoofed mammals, Artiodactyla and Perissodactyla. *Oikos*  
398 115: 526–536.

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

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

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

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

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

409 Gilbert AT, Fooks AR, Hayman DTS, Horton DL, Müller T, Plowright R, Peel AJ, Bowen R,  
410 Wood JLN, Mills J, Cunningham AA, Rupprecht CE. 2013. Deciphering serology to  
411 understand the ecology of infectious diseases in wildlife. *EcoHealth* 10: 298–313.

412 Guth S, Visher E, Boots M, Brook CE. 2019. Host phylogenetic distance drives trends in virus  
413 virulence and transmissibility across the animal-human interface. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 374: 20190296.

415 Han BA, Kramer AM, Drake JM. 2016. Global Patterns of Zoonotic Disease in Mammals.  
416 Trends in parasitology 32: 565–577.

417 Han BA, Schmidt JP, Bowden SE, Drake JM. 2015. Rodent reservoirs of future zoonotic  
418 diseases. *Proceedings of the National Academy of Sciences of the United States of  
419 America* 112: 7039–7044.

420 Jetz W, McGeoch MA, Guralnick R, Ferrier S, Beck J, Costello MJ, Fernandez M, Geller GN,  
421 Keil P, Merow C, Meyer C, Muller-Karger FE, Pereira HM, Regan EC, Schmeller DS, Turak  
422 E. 2019. Essential biodiversity variables for mapping and monitoring species populations.  
423 *Nature ecology & evolution* 3: 539–551.

424 Johnson CK, Hitchens PL, Pandit PS, Rushmore J, Evans TS, Young CCW, Doyle MM. 2020.  
425 Global shifts in mammalian population trends reveal key predictors of virus spillover risk.  
426 *Proceedings. Biological sciences / The Royal Society* 287: 20192736.

427 Lachish S, Murray KA. 2018. The Certainty of Uncertainty: Potential Sources of Bias and  
428 Imprecision in Disease Ecology Studies. *Frontiers in veterinary science* 5: 90.

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

432 Lowndes JSS, Best BD, Scarborough C, Afflerbach JC, Frazier MR, O'Hara CC, Jiang N,  
433 Halpern BS. 2017. Our path to better science in less time using open data science tools.  
434 *Nature ecology & evolution* 1: 160.

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

438 Mollentze N, Streicker DG. 2020. Viral zoonotic risk is homogenous among taxonomic orders of  
439 mammalian and avian reservoir hosts. *Proceedings of the National Academy of Sciences of*

440 the United States of America 117: 9423–9430.

441 Nunn CL, Altizer SM. 2005. The global mammal parasite database: An online resource for  
442 infectious disease records in wild primates. *Evolutionary Anthropology: Issues, News, and*  
443 *Reviews* 14: 1–2.

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

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

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

451 Pereira HM, Ferrier S, Walters M, Geller GN, Jongman RHG, Scholes RJ, Bruford MW,  
452 Brummitt N, Butchart SHM, Cardoso AC, Coops NC, Dulloo E, Faith DP, Freyhof J,  
453 Gregory RD, Heip C, Höft R, Hurt G, Jetz W, Karp DS, McGeoch MA, Obura D, Onoda Y,  
454 Pettorelli N, Reyers B, Sayre R, Scharlemann JPW, Stuart SN, Turak E, Walpole M,  
455 Wegmann M. 2013. Ecology. Essential biodiversity variables. *Science* 339: 277–278.

456 Plowright RK, Becker DJ, Crowley DE, Washburne AD, Huang T, Nameer PO, Gurley ES, Han  
457 BA. 2019. Prioritizing surveillance of Nipah virus in India. *PLoS neglected tropical diseases*  
458 13: e0007393.

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

461 Poisot T, Baiser B, Dunne JA, Kéfi S, Massol F, Mouquet N, Romanuk TN, Stouffer DB, Wood  
462 SA, Gravel D. 2016. mangal - making ecological network analysis simple. *Ecography* 39:  
463 384–390.

464 Poisot T. 2020. NCBI taxonomy.jl: Interact with the NCBI Taxonomy backbone from Julia.  
465 <https://doi.org/10.5281/zenodo.4282820>

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

470 Shaw LP, Wang AD, Dylus D, Meier M, Pogacnik G, Dessimoz C, Balloux F. 2020. The  
471 phylogenetic range of bacterial and viral pathogens of vertebrates. *Molecular ecology* 29:  
472 3361–3379.

473 Shaw LP, Wang AD, Dylus D, Meier M, Pogacnik G, Dessimoz C, Balloux F. 2020. Data from:  
474 The phylogenetic range of bacterial and viral pathogens of vertebrates.  
475 [https://figshare.com/articles/dataset/The\\_phylogenetic\\_range\\_of\\_bacterial\\_and\\_viral\\_patho](https://figshare.com/articles/dataset/The_phylogenetic_range_of_bacterial_and_viral_patho)  
476 [gens\\_of\\_vertebrates\\_dataset\\_and\\_supplementary\\_material/8262779](https://figshare.com/articles/dataset/The_phylogenetic_range_of_bacterial_and_viral_patho)

477 Stephens PR, Pappalardo P, Huang S, Byers JE, Farrell MJ, Gehman A, Ghai RR, Haas SE,  
478 Han B, Park AW, Schmidt JP, Altizer S, Ezenwa VO, Nunn CL. 2017. Global Mammal

479        Parasite Database version 2.0. *Ecology* 98: 1476.

480        Wardeh M, Risley C, McIntyre MK, Setzkorn C, Baylis M. 2015. Database of host-pathogen and  
481        related species interactions, and their global distribution. *Scientific data* 2: 150049.

482        Wardeh M, Sharkey KJ, Baylis M. 2020. Integration of shared-pathogen networks and machine  
483        learning reveals the key aspects of zoonoses and predicts mammalian reservoirs.  
484        *Proceedings. Biological sciences / The Royal Society* 287: 20192882.

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

487        Wille M, Geoghegan JL, Holmes EC. 2020. How accurately can we assess zoonotic risk?  
488        bioRxiv preprint <https://doi.org/10.1101/2020.08.17.254961>

489        Wyborn C, Louder E, Harrison J, Montambault J, Montana J, Ryan M, Bednarek A, Nesshöver  
490        C, Pullin A, Reed M, Dellecker E, Kramer J, Boyd J, Dellecker A, Hutton J. 2018.  
491        Understanding the Impacts of Research Synthesis. *Environmental Science & Policy* 86: 72–  
492        84.

493

494

495

496

497

498 **Table 1.** Available “big data” on host-virus associations, and major features of each  
499 dataset. Numbers of unique association records and host, virus, and pathogen species  
500 are all derived from the reconciled version presented in the CLOVER database, and  
501 therefore these numbers may differ from those presented in the main text (which are  
502 taken from the source data, or from self-reporting by the data curators). \*Number of  
503 associations and taxa accurate as of 2015 static release in *Scientific Data* paper.

504

505

Dataset	GMPD2	EID2*	HP3	Shaw
Source	U. Georgia	U. Liverpool	EcoHealth Alliance	Shaw LP, <i>et al.</i> <i>Molecular Ecology</i> (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	<a href="http://onlinelibrary.wiley.com/doi/10.1002/ecy.1799/supplinfo">http://onlinelibrary.wiley.com/doi/10.1002/ecy.1799/supplinfo</a>	<a href="https://eid2.liverpool.ac.uk/">https://eid2.liverpool.ac.uk/</a>	<a href="https://github.com/ecohealth-alliance/HP3">https://github.com/ecohealth-alliance/HP3</a>	<a href="https://doi.org/10.6084/m9.figshare.8262779">https://doi.org/10.6084/m9.figshare.8262779</a>

506

507

508 **Box 1. Glossary.**

509  
510 **Association data:** a format that records ecological interactions between a host and  
511 symbiont (an *association*) in the form of an edgelist.

512  
513 **Data provenance:** The primary literature origin of a particular record or set of records in  
514 a synthetic dataset.

515  
516 **Data reconciliation:** the task of harmonizing the language of a given dataset's fields and  
517 metadata to allow a researcher to merge data of different provenance, and generate a  
518 new synthetic product.

519  
520 **Edgelist:** a table, spreadsheet, or matrix of “links” in a host-symbiont network, where  
521 each row records the known association of a different host-symbiont pair.

522  
523 **Flat file:** a static document in Excel or similar spreadsheet or data format, with no  
524 dynamic component (no updating) and all data available from a single file rather than a  
525 queryable interface.

526  
527 **Metadata:** additional data describing focal data of interest and that is relevant to  
528 interpretation and analysis. Important examples for host-virus associations include  
529 sampling method (for example, serological assay, PCR or pathology), date and  
530 geographical location of sampling, and standardized information on host and virus  
531 taxonomy.

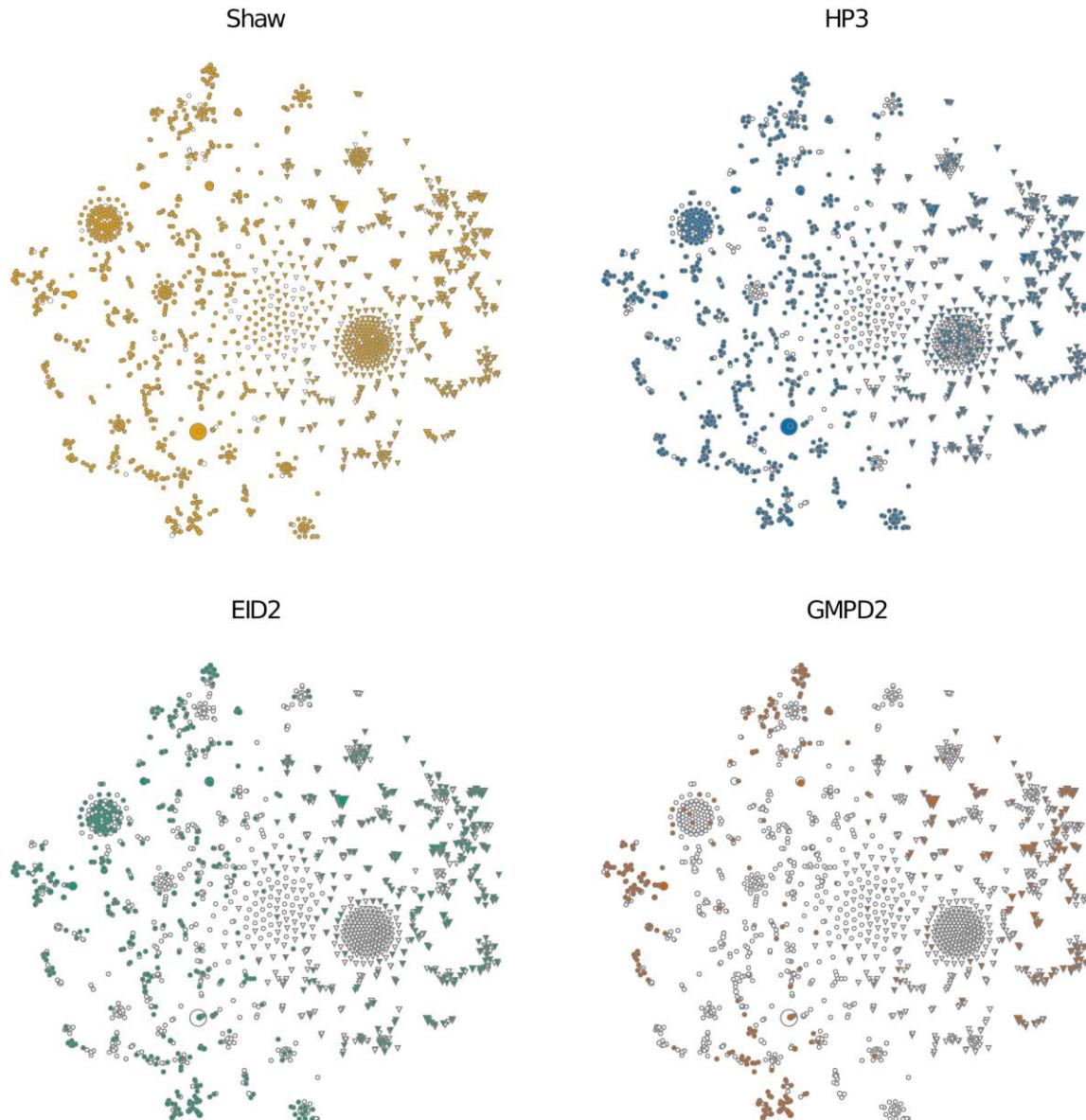
532  
533 **Open data:** data that is directly and freely accessible for reuse and exploration without  
534 impediment, gatekeeping, or cost restriction.

535  
536  
537  
538  
539  
540  
541  
542

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

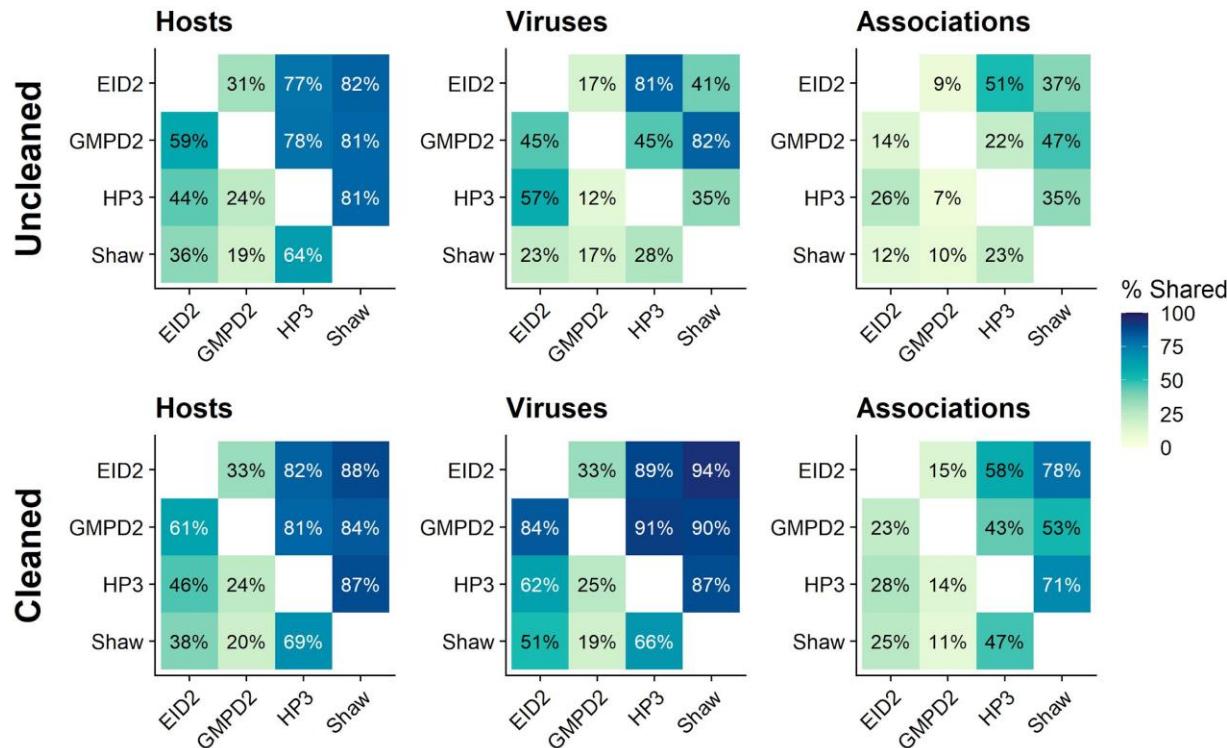
550

551



552

553 **Figure 2.** Proportional overlap before and after host taxonomic updating. The  
554 percentages and fill colours in these tiles can be interpreted as “% y axis was contained  
555 in x axis”; for example, 32% of uncleaned EID2 hosts were also represented in GMPD2,  
556 while 47% of cleaned Shaw associations were also contained in HP3. Darker colours  
557 represent greater overlap.  
558  
559



560  
561  
562  
563  
564  
565

566 **Figure 3.** Temporal trends in reporting of host-virus associations in the CLOVER  
567 dataset. Bar graphs show, for each year, the total number of reported associations  
568 coloured by source database (which can include duplicates of the same association  
569 reported over multiple years; top graph) and the number of novel unique associations  
570 (i.e. previously unreported; bottom graph). Years reflect the date when an association  
571 was reported, either in a published paper or report (for literature-based records) or to  
572 the NCBI Nucleotide database (EID2 only).

573

