

## 1 Abstract

2 Curated databases of species interactions are instrumental to exploring and un-  
3 derstanding the spatial distribution of species and their biotic interactions. In the  
4 process of conducting such projects, data development and curation efforts may  
5 give rise to a data product with utility beyond the scope of the original work,  
6 but which becomes inaccessible over time. Data describing insect host-pathogen  
7 interactions are fairly rare, and should thus be preserved and curated with appro-  
8 priate metadata. Here, we introduce the `insectDisease R` package, a mechanism  
9 for curating, updating, and distributing data from the *Ecological Database of the*  
10 *World's Insect Pathogens*, a database of insect host-pathogen associations, includ-  
11 ing attempted inoculations and infection outcomes for insect hosts and pathogens  
12 (bacteria, fungi, nematodes, protozoans, and viruses). This dataset has been uti-  
13 lized for several projects since its inception, but without a well-defined, curated  
14 and permanent repository, its existence and access have been limited to word-of-  
15 mouth connections. The current effort presented here aims to provide a means to  
16 preserve, augment, and disseminate the database in a documented and versioned  
17 format. This project is an example of the type of effort that will be necessary to  
18 maintain valuable databases after the original funding disappears.

19 **Running title:** Ecological Database of the World's Insect Pathogens (EDWIP)

20

## 21 Introduction

22 There are a number of data sources documenting host-pathogen associations, es-  
23 pecially for pathogens of mammals (Gibb et al., 2021, Patrick et al., 2017), birds  
24 (Bensch et al., 2009), and fish (Strona and Lafferty, 2012). Recent work from  
25 the Verena Consortium has developed a dynamically updated host-virus associa-  
26 tion database for all vertebrate hosts (VIRION) (Carlson et al., 2021), representing  
27 the largest collection of host-virus association data to date. These resources have  
28 been fundamental to our understanding of what determines pathogen host range,  
29 pathogen species richness across a set of hosts, and overall host-pathogen network  
30 structure (e.g, Carlson et al. (2020), Dallas et al. (2018)). But while some host  
31 groups are well-studied, there are taxonomic gaps in our understanding of host-  
32 pathogen associations. Insect host-pathogen relationships have considerably less  
33 open-source data available, despite their inherent importance to scientific studies  
34 and assessments of impacts to agricultural crops and spread of vector-borne dis-  
35 ease, in addition to the sheer numerical dominance of insect species over other  
36 taxa (Stork et al., 2015). This is a clear knowledge gap.

37 Many of the existing species interaction databases have dedicated researchers,  
38 resources, and infrastructure to enable data deposition and curation in openly  
39 accessible formats. However, some data have not been as lucky, at no fault of the  
40 original data curators. These data run the risk of disappearing into a file drawer or  
41 on an external hard drive, potentially shared with a small number of researchers

42 but not accessible to the scientific community at large. One data resource arguably  
43 close to this point of disappearance is the *Ecological Database of the World's Insect*  
44 *Pathogens* (EDWIP) (Onstad, 1997).

45 The EDWIP data consist of experimental infections and field observations of  
46 the interactions between insect hosts and a number of bacterial, fungal, nema-  
47 tode, protozoan, and viral pathogens (Braxton et al., 2003). One particularly  
48 unique component of EDWIP is the existence of negative associations – attempts  
49 to inoculate a host with a given pathogen that failed to infect – for some host  
50 groups (Figure 1). Failed infections represent *true* absences or incompatibilities  
51 between a given host and pathogen. These data are incredibly useful to pathogen  
52 host range estimation and host-pathogen interaction modeling, but we rarely have  
53 data on these known non-interactions.

54 Initially created in 1992, the data have been updated prior to 2000, but no clear  
55 semantic versioning was used. As such, it is unclear how long or how frequently  
56 this updating and curation continued, and thus, how many different versions of  
57 the data may be in existence presently. The database we present here, as the  
58 backbone of this R package, represents the most up-to-date version that we know  
59 of, though this may differ slightly from previous descriptions of the data (Braxton  
60 et al., 2003). Generally, we have attempted to preserve all of the original data in  
61 the original format.

## **Solution statement**

To preserve these data in a format that is well-documented, openly accessible, versioned, and flexible for continued development, we created the `insectDisease` R package. In doing so, we implicitly adhere to the FAIR (Findable, Accessible, Interoperable, Resuable) guidelines for managing data (Wilkinson et al., 2016). By hosting the data openly on GitHub, and versioning releases of the data with a permanent identifier (DOI), we ensure the longevity and versioned curation of this data resource. Finally, the incorporation of taxonomic data through `taxize` (Chamberlain and Szöcs, 2013) ensures that host and pathogen taxonomic names are updated periodically to accommodate for dynamic data or changing taxonomies.

## **Data specification**

**Package structure** Data products are broken down by pathogen group; nematodes (`data(nematode)`), viruses (`data(viruses)`), and non-viral pathogens, which include protozoan, fungi, and bacteria (`data(nvpassoc)`). Data on negative associations is stored collectively instead of being delineated by pathogen group (`data(negative)`), but information on pathogen group is provided within each of these files, allowing for sorting of negative interactions based on the initial pathogen groupings (Table 1). This data structure is inherited from the original structure of the EDWIP data files, and code to process and join these different

82 data files is provided in the *R* package vignette.

83 Each of the pathogen groups differs slightly in the available ancillary data on  
84 experimental infections. For instance, nematode infections contain information on  
85 soil type and associated bacteria, virus infection data have information on viral  
86 dose, and non-viral pathogens (protozoans, fungi, and bacteria) have information  
87 on intermediate host species. We recommend the user explore these data and  
88 associated metadata from within *R*, as the metadata and data are neatly in the  
89 same place.

90 Data are also available on the insect host species themselves (e.g., `data(hosts)`).  
91 These data contain some information on the Canadian province where the host is  
92 found (`ProvinceI` column), what it eats (`Food` column), and what type of habitat  
93 it is found in (`Habitat` column). Additionally, a column on host insect pest status  
94 is present, offering the opportunity to explore study effort and pathogen specificity  
95 dependent on the pest status of the insect host.

96 **FAIR data** The FAIR principles represent guidelines for making data more per-  
97 sistent, findable, and well-documented. Structuring the data as an *R* package  
98 ensures that metadata and data are packaged together, where *R* manual files con-  
99 tain column names and data descriptions for each data product (*Findable*). All  
100 code to take data from the raw data (`data-raw` folder) to the end product `.RData`  
101 and `.csv` files is contained in the versioned *R* data package, and integration with

102 Zenodo (<https://doi.org/10.5281/zenodo.5821896>) provides a DOI for each  
103 release (*Accessible*). Metadata are available in redundant forms, both from within  
104 the R package as `man` files, and in the project README file such that installa-  
105 tion of the package (or navigation into the `man` folder) is not necessary. Apart  
106 from providing data in these multiple formats, user access is aided by structuring  
107 the data as a package in a very popular computing language among biologists  
108 (and other folks too) and providing all code for data processing and serving in an  
109 open and public-facing repository (*Interoperable*). Having all code and data in a  
110 streamlined, open, and versioned format, serving the data through an interactive  
111 web portal, and publishing this software note collectively serve to promote the use  
112 of this data resource (*Reusable*).

113 **Metadata and package documentation** Differences in features across the  
114 data on different pathogen types (e.g., `?nematodes` relative to `?viruses`) make  
115 combining these data non-straightforward, without a degree of loss of information.  
116 We provide some example code in the package `vignette` on how to go about  
117 combining or linking the data across types, with the caveats of information loss,  
118 and have standardized some key column names across the different data products.  
119 Further, we have documented each data resource using *R* package documentation,  
120 allowing the metadata of each data product to be examined directly from R using  
121 the `help()` function or the question mark notation (e.g., `?viruses`). These same  
122 metadata are also provided in the README file in the top-level of the GitHub

123 repository.

124 **Data cleaning and taxonomic resolution** We attempted to maintain as much  
125 of the original data structure from the raw data files provided by David Onstad,  
126 principal maintainer of the EDWIP data resource (Onstad, 1997). This includes  
127 files such as `new_assoc`, as this was likely a test file containing pathogen species  
128 such as “wormy thing”, and `newnema`, a dataset identical to `nematode`. We docu-  
129 ment these idiosyncrasies in the metadata for each data product, providing a clear  
130 overview of the state of each data subproduct.

131 The first, and perhaps most important, novel augmentation, is the resolution of  
132 host and pathogen taxonomic information. We achieved this by using the R pack-  
133 age `taxize`, specifically the NCBI taxonomic backbone (Chamberlain and Szöcs,  
134 2013), making the data interoperable with existing data efforts by the Verena  
135 Consortium (e.g., VIRION; Carlson et al. (2021)). Cached versions of host and  
136 pathogen taxonomic information are provided (`data(hostTaxonomy)` and  
137 `data(pathTaxonomy)`), and the *R* code to generate these taxonomic backbones  
138 and clean the data are provided in the package vignette. This taxonomic backbone  
139 serves to both standardize host and pathogen nomenclature, while also correcting  
140 any taxonomic changes that have occurred in the past couple decades. This in-  
141 cludes the consideration of microsporidian parasites as fungi, not protozoans, a  
142 change affecting a large set of records in the EDWIP data. All of the data within

143 the **data** and **csv** folders have already gone through these data cleaning steps.  
144 However, these data may be dynamic, such that some form of continuous integra-  
145 tion or updating of the host and pathogen taxonomy may be necessary. As such,  
146 we provide a vignette which transparently shows the steps to clean and augment  
147 the data resource, as well as reproduce figures from this manuscript. Finally, we  
148 opt to store processed data in the **csv** folder, which contains all data files in **.csv**  
149 format. This allows non-*R* users to access the csv-formatted data easily, and en-  
150 sures long-term stability of the data, as **csv** is a stable text file format. These data  
151 are also provided as **.rda** files in the **data** folder.

152 Maintaining the data dynamically as described above allows users to access  
153 the data programmatically or as versioned flatfiles (i.e., **.csv** files). However, for  
154 users who do not wish to download the entire data resource, and simply want to  
155 quickly query a static version of the database, there is also a standalone web user  
156 interface (<https://edwip.ecology.uga.edu/>) that allows users to easily subset  
157 and explore the data. The web interface serves arguably the most important  
158 subset of the overall data (data files **nematode**, **viruses**, **nvpassoc**, **negative**, and  
159 **hosts**). This interface allows users to quickly query based on host or parasite  
160 taxonomy as a dropdown list. This is perhaps more useful as a teaching tool or  
161 for initial exploration of the data, while the programmatic interface and dynamic  
162 data may be more useful for more rigorous analysis. This version of the EDWIP  
163 data will also only be deployed with a single static copy of the data, such that



164 users wanting to benefit from versioned and dynamic data will need to access the  
165 data through the GitHub repository. Future efforts to integrate the web interface  
166 and the existing dynamic data structure will be explored, but this is not currently  
167 integrated.

## 168 **Case study: covariance among pathogen groups in parasite** 169 **species richness**

170 Hosts that are infected by more pathogens of one type may also be more infected by  
171 pathogens of another type, mediated by host life history traits, metabolic demands,  
172 geographic distribution, and intensity of scientific study (Dallas and Becker, 2021).  
173 We explore this in the EDWIP data by measuring the number of known positive  
174 associations of each of the pathogen groups for each insect host species, visualizing  
175 the relationship between the number of pathogens per insect host as a correlation  
176 matrix (Figure 2). We find very little evidence that pathogen groups have positive  
177 covariance, which would be expected if host species traits or trait-based sampling  
178 biases drove infection process across pathogen groups in the same manner. The  
179 failure to detect strong positive relationships, and indeed some negative relation-  
180 ships appearing, could be a signal of the targeted nature of data collection, as  
181 many insect host species were selected to study due to their potential as a crop  
182 pest, and many pathogens were selected to study based on their potential use as  
183 biocontrol or perhaps for their ease of culture.

184 This potential sampling bias among insect host species would be evident if there  
185 were a positive relationship between the number of positive interactions and the  
186 number of negative interactions for a host species, as it would indicate that host  
187 species with lots of known interactions also tended to appear in many studies and  
188 have some negative interactions as well. We find evidence for a significantly neg-  
189 ative relationship based on a Spearman’s rank correlation ( $\rho = -0.1$ ,  $p < 0.0001$ ),  
190 indicating no discernible influence of this relationship. This does not imply that  
191 there is no sampling bias in the insect host species researchers opt to study, but  
192 that such bias was not so strong as to be clearly detected.

## 193 **Concluding comments**

194 While ecological data are growing in availability, size, accessibility, and stability,  
195 there are still data resources that are aging in place, and should not be allowed  
196 to fade out of existence. The EDWIP data provided to the authors were in a  
197 proprietary format (‘Claris FileMaker Pro 5’) that was already over 10 major  
198 versions behind. With limited inter-version operability (e.g., `.fmp5` files cannot  
199 be opened in more recent versions of the software, or require multiple conversion  
200 steps), these data seemed as if headed towards obsolescence. The `insectDisease`  
201 package ensures that these data will be available to the broadest set of researchers,  
202 be bound to relevant metadata, and be properly versioned. By hosting the data  
203 openly, we welcome contributions from researchers interested in augmenting the

204 data or building off the existing resource.

## 205 **Data accessibility**

206 The `insectDisease` R package is currently available on GitHub  
207 ([github.com/viralemergence/insectDisease](https://github.com/viralemergence/insectDisease)), with ‘.csv’ files in the `csv` directory  
208 for long-term data stability. GitHub releases of the data ensure versioning is  
209 maintained and all versions are accessible. At the time of this writing, the current  
210 version is 1.2.0 (available at  
211 <https://github.com/viralemergence/insectDisease/releases/tag/1.2.0>).  
212 Releases are given a DOI through integration with Zenodo (available at  
213 <https://doi.org/10.5281/zenodo.5821896>).

## References

- Bensch, S. et al. 2009. MalAvi: a public database of malaria parasites and related haemosporidians in avian hosts based on mitochondrial cytochrome b lineages. – *Molecular Ecology Resources* 9: 1353–1358.
- Braxton, S. et al. 2003. Description and analysis of two internet-based databases of insect pathogens: EDWIP and VIDIL. – *Journal of Invertebrate Pathology* 83: 185–195.
- Carlson, C. J. et al. 2020. What would it take to describe the global diversity of parasites? – *Proceedings of the Royal Society B* 287: 20201841.
- Carlson, C. J. et al. 2021. The Global Virome in One Network (VIRION): an atlas of vertebrate-virus associations. – *bioRxiv* .
- Chamberlain, S. A. and Szöcs, E. 2013. taxize: taxonomic search and retrieval in R. – *F1000Research* 2.
- Dallas, T. A. and Becker, D. J. 2021. Taxonomic resolution affects host- parasite association model performance. – *Parasitology* 148: 584–590.
- Dallas, T. A. et al. 2018. Gauging support for macroecological patterns in helminth parasites. – *Global Ecology and Biogeography* 27: 1437–1447.
- Gibb, R. et al. 2021. Data proliferation, reconciliation, and synthesis in viral ecology. – *BioScience* : in press.

- 233 Onstad, D. W. 1997. Ecological Database of the World's Insect Pathogens (ED-  
234 WIP). – Illinois Council on Food and Agricultural Research.
- 235 Patrick, R. et al. 2017. Global mammal parasite database version 2.0. – Ecology .
- 236 Stork, N. E. et al. 2015. New approaches narrow global species estimates for bee-  
237 tles, insects, and terrestrial arthropods. – Proceedings of the National Academy  
238 of Sciences 112: 7519–7523.
- 239 Strona, G. and Lafferty, K. D. 2012. FishPEST: an innovative software suite for  
240 fish parasitologists. – Trends in Parasitology 28: 123.
- 241 Wilkinson, M. D. et al. 2016. The FAIR Guiding Principles for scientific data  
242 management and stewardship. – Scientific data 3: 1–9.

## Tables

Table 1: Files associated with the EDWIP data resource. Metadata is stored in *R* package documentation, allowing the data and metadata to be intrinsically linked. For instance, users can use the help functionality from within *R* to see more information on data columns and unit (e.g., `?nematode`).

filename	rows	columns	description
<code>assocref</code>	11005	16	references for some host-pathogen associations
<code>citation</code>	1966	7	references but no host-pathogen association information
<code>hosts</code>	4392	21	insect host trait data
<code>hostTaxonomy</code>	4489	7	host taxonomic data updated with the <code>getNCBI()</code> function
<code>negative</code>	529	21	information on negative host-pathogen associations
<code>nemaref</code>	338	5	references from nematode pathogens
<code>nematode</code>	234	24	host-nematode interaction data
<code>new_asso</code>	19	25	likely a training document (perhaps do not use)
<code>noassref</code>	569	16	references for some host-pathogen associations
<code>nvpassoc</code>	7164	23	non-viral pathogen infection data
<code>pathogen</code>	2041	9	pathogen trait data
<code>pathTaxonomy</code>	2282	7	pathogen taxonomic data updated with the <code>getNCBI()</code> function
<code>viraref</code>	2124	16	references from viral infections
<code>viruses</code>	1659	25	host-viral interaction data

## Figures

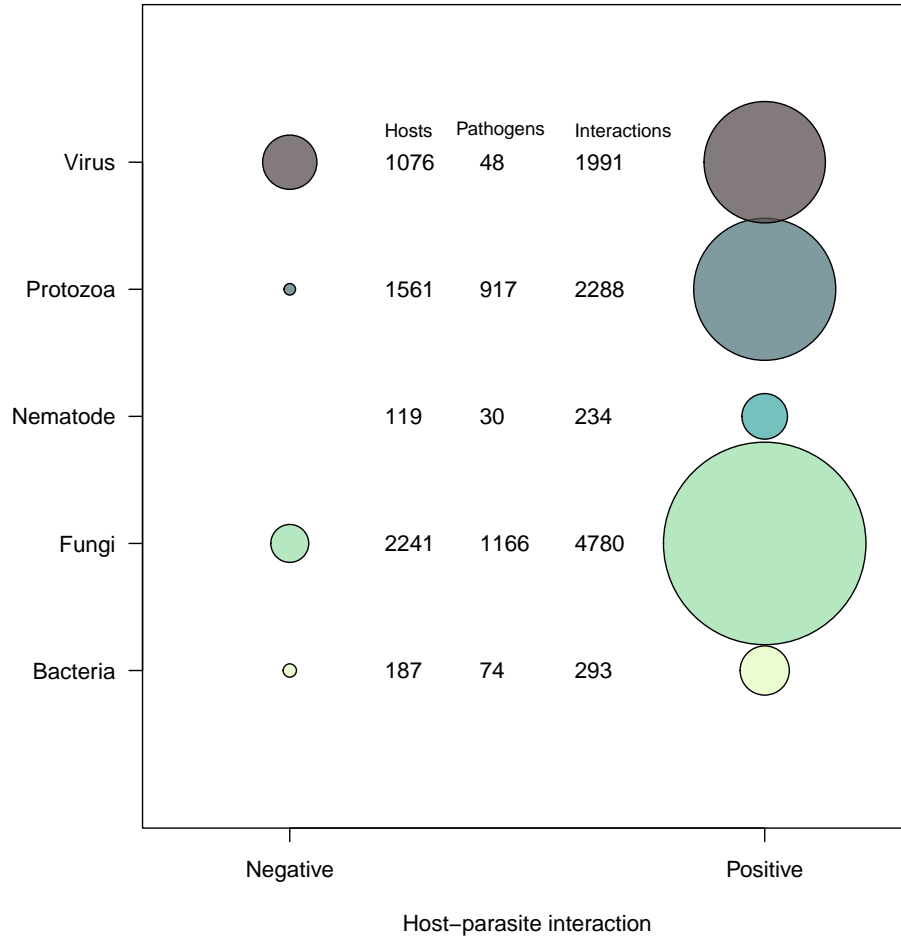


Figure 1: The number of known non-interactions (*negative* left panel) and known interactions (*positive* right panel) for the set of bacterial, fungal, nematode, protozoan, and viral pathogens (*y*-axis). Bubble size is proportional to the total number of interactions associated with that pathogen group and interaction type (i.e., *negative* or *positive*). Numeric columns correspond to the number of unique host species, pathogen species, and interactions for each pathogen group.

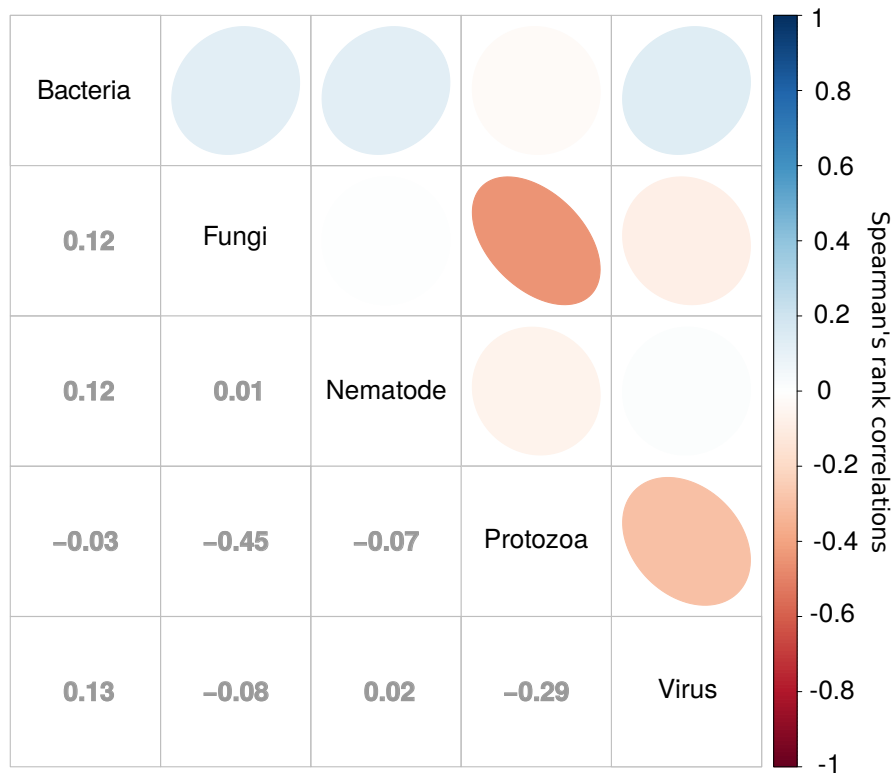


Figure 2: Correlations between each pathogen group in terms of pathogen richness of insect host species, where color corresponds to Spearman's rank correlation values (provided in the lower diagonal matrix). Fungal and protozoan pathogens were negatively related, as were viruses and protozoans. Understanding to what extent this is driven by sampling effects or insect host ecology is an outstanding research question that these could be used to begin addressing.