Title: helminthR: An R interface to the London Natural History Museum's Host-Parasite Database

Running title: Global helminth data access

Author: Tad Dallas

Affiliation: University of Georgia, Odum School of Ecology, Athens Ga 30602

Email: tdallas@uga.edu

Abstract

The understanding of the diversity and distribution of helminth parasites is currently constrained by the limited number of host-parasite interaction databases, and the difficulty in accessing existing data. The London Natural History Museum’s Host-Parasite Database represents one such underutilized database, containing over a quarter million helminth parasite occurrence records, accessible through a web interface. To enable users to programmatically search and manipulate data from this database, I developed an `R` package called `helminthR`. Here, I introduce the core functions of the package, and detail how `helminthR` can be used to obtain host-parasite interaction records, citations for interactions, and host taxonomic data.

Introduction

Helminth parasites are one of the most common infectious agents to humans (Stoll 1947, Hotez et al. 2008, De Silva et al. 2003), wild animals (Poulin and Valtonen 2002, Jolles et al. 2008), and livestock (Over et al. 1992, Morgan et al. 2013). Limitations in data availability have hampered our understanding of the spatial distribution of helminth parasites, and associations between helminth parasites and both human and wildlife hosts. Further, there is a need for basic scientific research into the community ecology and macroecology of host-helminth associations (Rohde 2002). Such efforts could provide tests of principles from community ecology, and macroecological patterns in parasites.

To address these research concerns, data on host-helminth associations across broad spatial scales are needed. Efforts to document known host-parasite associations in large databases are fairly recent, and represent valuable resources for researchers (see Strona et al. 2013, Gibson et al. 2005, Nunn and Altizer 2005). However, a portion of these databases are not openly accessible, requiring users to contact database administrators or to copy data from web interfaces. These methods of accessing databases may lead to transcription errors, duplicated efforts among labs, and create static copies of the data that are difficult to update if and when new data are added. Allowing host-parasite databases to be open and easy to access may promote open and reproducible science, and would potentially promote the discovery of “general laws” in parasite ecology (Poulin 2007).

To this end, I have developed an R package capable of extracting information from a large global database of host-helminth parasite occurrence records maintained by the London Natural History Museum (NHM; Gibson et al. 2005). This curated database includes more than 250,000 host-helminth records from over 28,000 published peer-reviewed articles. However, the web interface of the database makes data analysis difficult, which subsequently limits the use of this data resource by researchers (but see Strona and Fattorini (2014) and Wells et al. (2015)). The goal of the `helminthR` package is to make all the data contained in the London Natural History Museum's database accessible from R, a commonly used open source statistical programming environment (R core team 2015).

Core package functionality

Here, I explore the core functions of the `helminthR` package, and then demonstrate the utility of `helminthR` for creating host-parasite interaction networks. `helminthR` relies on several packages that interface with html and xml, including `rvest` (Wickham 2015) and `xml2` (Wickham 2015b). Currently, `helminthR` is available on Github, and is hosted by the rOpenSci collective, a group of scientists and developers committed to creating packages to promote open science, including the creation of packages to access online data sources. The package can be easily downloaded using the `devtools` package, using the following R code.

```
devtools::install_github('ropensci/helminthR')
library('helminthR')
```

Downloading and using this package does not require the user to have a Github account, unless they would like to actively contribute to package functionality, or file an issue.

Querying the database

Host-parasite records in the NHM database contain information on host and parasite species, one or more citations for the host-parasite association, and the location of the interaction georeferenced to the country, state (for the United States), or water body (e.g. Lake Erie) level. Queries can be made to find all interactions of a known host species (`findHost`), all interactions of a known parasite species (`findParasite`), or all interactions at a specific geographic location (`findLocation`). Links to citations for a given helminth record can be obtained from any of the functions listed above by setting the `citation` argument to `TRUE`.

When querying the database for known hosts or helminths, the user can input genus and/or species name in order to query different taxonomic levels of host or parasite. Further, `findParasite` can find host-helminth records given a parasite group (Cestodes, Acanthocephalans, Monogeneans, Nematodes, Trematodes, or Turbellarian) or subgroup. The following example code would find all interactions of nematodes in the genus *Strongyloides*.

```
StrongHosts <- findParasite(genus='Strongyloides', validateHosts=FALSE)
```

The resulting structure of `strongHosts` is a host-parasite matrix in the form of a three (or four) column `data.frame` containing host and parasite names, parasite full name, and citation (if the `citation` argument is set as `TRUE`). The argument `validateHosts` provides taxonomic information on hosts from the Catalogue of Life (Roskov et al. 2015). While slightly slow, this removes questionable hosts, and validates species names (when `validateHosts=TRUE`), returning a `list` object containing the `data.frame` described above, and the taxonomic information for all hosts. This structure is maintained when querying using any of the “find*” functions, including `findHost`, `findParasite`, and `findLocation`. The following code demonstrates the `findHost` function in order to find helminth occurrence records in wild individuals of *Gorilla gorilla* (using the `hostState` argument). The user can also query captive hosts, domesticated hosts, or hosts used in commercial applications.

```
gorillaParasites <- findHost(genus = 'Gorilla', species = 'gorilla',
                             hostState = 1 )
```

The final core function in the `helminthR` package queries all host-parasite interactions for a given geographic location. A list of locations capable of being queried is provided by the `listLocations` function, and a cached copy of these data is provided as a data object (using the command `data(locations)`). Georeferencing of these data is performed using the `geocode` function in the `ggmap` package (Kahle & Wickham 2013). The user is responsible for ensuring the accuracy of the provided latitude and longitude coordinates. Further care should be taken when searching by location, as some locations may be nested within others (e.g. “South America” is a valid location query, but many countries in South America are also valid queries). Below, I demonstrate the functionality by finding all host-parasite associations recorded in France where the host was “in the wild” (i.e., `hostState = 1`), removing occurrence records where the host or parasite has parantheses (e.g. “(freshwater_fish)”) or is identified to be at the genus level (e.g. “Sanguinicola spp.”) by setting the argument `speciesOnly` to be `TRUE`. The result is a host-parasite association list containing information on host-helminth associations, including links to the original citations. It is important to note that not all interactions will be unique, so the user must use the `unique` function on the `Host` and `Parasite` columns of the output `data.frame`.

```
# Find all host-helminth associations occuring in France
FrenchHostPars <- findLocation(location = 'France', speciesOnly = TRUE,
                               citation = TRUE)

# Find unique host-parasite associations
FrenchHostParsUnique <- unique(FrenchHostPars[,1:2])
```

Visualizing host-parasite networks

The above code demonstrates the functionality of the `helminthR` package for querying host-parasite interactions by host and parasite genus and/or species, and also for locating all host-parasite interactions in a given country or locality. Using the `findLocation` function, I queried the database for all host-parasite interactions occurring within Lake Erie, one of the US Great Lakes, and visualized the resulting host-parasite interaction network (Figure 1) using the `igraph` `R` package (Csardi & Nepusz 2006). Detailed code to create this type of visualization is provided in the supplement.

Data limitations

The data contained in the London Natural History Museum’s Host-Parasite Database represent a valuable resource, but are not without limitation. First, the data are from studies published anytime after 1922, and the data owners themselves accept no responsibility for data accuracy.

Second, the data are only georeferenced to the country level in most cases, which limits their application. However, citations are given for each host-parasite association, and an attempt has been made to obtain latitude and longitude values for the centroids of countries (using the command `data(locations)`). While this may be time consuming, the examination of original references would help assure data quality, and provide more fine georeferencing. Nevertheless, the data can still be used to address many macroecological patterns in their current form. For example, data on aquatic and marine parasites are georeferenced to coastal areas (e.g. “Coast of New Guinea”) or larger bodies of water (e.g. “Aral sea”), providing a way to apply macroecological theory to largely unexplored questions related to the diversity and distribution of marine parasites (Rohde 2002, Rhode 2010).

Conclusions

In this paper I have shown how the R package `helminthR` permits the programmatic access of the Natural History Museum Host-Parasite Database, making it easy to generate host-parasite networks at different geographical scales spanning from local to global. This database represents one of the most complete aquatic host-parasite databases (but see Strona et al. 2013), providing data on parasite occurrences for both terrestrial and aquatic hosts. With any luck, `helminthR` will promote the application of concepts from community ecology and macroecology to parasite communities at a broader spatial scale. This project is hosted on Github, and uses TravisCI for continuous integration of the package on different `R` versions. Issues or improvements can be suggested at this link (https://github.com/ropensci/helminthR/issues).

To cite `helminthR` or acknowledge its use, cite the original data source (Gibson et al. 2005), and this Software note as follows, substituting the version of the application that you used for `ver. xxx`:

T. Dallas 2015. `helminthR`: An R interface to the London Natural History Museum's Host-Parasite Database - Ecography (ver. xxx).

Acknowledgments

I thank the London Natural History Museum, and specifically D.A. Baylis, the original curator, and the current curation team (D. Gibson, R. Bray, and E. Harris). Colin Carlson, Kevin Burgio, Alexa McKay, Maxwell Joseph, and Giovanni Strona provided thoughtful comments on earlier drafts. I thank my advisor, John Drake, and the developers at rOpenSci for their guidance, support, and general views on open science.

References

Crompton, D. W. T. and Nesheim, M. 2002. Nutritional impact of intestinal helminthiasis during the human life cycle. -Annual Review of Nutrition 22: 35-59.

Csardi G, Nepusz T. 2006. The igraph software package for complex network research. -InterJournal, Complex Systems 1695. http://igraph.org

De Silva, N. R. et al. 2003. Soil-transmitted helminth infections: updating the global picture. -Trends in Parasitology 19: 547-551.

Gibson, D. et al. 2005. Host-parasite database of the natural history museum, London. http://www.nhm.ac.uk/research-curation/scientific-resources/taxonomy-systematics/host-parasites/database/index.jsp .

Hotez, P. J. et al. 2008. Helminth infections: the great neglected tropical diseases. -Journal of Clinical Investigation 118: 1311-1321.

Jolles, A. E. et al. 2008. Interactions between macroparasites and microparasites drive infection patterns in free-ranging african buffalo. -Ecology 89: 2239-2250.

Kahle, D. and H. Wickham. 2013. ggmap: Spatial Visualization with ggplot2. -The R Journal, 5(1): 144-161. http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

Morgan, E. R. et al. 2013. Global change and helminth infections in grazing ruminants in Europe: impacts, trends and sustainable solutions. -Agriculture 3: 484-502.

Nunn, C. L. and Altizer, S. M. 2005. The global mammal parasite database: an online resource for infectious disease records in wild primates. -Evolutionary Anthropology: Issues, News, and Reviews 14: 1-2.

Over, H. J. et al. 1992. Distribution and impact of helminth diseases of livestock in developing countries. 96. -Food & Agriculture Organization.

Poulin, R. and Valtonen, E. T. 2002. The predictability of helminth community structure in space: a comparison of fish populations from adjacent lakes. -International Journal for Parasitology 32: 1235-1243.

Poulin, R. 2007. Are there general laws in parasite ecology?.-Parasitology 134(06): 763-776.

R Core Team 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rohde, K. 2002. Ecology and biogeography of marine parasites. - Advances in Marine Biology 43: 1-83.

Rohde, K. 2010. Marine parasite diversity and environmental gradients. In: Morand, S, and BR Krasnov, (eds.) The biogeography of host-parasite interactions. Oxford, UK: Oxford University Press. pp. 73-88.

Roskov Y. et. al. 2015. Species 2000 & ITIS Catalogue of Life, 2015 Annual Checklist. Digital resource at www.catalogueoflife.org/annual-checklist/2015. Species 2000: Naturalis, Leiden, the Netherlands. ISSN 2405-884X.

Stephenson, L. S. et al. 2000. Malnutrition and parasitic helminth infections. -Parasitology 121: S23-S38.

Stoll, N. R. 1947. This wormy world. -The Journal of Parasitology 33: 1.

Strona, G. and Fattorini, S. 2014. Parasitic worms: how many really? -International Journal for Parasitology 44: 269-272.

Strona, G. et al. 2013. Host range, host ecology, and distribution of more than 11,800 fish parasite species: Ecological archives e094-045. -Ecology 94: 544-544.

Wells, K. et al. 2015. The importance of parasite geography and spillover effects for global patterns of host -parasite associations in two invasive species. -Diversity and Distributions 14 21: 477-486.

Wickham, H. 2015. rvest: Easily Harvest (Scrape) Web Pages. R package version 0.3.1. http://CRAN.R-project.org/package=rvest

Wickham, H. 2015b. xml2: Parse XML. R package version 0.1.2. http://CRAN.R-project.org/package=xml2

Figure captions

FIG. 1. The host-parasite association network for Lake Erie, one of the Great Lakes located in the Northern United States. Grey lines between boxes represent interactions between hosts (larger blue dots) and helminth parasites (smaller black dots).