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

Tad A Dallas[a,b,*] and Colin Carlson[c]

[a] *Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, USA*

[b] *Department of Biological Sciences, University of South Carolina, Columbia, SC, USA*

[c]

[d]

*Corresponding author: 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: We sincerely thank all individuals associated with the collection and curation of the *Ecological Database of the World's Insect Pathogens* data, specifically David Onstad at the US Department of Agriculture.

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

Keywords: insect pathogens, crop pest, experimental infection,

Abstract

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

Introduction

There are a number of data sources documenting host-pathogen associations, especially for pathogens of mammals (Gibb et al., 2021, Patrick et al., 2017), birds (Bensch et al., 2009), and fish (Strona and Lafferty, 2012). These have been incredibly important to our understanding of what determines pathogen host range, pathogen species richness across a set of hosts, and overall host-pathogen network structure (e.g, (Carlson et al., 2020, Dallas et al., 2018)). But while some host groups are well-studied, there are taxonomic gaps in our understanding of host-pathogen associations. For instance, insect host-pathogen relationships have considerably less open-source data available, despite their inherent importance to agricultural crops and vector-borne disease, in addition to the numerical dominance of insect species over other taxa (Stork et al., 2015). This is a clear knowledge gap.

Additionally, all of the previous data sources listed before have dedicated researchers and government agencies responsible for data deposition and curation in openly accessible formats. However, some data have not been as lucky, at no fault of the original data curators. These data run the risk of disappearing into a file drawer or on an external hard drive, potentially shared with a small number of researchers but not accessible to the scientific community at large. One data resource arguably close to this point of disappearance is the *Ecological Database of the World's Insect Pathogens* (EDWIP) (Onstad, 1997).

The EDWIP data consist of experimental infections and field observations of the interactions between insect hosts and a number of nematode, viral, protozoan, fungal, and bacterial pathogens (Braxton et al., 2003). In the EDWIP data, a number of pathogen infection types were excluded, including vector-borne pathogens which do not infect the insect host. An interesting component of EDWIP is the existence of negative associations – attempts to inoculate a host with a given pathogen that failed to infect – for some host groups (Figure 1). Failed infections represent *true* absences or incompatibilities between a given host and pathogen. These data are incredibly useful to pathogen host range and host-pathogen interaction modeling, but we rarely have data on these known non-interactions. Initially created in 1996, the data have been dynamically updated over time. However, it is unclear how long this curation continued, and how many different versions of the data may be in existence without proper versioning. To this point, the numbers of insect-host pathogen interactions differs slightly from previously published versions of the data (Braxton et al., 2003), but these changes are relatively minor.

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. 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) makes sure that host and pathogen taxonomic names are updated.

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.

Each of the pathogen groups slightly differs in the available data on experimental infections. For instance, nematode infections contain information on soil type and associated bacteria, virus infection data has information on viral dose, and non-viral pathogens (protozoans, fungi, and bacteria) have information on intermediate host species.

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

Metadata and package documentation Differences in features across the data on different pathogen types (e.g., ?nematodes relative to ?viruses) make combining these data difficult without a loss of information. However, there is clear utility in having all of the data in a single standardized form. As such, we have documented each data resource using R package documentation, allowing the metadata of each data product to be examined directly from R using the help() function or the question mark notation (e.g., ?viruses).

Data cleaning and taxonomic resolution The initial data structure was maintained from the original raw data files provided by David Onstad, principal maintainer of the EDWIP data resource (Onstad, 1997). This includes files such as new_assoc, as this was likely a test file containing pathogen species such as "wormy thing". We clean and augment this existing data source programmatically, with much of this code in the insectDisease package vignette.

First, we resolve host and pathogen names using the R package taxize, using NCBI taxonomic backbone (Chamberlain and Szöcs, 2013). While this seems like a small change, it both standardizes host and pathogen nomenclature, and catches any taxonomic changes that have occurred in the past couple decades. This includes the consideration of microsporidian parasites as fungi, not protozoans, a change affecting a large set of records in the EDWIP data. Second, we provide a vignette which combines the different host-pathogen data together to form an informative set of host-pathogen associations, including both both known associations and failed experimental infections, as delineated by the `interaction` column in the resulting `edwip.csv` file. Finally, processed data by the vignette is output in `.csv` format in the vignettes folder, but all data stored as R data objects (`.rda`) are also converted to csv and placed in the `csv` folder, allowing non-R users to still access the data in a stable format.

Case study: covariance among pathogen groups in parasite species richness

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

This potential sampling bias among insect host species would be clear if there were a positive relationship between the number of positive interactions to the number of negative interactions for a host species, as it would indicate that insect host species with lots of known interactions also tended to appear in many studies and have some negative interactions as well. We find evidence for a significantly negative relationship based on a Spearman's rank correlation (ρ = -0.1, p $<$ 0.0001).

Concluding comments

Ecological data – while growing in availability, size, and stability – are still a resource that should not be allowed to fade out of existence. The EDWIP data provided to the authors were in a proprietary format ('Claris FileMaker Pro 5') that was already over 10 major versions behind. With limited inter-version operability (e.g., `.fmp5` files cannot be opened in more recent versions of the software, or require multiple conversion steps), these data seemed as if headed towards obsolescence. The insectDisease package ensures that these data will be available to the broadest set of researchers, be bound to relevant metadata, and be properly versioned. By hosting the data openly, we welcome contributions from researchers interested in augmenting the data or building off the existing resource.

Data accessibility

The insectDisease R package is currently available on GitHub https://github.com/viralemergence/ with '.csv' files in the csv directory for long-term data stability.

Finally, the data are periodically archived at major version changes via Zenodo ??????????????????

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.

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 .

Onstad, D. W. 1997. Ecological Database of the World's Insect Pathogens (ED-WIP). – Illinois Council on Food and Agricultural Research.

Patrick, R. et al. 2017. Global mammal parasite database version 2.0. – Ecology .

Stork, N. E. et al. 2015. New approaches narrow global species estimates for beetles, insects, and terrestrial arthropods. – Proceedings of the National Academy of Sciences 112: 7519–7523.

Strona, G. and Lafferty, K. D. 2012. Fishpest: an innovative software suite for fish parasitologists. – Trends in parasitology 28: 123.

Figures

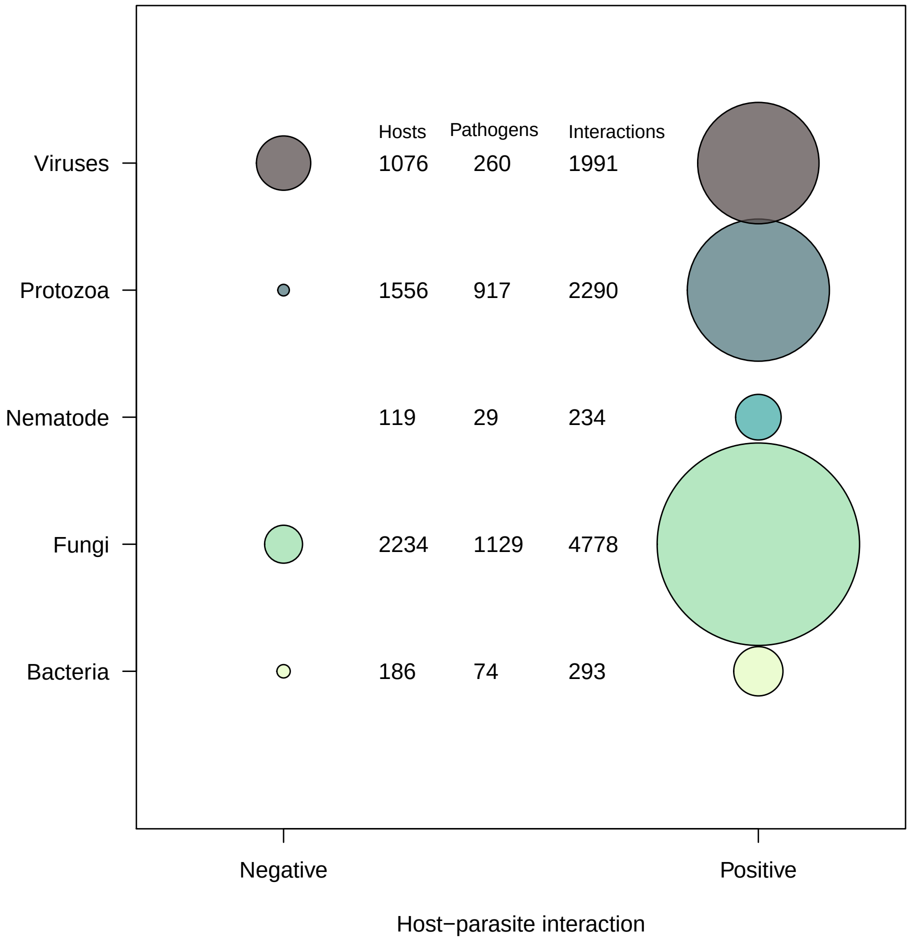

Figure 1: Bubble plot, where points are proportional to the total number of negative (on the left) and positive (on the right) host-pathogen interactions for each pathogen group (y-axis). Numeric columns correspond to the number of unique host species, pathogen species, and interactions for each pathogen group.

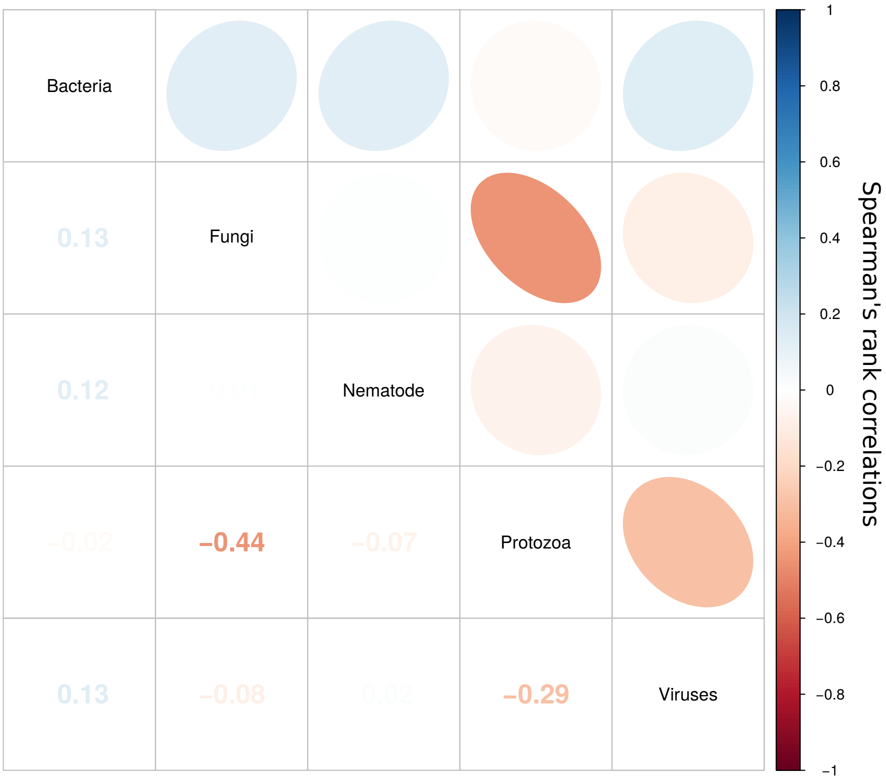

Figure 2: Correlations between each pathogen group in terms of pathogen richness of insect host species, where color corresponds to Spearman's rank correlation values (provided in the lower diagonal matrix). Fungal and protozoan pathogens seemed to negatively covary, as did viruses and protozoans. Understanding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