J. Parasitol., 99(1), 2013, pp. 6–10
© American Society of Parasitologists 2013

PREDICTING WHAT HELMINTH PARASITES A FISH SPECIES SHOULD HAVE USING PARASITE CO-OCCURRENCE MODELER (PaCo)

Giovanni Strona and Kevin D. Lafferty*

Department of Biotechnology and Biosciences, University of Milano Bicocca, Piazza della Scienza 2, 20126 Milan, Italy. *e-mail: giovanni.strona@unimib.it*

ABSTRACT: Fish pathologists are often interested in which parasites would likely be present in a particular host. Parasite Co-occurrence Modeler (PaCo) is a tool for identifying a list of parasites known from fish species that are similar ecologically, phylogenetically, and geographically to the host of interest. PaCo uses data from FishBase (maximum length, growth rate, life span, age at maturity, trophic level, phylogeny, and biogeography) to estimate compatibility between a target host and parasite species–genera from the major helminth groups (Acanthocephala, Cestoda, Monogenea, Nematoda, and Trematoda). Users can include any combination of host attributes in a model. These unique features make PaCo an innovative tool for addressing both theoretical and applied questions in parasitology. In addition to predicting the occurrence of parasites, PaCo can be used to investigate how host characteristics shape parasite communities. To test the performance of the PaCo algorithm, we created 12,400 parasite lists by applying any possible combination of model parameters (248) to 50 fish hosts. We then measured the relative importance of each parameter by assessing their frequency in the best models for each host. Host phylogeny and host geography were identified as the most important factors, with both present in 88% of the best models. Habitat (64%) was identified in more than half of the best models. Among ecological parameters, trophic level (41%) was the most relevant while life span (34%), growth rate (32%), maximum length (28%), and age at maturity (20%) were less commonly linked to best models. PaCo is free to use at www.purl.oclc.org/fishpest.

Knowing which parasite species have been reported from a host has many uses in ecology. Determining what additional parasite species are likely to parasitize a host is important because most hosts have never been sampled for parasites, and the remaining hosts are often undersampled. Having a list of likely parasite species can guide efforts to sample for additional parasite species in a host and may also indicate something about the factors that determine parasite communities and distributions. However, such lists do not supplant traditional dissections and identifications.

There are 2 ways one could hypothesize about the potential parasites of a host. First, one might use a list of parasite species known from the target host to specify lists of additional parasite species associated with the known parasites in other host species. The length of the parasite list from the known host heavily influences this list-centric approach. For instance, if the host has never been sampled for parasites, it becomes impossible to hypothesize potential parasites. Unfortunately, unsampled hosts are precisely those for which it is most difficult to predict parasites. In addition, the list-centric approach tells us little about parasite community structure.

Alternatively, if the ecology, phylogeny, and biogeography of a host help shape its parasite community, we might expect that similar hosts would share parasite species. Here, we took a second, host-centric, approach to identify potential parasite species. Unlike the list-centric approach, a host-centric approach works even for hosts with no known parasite records, and it avoids the potential circularity of choosing parasites based on those parasites already known from a host. Finally, it allows researchers to test hypotheses regarding factors that determine parasite communities. The challenge of implementing a host-centric approach is that the modeling requires access to vast quantities of data which, until recently, were not available to researchers.

Parasite Co-occurrence Modeler (PaCo) meets the data needs of a host-centric approach and creates lists of potential parasite species for more than 27,400 fish species. PaCo is a dynamic web system that uses the open-source scripting language Python (van Rossum and de Boer, 1991) and the Python-based web framework Django (Holovaty and Kaplan-Moss, 2009). It is a component of Fish PEST (Fish Parasite Ecology Software Tools), which is a web project designed to integrate parasitology data with FishBase (Strona and Lafferty, 2012a, 2012b). From this point, for any specific term or subject directly regarding FishBase, readers should refer to Froese and Pauly (2000) or to FishBase online documentation at www.fishbase.org/manual/Key%20Facts.htm.

The host–parasite records are a validated subset of 64,000 records of helminths from fishes that we acquired from Hewitt and Hine (1972), Williams and Bunkley-Williams (1996), Holland and Kennedy (1997), Kohn and Cohen (1998), Gibson et al. (2005), Kohn et al. (2006), Salgado-Maldonado (2006), Cohen and Kohn (2008), Harris et al. (2008), Salgado-Maldonado (2008), Strona et al. (2009b), and Lichtenfels et al. (2011). Throughout, we refer to these helminth species as "parasites," but remind the reader that protozoan, crustacean, and other parasitic groups are not yet part of the database. Host scientific names were validated according to FishBase (Froese and Pauly, 2012), whereas parasite scientific names were validated according to the Catalogue of Life (Bisby et al., 2012) and WoRMS (Appeltans et al., 2011). Invalid synonyms were replaced with the corresponding current valid names, and we excluded any ambiguous records or any host or parasite records at a taxonomic level higher than species. Parasite records of species not listed in either the Catalogue of Life or in WoRMS were excluded as well. We acquired approximately 20,000 valid host–parasite records, but these 2 databases are far from complete and are constantly updated. The absence of a parasite species from the 2 aforementioned databases does not imply that the corresponding record is incorrect; therefore, as an option, PaCo includes a list of 40,000 host–parasite records where only host names were validated according to FishBase (Froese and Pauly, 2012). Still, readers should be aware that taxonomy is a dynamic field and parasite lists are not comprehensive. There will be parasites that have been reported from fish species that are not included in the PaCo database.

PaCo relies on existing databases for analysis. Therefore, the choices made in constructing the model were constrained by the

Received 17 March 2012; revised 6 June 2012; accepted 25 July 2012.

* U.S. Geological Survey, Western Ecological Research Center, Marine Science Institute, University of California, Santa Barbara, California 93106.

DOI: 10.1645/GE-3147.1

structure or contents in the database. Ecological information for the fish species (>27,400) was obtained from FishBase Species Ecology Matrices. Data in the Species Ecology Matrices were extracted using a script based on the Python HTML/XML parser Beautiful Soup (www.crummy.com/software/BeautifulSoup/). FishBase has many data fields but some are highly correlated, several have outputs that are difficult to analyze, and others are inconsistently reported across species. Of the fish ecology parameters available from the Species Ecology Matrices in FishBase, PaCo uses maximum length (max L), growth rate (K), i.e., the rate at which the asymptotic length is approached, life span (Y), age at first maturity (Ym), and trophic level (T). Habitat and geographical information for each host species were also collected from FishBase. Our selection of these factors does not imply that they necessarily affect parasite communities. For instance, it is not clear how age at maturity or growth rate of a species would affect parasites. In other cases, however, links between parasites and ecology are better known; parasites accumulate with fish size, increase with fish trophic level, and vary by habitat and geography (Poulin, 1994, Rohde et al., 1995).

A good model should identify parasites that are already known from the target host. When proposing a list of probable parasites for a particular fish species, PaCo ignores the parasites of the target host. Candidate parasite species come from non-target hosts. Any parasites suggested are based on their presence in similar host species. Therefore, one way to evaluate PaCo is to consider the fraction of parasites already known from a target host that are on the proposed list of parasites for that target host. However, because PaCo only uses parasites from other hosts as candidates, parasites specific to the target host cannot be proposed. This affects how one evaluates the performance of PaCo. Specifically, when evaluating how well PaCo predicts known parasites for a host, it is important to exclude parasites specific to the target host from consideration, i.e., they should not count in the denominator of the fraction of known parasites.

The algorithm underlying PaCo assumes a host would be most likely to share generalist parasites with similar fishes, where similarity can be expressed in terms of ecology, phylogeny, geography, and habitat. The PaCo algorithm considers these factors in 3 steps. First, it computes the ecological similarity (ES) between the target host (H) and all the other fish species in the internal database (h_i). ES for each host is calculated as follows:

$$ES = 1/(Euc(H,h_i) + 1),$$

where Euc(H, h_i) is the Euclidean distance between H and h_i calculated on the basis of user-selected niche dimensions, i.e., max L, K, Y, Ym, and T (see above). Second, it weights ES values by applying a user-selected combination of habitat, geography and phylogenetic filter. Third, it assigns to each parasite species a C-score which is obtained by summing up all the weighted similarity values of the hosts (excluding the target host) that the parasite infects. By summing ES values, PaCo considers both the ecological similarity between the target host and any other fish species where candidate parasites occur and the host specificity of these parasites.

The filters work as follows:

Habitat (HAB): ES of a fish species is multiplied by the coefficient HAB, which represents the fraction of preferences in habitat usage the fish shares with the target host. HAB of a species is computed as: HAB = (M + B + F)/HH, where M, B, and F indicate the respective overlaps in marine, brackish, and freshwater habitat preference between the fish species and the target host and HH indicates the total number of habitats where the target host occurs. HAB may, therefore, vary between 0 (no overlap) and 1 (maximum overlap). The use of very broad environmental categories is a tradeoff between data availability (finer habitat data are not available for all species) and the necessity to include in the model the fundamental effect of salinity in the determination of host–parasite assemblages (e.g., Poulin, 1994).

Geography (GEO): ES of a host species is multiplied by the coefficient GEO, which represents the average overlap in the geographic distribution between the fish species and the target host. GEO of a fish species is computed as GEO = GO/GH, where GO indicates the number of localities where both the fish species and the target host occur, while GH indicates the total number of localities where the target host occurs. GEO may, therefore, vary between 0 (no overlap) and 1 (maximum overlap).

Phylogeny (PHY): ES of a host species is multiplied by the coefficient PHY, which represents the average phylogenetic overlap between the fish species and the target host. PHY can be scaled taxonomically. The recommended option is PHY = (s + g + f)/3N, where s, g, and f indicate the match between family (f), genus (g), and species (s) of a fish in the database and those of target host. A phylogenetic overlap at a taxonomic level is scored as 1; otherwise, the value is 0. A value of 1 for s will lead to a 1 for g and f, and a value of 1 in g will lead to a 1 for f (but not vice versa), providing a balance among different taxonomical levels. PHY may, therefore, vary between 0 (no overlap) and 1 (maximum overlap). As an option, PaCo includes other sets of taxonomic levels that embrace order, class, or both (note that the procedures and the results reported in the following paragraphs refer to the default setting of the filter, s-f-g).

Habitat, geography, and phylogeny are used as filters (and not as ecological parameters) in order to maximize their effect on ES scores. For example, if HAB filter is applied, hosts inhabiting environments different from those inhabited by target host will get a weighted ES score equal to 0 independently from their ecology, phylogeny, and geography.

The user is free to manipulate the model parameters to make the output list more-restrictive or less-restrictive according to variable influences of each parameter. Each computed model indicates the most probable parasites for a fish host under the hypothesis that the user-selected parameters influence the distribution of parasites on that host. One way to assess the validity of the model is to compare its output with the known list of non-host specific parasites for a target host species. For target host species with reported parasites, PaCo computes a Model Evaluation Value (MEV). This value comes from the list of proposed parasites, and is computed as:

$$1 - (\Delta_C/Cmax),$$

where Δ_C is the average difference between the maximum C-score and the C-score of each reported parasite and Cmax is the maximum C-score. MEV may vary between 0 and 1. The closer MEV is to 1, the better the model. To improve a model, users can experiment with different filters and weights and compare their effects on the MEV.

Although users can choose the minimum and maximum number of suggested hosts that should show up in the output

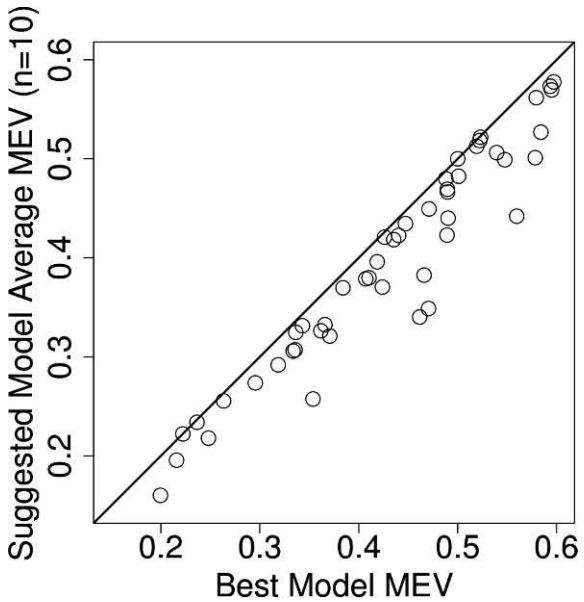

FIGURE 1. Overall performance of PaCo optimal model search procedure illustrated by plotting the average MEVs of 10 suggested models for 58 fish hosts, known to harbor at least 30 parasite species, against the corresponding highest achievable MEVs (obtained by testing all the 248 possible combinations of filters and niche dimensions). The continuous line represents the line of equality.

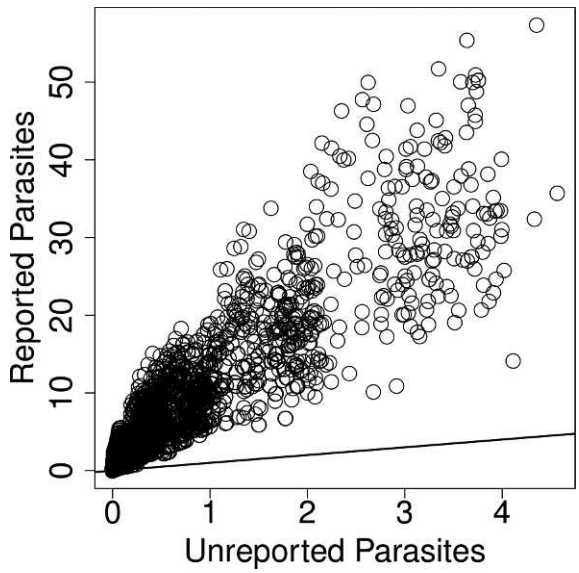

FIGURE 2. Performance of the PaCo algorithm, illustrated by plotting the averaged C-scores of reported and unreported parasite for each of the 248 possible combinations of filters and niche dimensions (means were calculated for each combination of parameters on 58 host species known to harbor at least 30 parasite species). The continuous line represents the line of equality.

list, and the minimum percentage of known parasites of the target host that should be included in the output list, MEV is always computed on the basis of the complete list of known parasites.

MEV is reported on the PaCo output page together with the average C-score of all the parasite species of the internal database belonging to the user-selected parasite taxon–taxa. Additionally, the output page reports the average of the C-scores of parasites reported from the target host and the average of the C-scores of parasites not reported from the target host. These values should be used as an indication of model robustness by evaluating the difference between the average C-score of reported parasites with respect to the average C-score of all the other parasites: a high value of the first with respect to the latter indicates that the model has done well at identifying potential parasite species for the target host. However, because C-scores are not standardized, i.e., they cannot be used to compare different models, we recommend that users refer mainly to the MEV as an estimate of model goodness of fit.

While users can evaluate models based on trial and error, PaCo can attempt to find the optimal model for host species with at least 5 known parasite species from the user-selected parasite taxon–taxa. First, a genetic algorithm is used to identify an optimal combination of niche dimensions. The genetic algorithm works as follows: A set of 6 models is created by randomly extracting host niche dimensions; MEVs of the 6 models are computed; and niche dimensions are randomly extracted from the 6 models, with a probability proportional to the corresponding MEV, and used to populate a second generation of 5 models. With a probability of 0.1 ("mutation rate"), niche dimensions are randomly assigned instead of being extracted. This allows the introduction of niche dimensions not included in the first generation of models, i.e., MEVs of the second generation models are computed; as before, niche dimensions of the second generation models are used to populate a third generation of 4 models; MEVs of the third generation models are then computed; and the 15 models generated in the 3 generations are ranked according to their MEVs. At the second step, 7 models are created by applying all possible combinations of filters (HAB, GEO, and PHY) to the most robust model found in step 1. Then, the most robust of the 7 models is returned as output. There are many ways for seeking an optimal model. PaCo balances the tradeoff between computation speed and efficiency and nearly always arrives at the best or next-best model (Fig. 1).

We used MEVs to test how different combinations of filters (HAB, GEO, and PHY) and ecological parameters (max L, K, Y, Ym, and T) affect model outputs. Although MEV is the most direct measure of reliability of the model, it is informative only for hosts where there is a substantial list of known parasites. For this reason, we limited the validation by randomly selecting 50 fish species with at least 30 known parasite species. The resulting set included 17 freshwater species, 17 marine species, and 16 euryhaline species belonging to 23 different families (with the Salmonidae and the Cyprinidae being the most represented with 9 species each). For each of the 248 combinations of model variables, we generated lists of potential parasites for the 50 host species. Then, for each model of each target host species, we used the MEV. We considered all parasite groups in our analysis, but underscore that each parasite group could respond differently to the model setup (we hope that users will use PaCo to explore these differences).

The overall performance of PaCo algorithm is shown in Figure 2, where the average C-scores of reported (non-host specific) parasites for each of the 248 models for the 58 target hosts is plotted against that of the average C-scores of parasite species not reported from the respective target hosts. Although there was

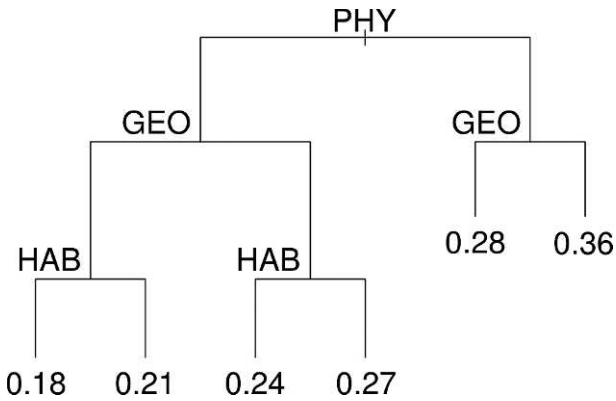

FIGURE 3. Regression tree of the presence–absence of a model filter–niche dimension (for each of the 248 possible combinations) vs. the corresponding average MEV (means were calculated for each combination of parameters on 58 host species known to harbor at least 30 parasite species). The rule is moved to the right when the filter–parameter is included. PHY: phylogeny; GEO: geography; HAB: habitat.

variation in the robustness of the models based on the differential inclusion of filters and ecological parameters, only a few C-scores of the reported parasites were less than the corresponding C-scores of unreported parasites (i.e., fell below the line of equality). This demonstrates that the PaCo algorithm tends to attribute higher C-scores to reported parasites, suggesting that the included filters and parameters are relevant in determining the distribution of parasites. However, because this does not provide definitive evidence of a model's detection power, we considered the MEV of the best models for the 50 hosts. On average, the MEV for the best model was 0.43, which is almost twice the mean MEV value for any random combination of filters and parameters (0.27), indicating that users can do better than a random model.

Some host characteristics were more useful than others in proposing potential parasites. Among filters, phylogeny (PHY) and geography (GEO) were the most effective, i.e., present in 88% of the best models, while habitat (HAB) (68%) was less useful. The niche dimension most common in the best models was trophic level (T, 41%), while life span (Y, 34%), growth rate (K, 32%), maximum length (L, 0.28%), and age at maturity (Ym, 20%) were less relevant. A regression tree of model filter–parameter combinations vs. corresponding MEVs (Fig. 3) provides a clearer picture of the patterns described above, highlighting the benefit of considering phylogeny, geography, and habitat when proposing parasites for a target host.

PaCo is the first method available that systematically generates proposed lists of parasites (in order of confidence) for fish hosts. Lists of proposed parasites are meant as a guide for what parasites researchers should investigate. Again, they are by no means a full set of possibilities nor are they intended to substitute for actual parasite sampling and identification, which should be performed according to proper procedures (see, for example, Strona et al., 2009a).

In addition, users should note that the MEV provided by PaCo for a target host relies on the available parasitological information for that host. As a consequence, MEV cannot be computed for hosts that have not been sampled for parasites. Currently, only 12% of the fish in the internal database have parasites reported from them, but our hope is that users will contribute data to Fish PEST to help increase the number of verified host–parasite records that PaCo can use.

The different analyses we performed consistently indicated a strong positive effect of phylogeny, offering insight into the study of host–parasite coevolution which, until now, has been approached primarily by means of molecular ecology (see, for example, Hafner and Nadler, 1988; Boeger and Kritsky, 1989; Barker, 1994; Desdevises et al., 2002). Thus, we suggest that users of PaCo include PHY and GEO in their models and critically consider whether to exclude other parameters (especially age at maturity).

Given these general guidelines, we emphasize that one of PaCo's most important features is its flexibility, i.e., users can select the combination of filters and parameters they consider most appropriate for their specific needs. At the moment, the internal database includes only helminth species. However, PaCo can potentially indicate a potential list of parasites for any group with enough records. We hope that, in addition to helminths, other parasite groups (such as the Crustacea and Protozoa) will be made available to PaCo users in the future.

ACKNOWLEDGMENTS

Any use of trade, product, or firm names in this publication is for descriptive purposes only and does not imply endorsement by the U.S. government. The authors would like to thank Rainer Froese and Nicolas Bailly for their suggestions and support with FishBase data.

LITERATURE CITED

APPELTANS, W., P. BOUCHET, G. A. BOXSHALL, K. FAUCHALD, D. P. GORDON, B. W. HOEKSEMA, G. C. B. POORE, R. W. M. VAN SOEST, S. STÖHR, T. C. WALTER ET AL. 2011. World Register of Marine Species. Available at: www.marinespecies.org. Accessed 15 February 2012.

BARKER, S. C. 1994. Phylogeny and classification, origins, and evolution of host associations of lice. International Journal for Parasitology **24:** 1285–1291.

BISBY, F., Y. ROSKOV, A. CULHAM, T. ORRELL, D. NICOLSON, L. PAGLINAWAN, N. BAILLY, W. APPELTANS, P. KIRK, T. BOURGOIN ET AL. 2012. Species 2000 & ITIS Catalogue of Life. Available at: www.catalogueoflife.org/col/. Accessed 15 February 2012.

BOEGER, W. A., AND D. C. KRITSKY. 1989. Phylogeny, coevolution, and revision of the Hexabothriidae Price, 1942 (Monogenea). International Journal for Parasitology **19:** 425–440.

COHEN, S., AND A. KOHN. 2008. South American Monogenea—List of species, hosts and geographical distribution from 1997 to 2008. Zootaxa **1924:** 1–42.

DESDEVISES, Y., S. MORAND, O. JOUSSON, AND P. LEGENDRE. 2002. Coevolution between *Lamellodiscus* (Monogenea: Diplectanidae) and Sparidae (Teleostei): The study of a complex host-parasite system. Evolution **56:** 2459–2471.

FROESE, R., AND D. PAULY. 2000. Fishbase 2000: Concepts, design and data sources. R. Froese and D. Pauly (eds). ICLARM, Los Baños, Laguna, Philippines. 344 p.

———, AND ———. 2012. FishBase. World Wide Web electronic publication. Available at: www.fishbase.org. Accessed 15 February 2012.

GIBSON, D. I., R. A. BRAY, AND E. A. HARRIS. 2005. Host-parasite database of the Natural History Museum, London. Available at: www.nhm.ac.uk/. Accessed 15 February 2012.

HAFNER, M. S., AND S. A. NADLER. 1988. Phylogenetic trees support the coevolution of parasites and their hosts. Nature **332:** 258–259.

HARRIS, P. D., A. P. SHINN, J. CABLE, T. A. BAKKE, AND J. E. BRON. 2008. GyroDb: Gyrodactylid monogeneans on the web. Trends in Parasitology **24:** 109–111.

HEWITT, G. C., AND P. M. HINE. 1972. Checklist of parasites of New Zealand fishes and of their hosts. New Zealand Journal of Marine and Freshwater Research **6:** 69–114.

Holland, C. V., and C. R. Kennedy. 1997. A checklist of parasitic helminth and crustacean species recorded in freshwater fish from Ireland. Biology and Environment: Proceedings of the Royal Irish Academy **3:** 225–243.

Holovaty, A., and J. Kaplan-Moss. 2009. The definitive guide to Django: Web development done right. Apress, New York, 361 p.

Kohn, A., and S. Cohen. 1998. South American Monogenea—List of species, hosts and geographical distribution. International Journal for Parasitology **28:** 1517–1554.

———, ———, and G. Salgado-Maldonado. 2006. Checklist of Monogenea parasites of freshwater and marine fishes, amphibians and reptiles from Mexico, Central America and Caribbean. Zootaxa **1289:** 1–114.

Lichtenfels, R., E. P. Hoberg, and P. A. Pilitt. 2011. U.S. National Parasite Collection. U.S. Department of Agriculture, Agricultural Research Service, Biosystematics and National Parasite Collection Unit, Beltsville, Maryland. Available at www.anri.barc.usda.gov/bnpcu/parasrch.asp. Accessed 15 February 2012.

Poulin, R. 1994. Meta-analysis of parasite-induced behavioural changes. Animal Behaviour **48:** 137–146.

Rohde, K., C. Hayward, and M. Heap. 1995. Aspects of the ecology of metazoan ectoparasites of marine fishes. International Journal for Parasitology **25:** 945–970.

Salgado-Maldonado, G. 2006. Checklist of helminth parasites of freshwater fishes from Mexico. Zootaxa **1324:** 1–357.

———. 2008. Helminth parasites of freshwater fish from Central America. Zootaxa **1915:** 29–53.

Strona, G. and K. D. Lafferty. 2012a. How to catch a parasite: Parasite Niche Modeler (PaNic) meets FishBase. Ecography **35:** 1–6.

———, and ———. 2012b. Fish PEST: An innovative software suite for fish parasitologists. Trends in Parasitology doi:10.1016/j.pt.2012.02.001.

———, F. Stefani, and P. Galli. 2009a. Field preservation of monogenean parasites for molecular and morphological analyses. Parasitology International **58:** 51–54.

———, ———, and ———. 2009b. Monogenoidean parasites of Italian marine fish: An updated checklist. Italian Journal of Zoology **77:** 419–437.

van Rossum, G., and J. de Boer. 1991. Interactively testing remote servers using the Python programming language. CWI Quarterly **4:** 283–303.

Williams Jr., E. H., and L. Bunkley-Williams. 1996. Parasites of offshore big game fishes of Puerto Rico and the western Atlantic. Puerto Rico Department of Natural and Environmental Resources and the University of Puerto Rico, San Juan, Puerto Rico, 383 p.