

insectDisease: programmatic access to the *Ecological Database of  
the World's Insect Pathogens*

Tad A Dallas<sup>a,b,\*</sup>, Colin Carlson<sup>c,d</sup>, Patrick R Stephens<sup>e,f</sup>, Sadie J Ryan<sup>g, h, i</sup> and  
David W Onstad<sup>j</sup>

<sup>a</sup>*Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, 70806*

<sup>b</sup>*Department of Biological Sciences, University of South Carolina, Columbia, SC, 29208*

<sup>c</sup>*Center for Global Health Science and Security, Georgetown University Medical Center,  
Washington, D.C. 20007*

<sup>d</sup>*Department of Microbiology and Immunology, Georgetown University Medical Center,  
Washington, D.C. 20007*

<sup>e</sup>*Center for the Ecology of Infectious Disease and Odum School of Ecology, University of  
Georgia, Athens, GA 30602*

<sup>f</sup>*Department of Integrative Biology, Oklahoma State University, Stillwater, OK 74078*

<sup>g</sup> *Department of Geography, 3125 Turlington Hall, University of Florida, Gainesville, FL,  
32611 USA.*

<sup>h</sup>*Emerging Pathogens Institute, P.O. Box 100009, 2055 Mowry Road, University of Florida,  
Gainesville, FL 32610 USA.*

<sup>i</sup>*School of Life Sciences, University of KwaZulu-Natal, Private Bag X54001, Durban, 4000,  
South Africa*

<sup>j</sup>*Corteva Agriscience, Johnston, IA, 50131*

\*Corresponding author: [tad.a.dallas@gmail.com](mailto:tad.a.dallas@gmail.com)

**Author contributions:** TAD developed the R package, with taxonomic functions developed by CC. All authors contributed to manuscript writing.

**Acknowledgements:** This work was supported by funding to the Viral Emergence Research Initiative (VERENA) consortium including NSF BII 2021909. For more information on the Verena Consortium, see <https://viralemergence.org>. Earlier development of these data was supported by the Macroecology of Infectious Disease Research Coordination Network (NSF/NIH/USDA DEB 131223).

**Conflict of interest:** The authors have no conflicts of interest to declare.

**Keywords:** insect pathogens, crop pest, experimental infection, nematode, virus, protozoa, bacteria

1 **insectDisease**: programmatic access to the *Ecological Database of the World's*  
2 *Insect Pathogens*

### 3 **Abstract**

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

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

22

## 23 Introduction

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

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

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

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

56 Initially created in 1992, the data have been updated prior to 2000, but no clear  
57 semantic versioning was used. As such, it is unclear how long or how frequently  
58 this updating and curation continued, and thus, how many different versions of  
59 the data may be in existence presently. The database we present here, as the  
60 backbone of this R package, represents the most up-to-date version that we know  
61 of, though this may differ slightly from previous descriptions of the data (Braxton  
62 et al., 2003). Generally, we have attempted to preserve all of the original data in  
63 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 (<https://github.com/viralemergence/insectDisease/>) – versioning releases of the data with a permanent identifier (DOI) on Zenodo – and on CRAN (<https://cran.r-project.org/web/packages/insectDisease/index.html>), 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

84 structure of the EDWIP data files, and code to process and join these different  
85 data files is provided in the *R* package vignette.

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

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

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

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

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



125 metadata are also provided in the README file in the top-level of the GitHub  
126 repository.

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

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

change affecting a large set of records in the EDWIP data. All of the data within the `data` and `csv` folders have already gone through these data cleaning steps. However, these data may be dynamic, such that some form of continuous integration or updating of the host and pathogen taxonomy may be necessary. As such, we provide a vignette which transparently shows the steps to clean and augment the data resource, as well as reproduce figures from this manuscript. Finally, we opt to store processed data in the `csv` folder, which contains all data files in `.csv` format. This allows non-*R* users to access the csv-formatted data easily, and ensures long-term stability of the data, as `csv` is a stable text file format. These data are also provided as `.rda` files in the `data` folder.

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

166 data will also only be deployed with a single static copy of the data, such that  
167 users wanting to benefit from versioned and dynamic data will need to access the  
168 data through the GitHub repository. Future efforts to integrate the web interface  
169 and the existing dynamic data structure will be explored, but this is not currently  
170 integrated.

## 171 **Case study: covariance among pathogen groups in parasite** 172 **species richness**

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

186 biocontrol or perhaps for their ease of culture.

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

## 196 **Concluding comments**

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

206 openly, we welcome contributions from researchers interested in augmenting the  
207 data or building off the existing resource.

## 208 **Data accessibility**

209 The `insectDisease` R package is currently available on GitHub  
210 ([github.com/viralemergence/insectDisease](https://github.com/viralemergence/insectDisease)), with ‘.csv’ files in the `csv` directory for  
211 long-term data stability. GitHub releases of the data ensure versioning is main-  
212 tained and all versions are accessible. At the time of this writing, the current  
213 version is 1.2.1 (available at  
214 <https://github.com/viralemergence/insectDisease/releases/tag/1.2.1>).  
215 Releases are given a DOI through integration with Zenodo (available at  
216 <https://doi.org/10.5281/zenodo.5821896>). The package (release version 1.2.1)  
217 is also available on CRAN  
218 <https://cran.r-project.org/web/packages/insectDisease/index.html>.

## 219 **References**

- 220 Bensch, S. et al. 2009. MalAvi: a public database of malaria parasites and related  
221 haemosporidians in avian hosts based on mitochondrial cytochrome b lineages.  
222 – Molecular Ecology Resources 9: 1353–1358.
- 223 Braxton, S. et al. 2003. Description and analysis of two internet-based databases  
224 of insect pathogens: EDWIP and VIDIL. – Journal of Invertebrate Pathology  
225 83: 185–195.
- 226 Carlson, C. J. et al. 2020. What would it take to describe the global diversity of  
227 parasites? – Proceedings of the Royal Society B 287: 20201841.
- 228 Carlson, C. J. et al. 2021. The Global Virome in One Network (VIRION): an atlas  
229 of vertebrate-virus associations. – bioRxiv .
- 230 Chamberlain, S. A. and Szöcs, E. 2013. taxize: taxonomic search and retrieval in  
231 R. – F1000Research 2.
- 232 Dallas, T. A. and Becker, D. J. 2021. Taxonomic resolution affects host- parasite  
233 association model performance. – Parasitology 148: 584–590.
- 234 Dallas, T. A. et al. 2018. Gauging support for macroecological patterns in helminth  
235 parasites. – Global Ecology and Biogeography 27: 1437–1447.
- 236 Gibb, R. et al. 2021. Data proliferation, reconciliation, and synthesis in viral  
237 ecology. – BioScience : in press.

- 238 Onstad, D. W. 1997. Ecological Database of the World's Insect Pathogens (ED-  
239 WIP). – Illinois Council on Food and Agricultural Research.
- 240 Patrick, R. et al. 2017. Global mammal parasite database version 2.0. – Ecology .
- 241 Stork, N. E. et al. 2015. New approaches narrow global species estimates for bee-  
242 tles, insects, and terrestrial arthropods. – Proceedings of the National Academy  
243 of Sciences 112: 7519–7523.
- 244 Strona, G. and Lafferty, K. D. 2012. FishPEST: an innovative software suite for  
245 fish parasitologists. – Trends in Parasitology 28: 123.
- 246 Wilkinson, M. D. et al. 2016. The FAIR Guiding Principles for scientific data  
247 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	16	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	26	host-viral interaction data



## Figures

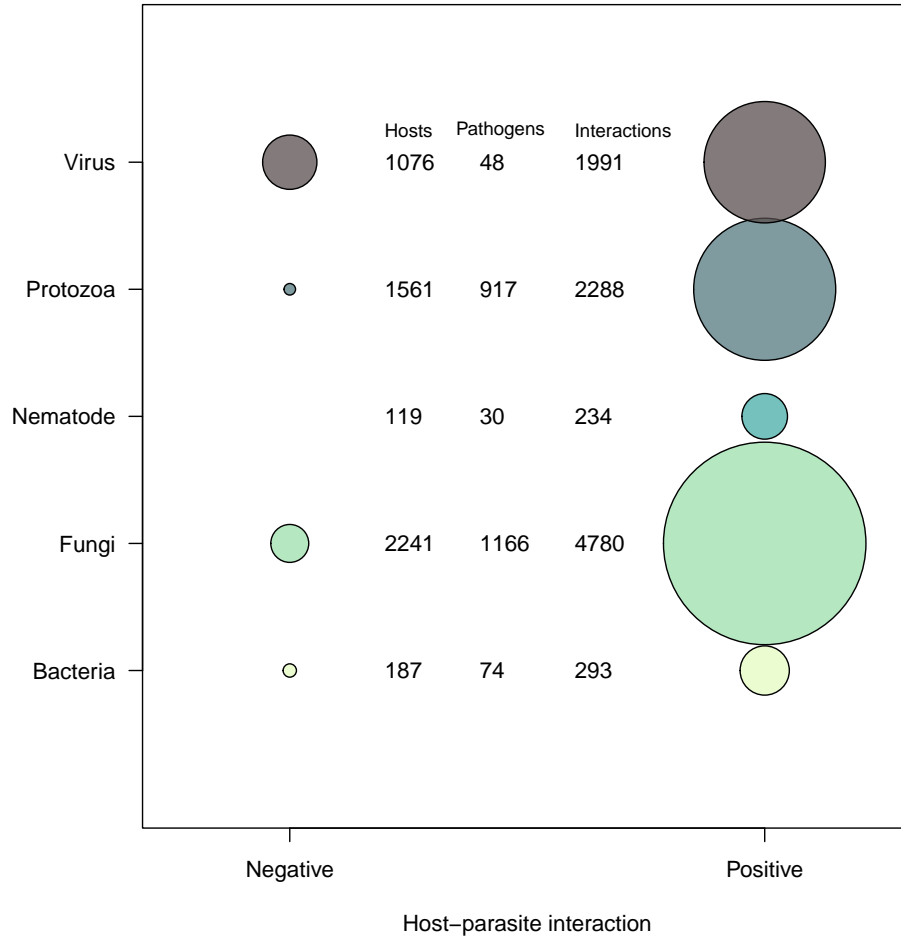


Figure 1: The number of known non-interactions (*negative* left panel) and known interactions (*positive* right panel) for the set of bacterial, fungal, nematode, protozoan, and viral pathogens (*y*-axis). Bubble size is proportional to the total number of interactions associated with that pathogen group 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.

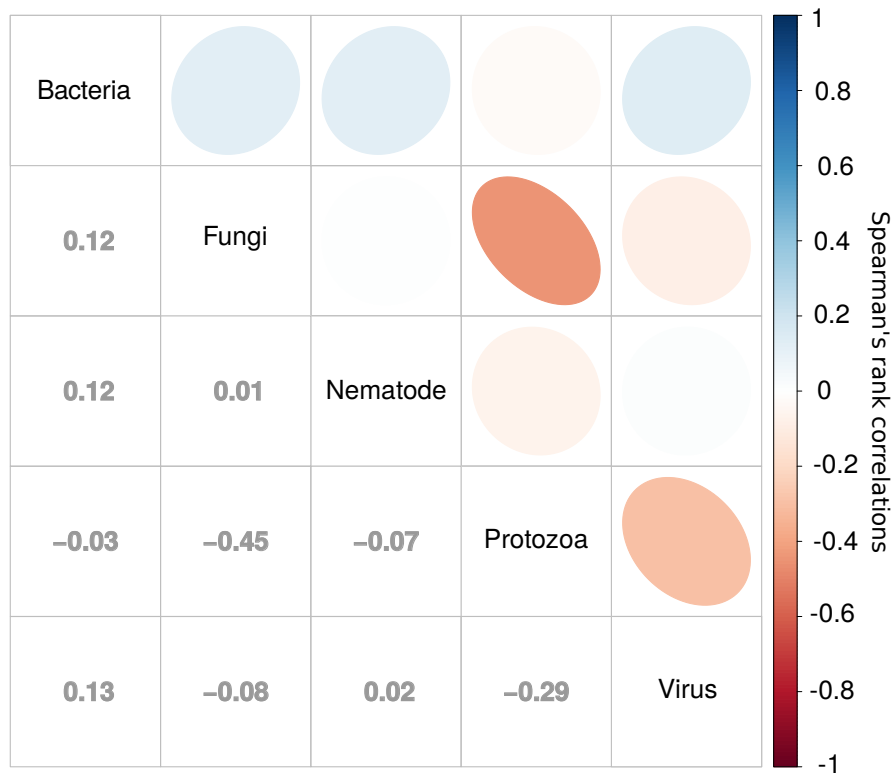


Figure 2: Correlations between each pathogen group in terms of pathogen richness of insect host species, where color corresponds to Spearman's rank correlation values (provided in the lower diagonal matrix). Fungal and protozoan pathogens were negatively related, as were viruses and protozoans. Understanding 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.