International Journal for Parasitology 49 (2019) 429–435

Contents lists available at ScienceDirect

International Journal for Parasitology

journal homepage: www.elsevier.com/locate/ijpara

# Taxonomic and geographic bias in the genetic study of helminth parasites

Robert Poulin *, Eleanor Hay, Fátima Jorge

*Department of Zoology, University of Otago, P.O. Box 56, Dunedin 9054, New Zealand*

ARTICLE INFO

*Article history:*
Received 8 November 2018
Received in revised form 19 December 2018
Accepted 19 December 2018
Available online 21 February 2019

*Keywords:*
Helminths
DNA sequences
GenBank
Natural History Museum London
Parasite biodiversity
Species richness

ABSTRACT

The use of genetic information is now fundamental in parasite taxonomy and systematics, for resolving parasite phylogenies, discovering cryptic species, and elucidating patterns of gene flow among parasite populations. The accumulation of available gene sequences per geographical area or per parasite taxonomic group is likely proportional to species richness, but not without some biases. Certain areas and certain taxonomic groups receive more research effort than others, possibly causing a deficit in the relative number of parasite species being characterized genetically in some areas or taxonomic groups. Here, we use data on the number of parasite records per country or helminth family from the London Natural History Museum host-parasite database, and matching data on the number of gene sequences available from the National Center for Biotechnology Information (NCBI) GenBank database, to determine how available gene sequences scale with species richness across countries or parasitic helminth families. Our quantitative analysis identified countries/regions of the world and helminth families that have received the most effort in genetic research. More importantly, it allowed us to generate lists (based on residuals from the statistical model) of the 20 countries/regions and the 20 helminth families with the largest deficit in available gene sequences relative to their helminth species richness. We propose these lists as useful guides toward future allocation of effort to maximise advances in parasite biodiscovery, systematics and population structure.

## 1. Introduction

In the last few decades, as the cost of gene sequencing has dropped dramatically, the use of genetic data in taxonomy and systematics has become de rigueur (Meier, 2008; Olson et al., 2016). This trend fits well with the principles of integrative taxonomy, which calls for the use of multiple and complementary sources of data for species characterization and delimitation (Dayrat, 2005; Padial et al., 2010). In parasite taxonomy, the combination of traditional morphological description and use of DNA sequences has become part of the accepted best practice to characterize species (Caira, 2011; Perkins et al., 2011; Blasco-Costa et al., 2016; Poulin and Presswell, 2016). The growing availability of parasite DNA sequences has also proven essential for progress in resolving parasite phylogenies (eg. Olson and Tkach, 2005; Caira et al., 2014), for uncovering cryptic parasite diversity (Nadler and Pérez-Ponce de León, 2011; Poulin, 2011; Pérez-Ponce de León and Poulin, 2018), and for elucidating patterns of connectivity and gene flow among populations (Criscione et al., 2005), all of which are crucial elements toward a full understanding of host-parasite coevolution (Hay et al., 2018; Jorge et al., 2018).

To maximize the efficiency of research effort toward the genetic characterization of parasites and improve broad-scale resolution of phylogenetic and biogeographic history, genetic data should be obtained from all major taxonomic groups or geographical areas. In addition, the extent of genetic information available should be directly proportional to the species richness of either spatial or taxonomic units, and not biased toward certain groups or regions. A positive correlation between the number of species and the number of sequences obtained across taxonomic groups or geographical areas should arise if sequencing effort is evenly, or even randomly, allocated relative to species richness. All else being equal, with a roughly uniform research effort, the more species in a parasite family or in a region, the more likely some of them have been characterized genetically.

However, research effort is unlikely to be allocated uniformly across taxa or regions. The availability of funding and facilities for genetic research, and the number of qualified researchers, vary from country to country; also, researchers invest disproportionate efforts toward the study of the regional biota near their home institution (Martin et al., 2012; Amano and Sutherland, 2013). This

* Corresponding author.
*E-mail address:* robert.poulin@otago.ac.nz (R. Poulin).

https://doi.org/10.1016/j.ijpara.2018.12.005
0020-7519/© 2019 Australian Society for Parasitology. Published by Elsevier Ltd. All rights reserved.

should cause a deficit in the relative number of parasite species being characterized genetically in some regions. Also, for various reasons, certain parasite taxa receive much more attention than others from taxonomists and systematists (Poulin, 2002). For example, taxa of medical, zoonotic, veterinary or conservation concern are generally better studied than others. Although these biases are generally understood and accepted, they have not been quantified to identify regions or taxa where greater efforts are needed. Specifically, where are the parasite faunas in greatest need of genetic research based on their species richness? And which parasite higher taxa are the most understudied genetically in relation to their diversity?

Here, we answer those questions and provide a clear guide for future efforts toward the genetic characterization of parasite species. Our focus is on endoparasitic helminths (trematodes, cestodes, nematodes, and acanthocephalans) of vertebrate hosts, although we expect the patterns we uncover to apply broadly to other groups of parasites. We tackle the above questions by combining information from two independent databases, to test the expected relationship between the number of known species and the number of available DNA sequences (per country or per helminth family), separately for different genetic markers. Our statistical models allow us to identify the countries and helminth families for which there is either an excess or a clear deficit in available gene sequences given the species richness of that country or family. These lists of best and worst studied regions or taxa can serve as a clear road map for future genetic research on parasites, accelerating progress toward large-scale understanding of diversification patterns, cryptic diversity, phylogeny and biogeography of parasites.

## 2. Methods

### 2.1. Data compilation

The two key variables we examine were obtained from two databases that were assembled completely independently, for different purposes and by different organizations; therefore, any relationship we find between them is unlikely to be spurious. First, data on the number of parasite species records by taxonomic group (family or superfamily) and by location were compiled from the global database of host–helminth parasite occurrence records maintained by the London Natural History Museum (NHM, UK) (Gibson, D.I., Bray, R.A., Harris, E.A. (Compilers), 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). The database includes over 250,000 host-helminth records from nearly 30,000 published peer-reviewed articles. The distribution of these records across taxonomic groups and geographic regions provides a reliable picture of the biogeography of parasite diversity emerging from published research activities. Each entry in the NHM database corresponds to a host-parasite species combination, and therefore not necessarily to a distinct parasite species; however, we did not count duplicate records of a particular parasite species in a region or family, i.e. parasite species were only counted once. In the majority of cases, multiple records in the NHM database of helminth species *X* on host species *Y* from locality *Z* come from multiple publications by the same research group; therefore, all reports of parasite *X* on host *Y*, other than the first one, do not represent new discoveries, and including all duplicate records would wrongly inflate the data. We consider the number of records we used to be a good proxy for the known number of parasite species in a family or region.

Our search was restricted to acanthocephalan, cestode, trematode and nematode parasites. Taxonomic family names (or superfamily names in the case of nematodes, due to NHM taxonomic subgroup assignment) for each parasite group were collected by directly accessing the NHM host-parasite database online (Gibson et al., 2005, cited earlier). Data collection was then performed in the R statistical computing environment (R Core Team, 2017). For each parasite family, we obtained and counted the number of unique records available in the NHM host-parasite database using the R package helminthR (Dallas, 2016) with the function *findParasite* specifying the family (or superfamily) as subgroup (see Supplementary Data S1 for the script used). Similarly, a list of geographical regions was created by accessing the data object 'locations' (using the command data (locations)) of the helminthR package. A query was performed for each location using the function *findParasite*, specifying each location and recording the number of unique records available in the NHM database for each parasite group (see Supplementary Data S1 for the script used). In most instances, regions corresponded to countries or recognized geopolitical areas. In some cases, however, records for geographical locations that were listed separately in the NHM database were pooled in order to organize data according to political borders (e. g, Spain is the sum of the NHM location list 'Spain+Andalusia', 'Balearic Islands', 'Gibraltar' and 'Canary Islands'). Finally, large bodies of water, i.e. seas and oceans, were treated as separate regions when the records identified these water bodies as the sampling location. The areas of the different regions/countries/oceans vary greatly, which affects how many parasite records are available from each of them (see Supplementary Fig. S1); however, since we used the number of records per region as our main predictor and not as a response variable (see below), these differences do not matter.

Second, to determine the number of parasite genetic sequences available by taxonomic group (family or superfamily) and by location, we searched the USA National Center for Biotechnology Information (NCBI) GenBank database. We used the R package *rentrez* (Winter, 2017) to obtain the number of entries for each family (or superfamily, in the case of nematodes) and location for each parasite group. We accepted the family name given in the databases, whether or not they reflect the latest taxonomic schemes; a subsequent check of all names against the Catalogue of Life (http://www.catalogueoflife.org/) indicates that most listed family names are accepted as valid (although some may represent a different level, e.g. subfamily, in the latest taxonomic scheme). The GenBank search was done separately for the barcode gene cytochrome oxidase subunit I (COI), the 18S ribosomal RNA (18S), the 28S ribosomal RNA (28S), and for all available nucleotide sequences, using the function *entrez_search*, and specifying the family name or location in the search term. Since we were interested in quantifying the overall effort in genetic research allocated to particular taxa or regions, all sequences from a particular parasite taxon were included and counted, even if many were for the same parasite species.

Given differences between the NHM database and GenBank in the taxonomy used (especially for nematodes), each case involving a large discrepancy between the two databases was verified by searching GenBank using names of lower taxonomic levels, e.g. genus. Specifically, this was done in the case of a family (or superfamily) with >20 NHM records but apparently zero GenBank sequences. Cases of 'false' zeros in GenBank were excluded from all analyses. Other inaccuracies due to the different taxonomic schemes may remain, but they would be difficult to detect and probably of little influence in our analyses. The full datasets arranged by regions and families are available as Supplementary Data S2 and S3.

### 2.2. Data analysis

Data on numbers of parasite records per region or family, and on the number of sequences per region or family, were all right-skewed, and had to be log $(x + 1)$ transformed prior to analysis to normalize their distribution. In all the analyses described below, we excluded double zeros, i.e. cases where there were no NHM records and no GenBank entries for a particular region or family. All analyses were conducted in JMP version 11.0 (SAS Institute Inc., Cary, NC, USA).

To test whether recorded parasite species richness in one higher helminth taxon covaried spatially with that in the other taxa, we computed all pairwise Pearson's correlation coefficients between the number of records for different taxa across countries/regions. Similarly, to test whether the effort in genetic research on different higher taxa covaried spatially, we computed pairwise correlations between the number of sequences for different taxa across countries/regions.

For our main analyses, we used mixed-effects models. Our main goal was to determine how the number of gene sequences available scales with the number of parasite records (a proxy for species richness) across different regions and among different helminth families, and to identify regions and families with disproportionately more or fewer sequences available relative to their richness. First, we analysed four models (one for COI, one for 18S, one for 28S, and one for all sequences) of the relationship between the number of sequences (response variable) and the number of NHM records (predictor) across countries/regions. The higher taxon (trematode, cestode, nematode, or acanthocephalan) to which parasites belonged was also included as a predictor, since the counts for each country/region were entered separately for each higher taxon. The continent or ocean to which a country/region belonged was included as a random factor, to account for uneven research resources across the world and other possible differences among continents, as well as to minimize the effect of spatial autocorrelation in the data.

Second, we ran four models (one for COI, one for 18S, one for 28S, and one for all sequences) of the relationship between the number of sequences (response variable) and the number of NHM records (predictor) across helminth families (superfamilies in the case of nematodes). The higher taxon (trematode, cestode, nematode, or acanthocephalan) to which a family belonged was included as a random factor, to account for phylogenetic non-independence among families.

Finally, to identify the countries/regions and families that received the most and the least attention with respect to genetic research on parasites, we obtained the residuals from the above models, i.e. the difference between the observed number of sequences in GenBank and the number predicted by our model based on the number of NHM entries. For calculating the residuals only, in the models of the relationship between the number of sequences and the number of records across countries/regions, we pooled all parasites by country/region instead of treating each higher taxon separately. Since we are interested in the overall effort in genetic research and different genes are often used to study different helminth taxa, we used the residuals from the models considering the number of all available gene sequences combined, and not those for particular genes such as the COI or 28S. Large positive residuals indicate more available gene sequences than expected based on the number of records, whereas large negative residuals indicate a deficit in gene sequences based on the number of records. A global map was generated to visualize differences in genetic research effort among countries, based on residual values, using the function *joinCountryData2Map* from the R package rworldmap (South, 2011); the few mismatches between our list of countries/regions and the package's list of countries were resolved by assigning the residual value of the larger geopolitical country to the region (e.g. Scotland assigned same value as England+Wales). Finally, when compiling lists of the highest and lowest residual values, we only considered countries/regions or families with at least 20 records in the NHM database, because residuals for countries/regions or families that are either species-poor or rarely encountered can be subject to error.

## 3. Results

In total, our analyses involved 51,344 records from the NHM host-parasite database, most of them representing nematodes and trematodes, and 2,779,536 sequences from the NCBI GenBank database (Table 1). Of the three specific genes considered, there were more COI sequences available for trematodes (mostly because in many cases, there are multiple COI sequences available per species), and more 18S and 28S sequences for nematodes (Table 1). We related the number of records to the number of sequences among 296 distinct helminth families (or superfamilies for nematodes), and across 190 countries/regions, although the exact numbers included in different models are lower than those maximum values (i.e., after exclusion of regions or families with zero values for a particular gene).

The number of records for any given higher helminth taxon covaried significantly across regions with that from the other taxa (all $r > 0.75$, $P < 0.001$) (see Supplementary Table S1). In other words, if there were many records of nematodes in one region, there were generally also many records of trematodes, cestodes and acanthocephalans in that region. Similarly, the number of sequences in GenBank for any given higher helminth taxon also covaried significantly across regions with that from the other taxa (all $r > 0.48$, $P < 0.001$) (see Supplementary Table S2).

Our analyses show that the number of gene sequences reported for helminths per country or region is strongly related to the number of NHM records for that country/region, whether considering specific genes (COI, 18S or 28S) or all sequences available (Table 2, Fig. 1). Independently of the number of NHM records, the higher taxon to which helminths belong also influenced the number of sequences in GenBank, but the effect was not nearly as strong. Approximately 40–44% of the variance not explained by the main predictors could be attributed to the broader geographical area (continent or ocean) to which a country/region belonged (Table 2). Despite the number of helminth sequences in GenBank being strongly related to the number of NHM records, some countries/regions have disproportionately more or fewer sequences available than expected based on their reported helminth species richness (Figs. 1 and 2). In general, most developing countries tend to show a low genetic research effort relative to their helminth richness (Fig. 2). The largest negative and positive residuals from the model (combining all helminth taxa; Fig. 1) provide lists of the countries/regions that have received either disproportionately low or disproportionately high effort in genetic research on helminth parasites (Table 3).

Similarly, the analyses indicate that the number of gene sequences reported per helminth family is strongly related to the number of NHM records for that family, whether for specific genes (COI, 18S or 28S) or all sequences available (Table 4, Fig. 3). Only about 7–16% of the variance not explained by the number of NHM records could be attributed to the higher taxon to which a family belonged (Table 4). As in the above analysis, despite the strong effect of the number of NHM records, some families have many more or fewer sequences available than expected based on their reported species richness (Fig. 3). The largest negative and positive residuals from the model for all sequences combined (Fig. 3) provide lists of families that have received either

**Table 1**
Summary of the dataset analysed, showing for each helminth higher taxon the number of unique records (i.e. species per country combinations) in the Natural History Museum (NHM, London, UK) database, the number of sequences for particular genes (cytochrome oxidase subunit I (COI), 18S, 28S) and the number of all sequences in GenBank, the number of countries/regions (out of 190) in which species from that taxon were recorded, and the number of families (or superfamilies for nematodes) per taxon included in our analyses.

| Taxon | No. records | COI | 18S | 28S | All sequences | No. regions | No. families |
|---|---|---|---|---|---|---|---|
| Trematodes | 17,354 | 7357 | 4524 | 9994 | 730,588 | 167 | 147 |
| Cestodes | 10,634 | 2154 | 2759 | 3922 | 226,205 | 165 | 79 |
| Nematodes | 21,328 | 5272 | 11,902 | 13,997 | 1,819,326 | 179 | 46 |
| Acanthocephalans | 2028 | 980 | 310 | 514 | 3417 | 116 | 24 |

**Table 2**
Summary of the mixed models exploring variation in the number of GenBank sequences of helminths (response variable) among countries/regions, showing the significance tests for each fixed effect and the percentage of the remaining variance accounted for by the random factor. The number of data points (countries/regions X higher taxon) included in each analysis excludes double zeros and is shown in parentheses.

| Gene marker | Fixed effect | df | *F* ratio | *P* value | Random factor | % variance |
|---|---|---|---|---|---|---|
| COI ($n = 630$) | No. NHM records | 1, 624.9 | 200.62 | <0.0001 | Continent/ocean | 43.7 |
| | Higher taxon | 3, 613.9 | 2.58 | 0.0529 | | |
| 18S ($n = 633$) | No. NHM records | 1, 627.7 | 286.26 | <0.0001 | Continent/ocean | 41.7 |
| | Higher taxon | 3, 616.3 | 3.98 | 0.0080 | | |
| 28S ($n = 633$) | No. NHM records | 1, 627.4 | 307.13 | <0.0001 | Continent/ocean | 40.6 |
| | Higher taxon | 3, 616.4 | 8.12 | <0.0001 | | |
| All sequences ($n = 643$) | No. NHM records | 1, 637.8 | 313.40 | <0.0001 | Continent/ocean | 42.6 |
| | Higher taxon | 3, 626.8 | 6.65 | 0.0002 | | |

COI, cytochrome oxidase subunit I; NHM, Natural History Museum (London, UK).

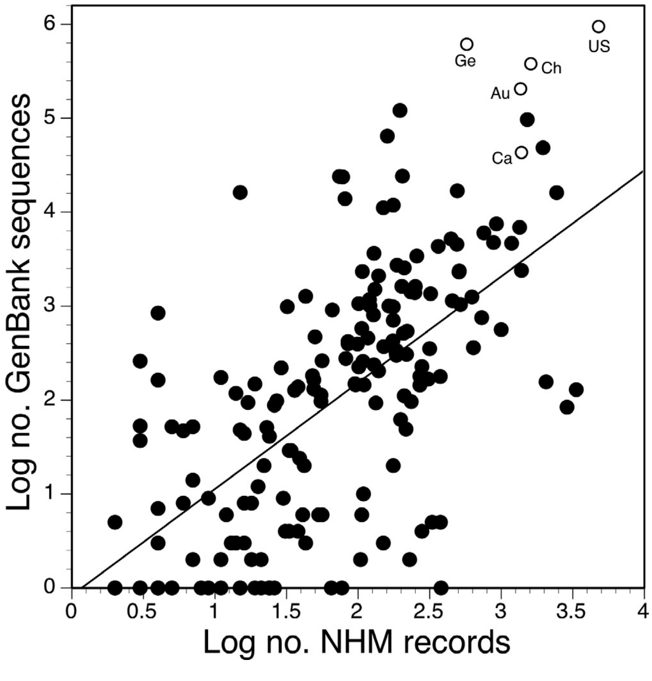

**Fig. 1.** Total number of helminth sequences in GenBank as a function of the number of unique species records in the Natural History Museum (NHM, London, UK) host-parasite database, across different countries/regions. Data on trematodes, cestodes, nematodes and acanthocephalans are pooled together. Examples of countries with many more gene sequences available than expected based on their reported species richness are shown on the plot (open circles): Au, Australia; Ca, Canada; Ch, China; Ge, Germany; US, United States.

disproportionately low or disproportionately high effort in genetic research on helminth parasites (Table 5).

## 4. Discussion

The genetic characterization of parasite species is fundamental for any investigation of their taxonomy, phylogenetic history, cryptic diversity or population structure. Here, using entries in the Natural History Museum host-parasite database as a proxy for species richness, we confirm that more gene sequences are obtained and deposited in GenBank from geographical areas or taxonomic families containing more species. However, for some countries/regions or helminth families, the number of sequences available is disproportionately lower or higher than what would be expected based on their species richness. These disparities are not surprising, as they reflect the priorities driving parasitological research. Our study does not only confirm this expected pattern, it goes one step further and identifies the countries/regions or helminth families with the greatest deficit in genetic research on their parasites, providing guidance for future research efforts.

We must begin with a word of caution about the data used here. The NHM host-parasite database (Gibson et al., 2005, cited earlier) provides a comprehensive list of parasite species recorded from various hosts and locations; it is the largest such database available at the moment. It can serve as a proxy for species richness per country/region, or per helminth family; the NHM database has indeed been used that way in recent analyses (e.g., Dallas et al., 2018). However, it also reflects study effort (i.e. there will be more records from a region if more research is conducted in that region). It is not possible to disentangle actual species richness from study effort in the NHM data. This is not really a problem, because in the absence of biases, we would expect genetic characterization of parasites to be proportional to both parasite species richness and basic parasitological research effort, and the NHM database captures both. A different problem arises when one tries to integrate data from the NHM database with data from GenBank, the largest publicly available depository of gene sequences. The two databases use different taxonomic schemes, creating multiple mismatches when searching for particular families or other taxonomic groups. Others have highlighted similar issues recently regarding the need to standardize taxonomic and geographic data to make full use of the GenBank database (Tahsin et al., 2017; Porter and Hajibabaei, 2018). We have followed a conservative approach, using a simple rule-of-thumb to exclude obvious cases where entries in the NHM database wrongly appeared to have no matches in GenBank; small errors may remain, although these are unlikely to affect our

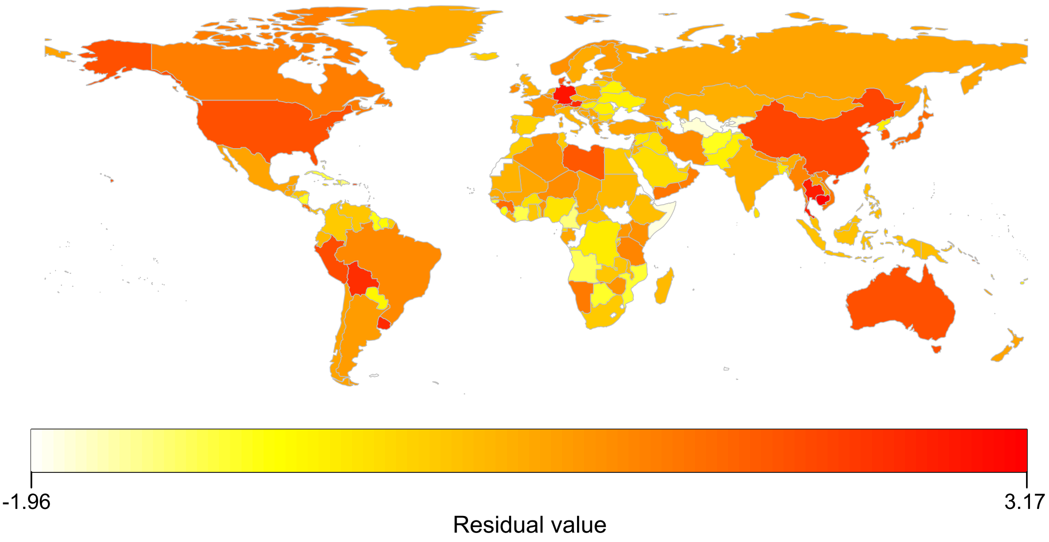

**Fig. 2.** Global map showing variation in the effort put into genetic research on helminth parasites relative to their species richness across countries of the world. Relative effort in genetic effort, ranging from low (light yellow) to high (dark red), is based on the residual values from the model relating the number of all GenBank sequences to the number of entries in the Natural History Museum (NHM, London, UK) database (shown in Fig. 1). Data for oceanic regions and Antarctica are not shown.

**Table 3**
List of the 20 countries/regions (with a minimum of 20 Natural History Museum (NHM, London, UK) entries) that have received the most effort in genetic research on helminth parasites relative to their species richness, and the 20 countries/regions that have received the least effort. Residuals are from our model relating the total number of sequences in GenBank to the number of entries in the NHM database per country/region.

| Lowest effort in genetic research | | Highest effort in genetic research | |
|---|---|---|---|
| Country/region | Residual | Country/region | Residual |
| Antilles | −2.719 | Germany | 2.853 |
| Antarctic Ocean | −2.155 | Thailand | 2.706 |
| Arctic Ocean | −2.015 | Austria | 2.539 |
| Tadzhikistan | −1.956 | Uruguay | 2.509 |
| Uzbekistan | −1.945 | Bolivia | 2.477 |
| Lesser Antilles | −1.890 | Puerto Rico | 2.226 |
| Atlantic Ocean | −1.847 | China | 2.109 |
| Somalia | −1.802 | Peru | 1.986 |
| Turkmenistan | −1.759 | USA | 1.940 |
| Kyrgyzstan | −1.745 | Australia | 1.928 |
| Pacific Ocean | −1.738 | Denmark | 1.809 |
| Indian Ocean | −1.403 | South Korea | 1.754 |
| Galapagos Islands | −1.325 | Guinea | 1.558 |
| Faroes | −1.284 | Japan | 1.545 |
| Dominican Republic | −1.284 | Trinidad & Tobago | 1.516 |
| Kerguelen Island | −1.284 | Costa Rica | 1.402 |
| Cameroon | −1.278 | Vietnam | 1.371 |
| South Shetlands | −1.215 | Namibia | 1.306 |
| Angola | −1.110 | Canada | 1.243 |
| Ivory Coast | −1.070 | Myanmar | 1.149 |

findings. In future, it would be useful for all databases to adopt a standard taxonomic structure, to make their coupling straightforward.

In our analysis of geographical patterns, we found that the number of NHM records per country/region for one higher taxon (trematodes, cestodes, nematodes or acanthocephalans) of helminths correlated strongly with the number of records for a different taxon, across all countries and regions. This is in sharp contrast to a similar spatial analysis using a different data set: the geographical coordinates where new helminth species were found and described in the last few decades (Poulin and Jorge, 2018). This earlier analysis indicated that there is no spatial congruence between higher helminth taxa with respect to areas where new species are found (Poulin and Jorge, 2018). It seems that original species discoveries and subsequent reports of species occurrence follow different patterns of spatial covariation among higher taxa.

More importantly, our analysis of geographical patterns allowed us to identify countries/regions with unusually large deviations from the number of available gene sequences expected from our statistical model (Table 3). Not surprisingly, countries with an excess of available sequences based on their helminth species richness include research powerhouses such as the United States, Canada, China, Japan, Germany and Australia. The most useful list for future research, however, is that of countries/regions with a deficit in available gene sequences relative to their helminth species richness. In particular, this list includes countries from Africa and central Asia, small island countries, and oceanic regions. Allocating greater resources and effort to genetic research on parasites from those areas would seem the best way to advance our understanding of parasite biodiversity. For instance, the discovery of cryptic species increases with greater sequencing effort, both in metazoans in general (Poulin and Pérez-Ponce de León, 2017) and in helminth parasites specifically (Poulin, 2011; Pérez-Ponce

**Table 4**
Summary of the mixed models exploring variation in the number of GenBank sequences (response variable) among helminth families, showing the significance tests for the fixed effect and the percentage of the remaining variance accounted for by the random factor. The number of families (superfamilies for nematodes) included in each analysis excludes double zeros and is shown in parentheses.

| Gene marker | Fixed effect | df | *F* ratio | *P* value | Random factor | % variance |
|---|---|---|---|---|---|---|
| COI (*N* = 294) | No. NHM records | 1, 291.9 | 149.96 | <0.0001 | Higher taxon | 7.9 |
| 18S (*N* = 296) | No. NHM records | 1, 292.9 | 269.20 | <0.0001 | Higher taxon | 16.5 |
| 28S (*N* = 296) | No. NHM records | 1, 293.6 | 241.32 | <0.0001 | Higher taxon | 10.4 |
| All sequences (*N* = 296) | No. NHM records | 1, 293.8 | 242.14 | <0.0001 | Higher taxon | 8.7 |

COI, cytochrome oxidase subunit I; NHM, Natural History Museum (London, UK).

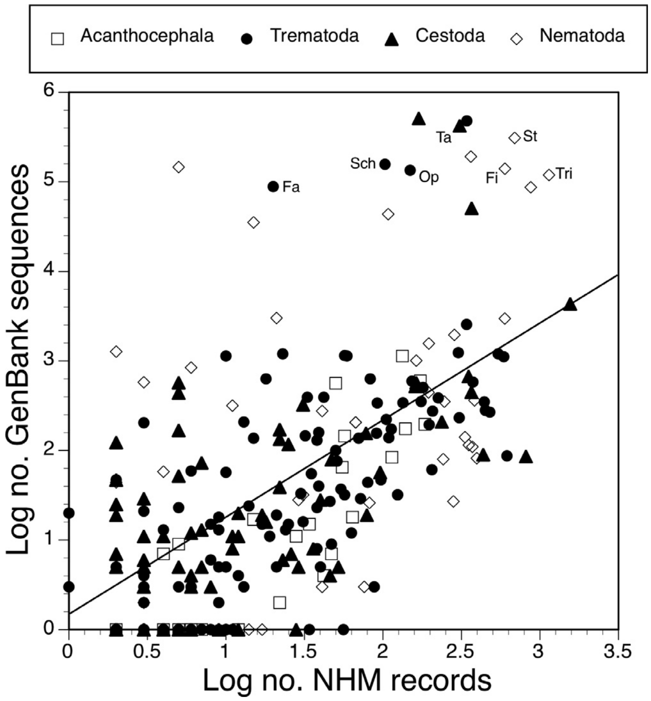

**Fig. 3.** Total number of sequences in GenBank as a function of the number of unique species records in the Natural History Museum (NHM, London, UK) host-parasite database, across different helminth families (superfamilies in the case of nematodes). Families of trematodes, cestodes, nematodes and acanthocephalans are indicated by different symbols. Examples of families with many more gene sequences available than expected based on their reported species richness are shown on the plot: Fa, Fasciolidae (trematodes); Fi, Filarioidea (nematodes); Op, Opisthorchiidae (trematodes); Sch, Schistosomatidae (trematodes); St, Strongyloidea (nematodes); Ta, Taeniidae (cestodes); Tri, Trichostrongyloidea (nematodes).

**Table 5**
List of the 20 helminth families (with a minimum of 20 Natural History Museum (NHM (London, UK) entries) that have received the most effort in genetic research on their parasites relative to their species richness, and the 20 families that have received the least effort. Residuals are from our model relating the total number of sequences in GenBank to the number of entries in the NHM database per family (or superfamily in the case of nematodes).

| Lowest effort in genetic research | | Highest effort in genetic research | |
|---|---|---|---|
| Family[a] | Residual | Family[a] | Residual |
| Cladorchidae (T) | −1.963 | Fasciolidae (T) | 3.503 |
| Pronocephalidae (T) | −1.714 | Diphyllobothriidae (C) | 3.187 |
| Proterodiplostomidae (T) | −1.711 | Schistosomatidae (T) | 2.924 |
| Subuluroidea (N) | −1.640 | Dicrocoeliidae (T) | 2.803 |
| Amphicotylidae (C) | −1.613 | Taeniidae (C) | 2.802 |
| Dilepididae (C) | −1.380 | Opisthorchiidae (T) | 2.672 |
| Acuarioidea (N) | −1.346 | Rhabditoidea (N) | 2.374 |
| Aproctoidea (N) | −1.328 | Mermithoidea (N) | 2.344 |
| Amabiliidae (C) | −1.261 | Strongyloidea (N) | 2.258 |
| Hemiuridae (T) | −1.235 | Aphelenchoidea (N) | 2.011 |
| Cephalobothriidae (C) | −1.226 | Filarioidea (N) | 1.987 |
| Arhythmacanthidae (A) | −1.216 | Anoplocephalidae (C) | 1.793 |
| Diplosentidae (A) | −1.190 | Trichostrongyloidea (N) | 1.590 |
| Opistholebetidae (T) | −1.094 | Ascaridoidea (N) | 1.587 |
| Davaineidae (C) | −1.044 | Cathaemasiidae (T) | 1.565 |
| Habronematoidea (N) | −1.037 | Clinostomidae (T) | 1.088 |
| Plagiorhynchidae (A) | −1.029 | Paragonimidae (T) | 1.066 |
| Telorchiidae (T) | −0.943 | Homalometridae (T) | 0.898 |
| Nematotaeniidae (C) | −0.932 | Pomphorhynchidae (A) | 0.846 |
| Brachycoeliidae (T) | −0.920 | Mesocestoididae (C) | 0.846 |

[a] A, acanthocephalans; C, cestodes; N, nematodes; T, trematodes.

de León and Poulin, 2018). We now have a list of those areas in greatest need of increased genetic research on parasites.

Similarly, our analysis of taxonomic biases in genetic research on helminths has also uncovered families with unusually large deviations from their expected number of available gene sequences based on their number of NHM records (Table 5). As expected, families with an excess of available sequences based on their helminth species richness show no particular phylogenetic affinities with each other, but mostly include those of medical or veterinary importance, e.g. the trematode families Schistosomatidae, Opisthorchiidae, Fasciolidae and Dicrocoeliidae, the cestode families Taeniidae and Diphyllobothriidae, and the nematode superfamilies Filarioidea, Strongyloidea and Ascaridoidea. In contrast, families with a deficit in available gene sequences relative to their species richness comprise mostly taxa parasitic in wildlife but of little veterinary concern. Yet these are the ones in greatest need of genetic characterization, from the perspective of biodiversity and phylogenetic research.

Parasite diversity is clearly not the only determinant of variation in genetic research effort across regions or parasite taxa. Our models revealed that the number of NHM records only explains a modest proportion of the variance in the number of GenBank sequences. There are many factors likely to influence genetic research effort. For example, across different countries, variation in population size, political stability, number of universities, GDP or research funding are all likely to affect the extent of genetic research on parasites. Future studies could explore the relative influence of these socio-economic factors.

Parasitological research in general suffers from taxonomic and geographical biases. For example, historically there have been more articles per year published on parasitic nematodes than on other helminths, even when taking into account the relative species richness of the different taxa (Poulin, 2002). Also, the effort allocated toward finding and describing new helminth parasite species shows a very poor geographic congruence with hotspots of host diversity, such that regions of the world with more vertebrates receive disproportionately lower parasite discovery effort than temperate regions (Jorge and Poulin, 2018). Therefore, the patterns we uncover here are not unique to genetic research on parasites, and they are likely to continue as genetic research scales up to whole-genome sequencing. We hope that the lists of understudied regions and helminth families we provide here, based on a quantitative analysis of available information, will be useful to future allocation of effort to maximise advances in parasite biodiscovery.

## Acknowledgements

This research was indirectly supported by an Otago University (New Zealand) Research Grant to RP. We thank the Natural History Museum (London, UK) for permission to access their database, Tad Dallas for invaluable assistance with the use of the helminthR package, and two anonymous reviewers for very useful comments on an earlier draft.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijpara.2018.12.005.

## References

Amano, T., Sutherland, W.J., 2013. Four barriers to the global understanding of biodiversity conservation: wealth, language, geographical location and security. Proc. R. Soc. B 280, 20122649.

Blasco-Costa, I., Cutmore, S.C., Miller, T.L., Nolan, M.J., 2016. Molecular approaches to trematode systematics: "best practice" and implications for future study. Syst. Parasitol. 93, 295–306.

Caira, J.N., 2011. Synergy advances parasite taxonomy and systematics: an example from elasmobranch tapeworms. Parasitology 138, 1675–1687.

Caira, J.N., Jensen, K., Waeschenbach, A., Olson, P.D., Littlewood, D.T., 2014. Orders out of chaos: molecular phylogenetics reveals the complexity of shark and stingray tapeworm relationships. Int. J. Parasitol. 44, 55–73.

Criscione, C.D., Poulin, R., Blouin, M.S., 2005. Molecular ecology of parasites: elucidating ecological and microevolutionary processes. Mol. Ecol. 14, 2247–2257.

Dallas, T., 2016. helminthR: an R interface to the London Natural History Museum's host-parasite database. Ecography 39, 391–393.

Dallas, T.A., Aguirre, A.A., Budischak, S., Carlson, C., Ezenwa, V., Han, B., Huang, S., Stephens, P.R., 2018. Gauging support for macroecological patterns in helminth parasites. Glob. Ecol. Biogeogr. 27, 1437–1447. https://doi.org/10.1111/geb.12819.

Dayrat, B., 2005. Towards integrative taxonomy. Biol. J. Linn. Soc. 85, 407–415.

Hay, E., Jorge, F., Poulin, R., 2018. The comparative phylogeography of shore crabs and their acanthocephalan parasites. Mar. Biol. 165, 69.

Jorge, F., Poulin, R., 2018. Poor geographical match between the distributions of host diversity and parasite discovery effort. Proc. R. Soc. B 285, 20180072.

Jorge, F., Perera, A., Poulin, R., Roca, V., Carretero, M.A., 2018. Getting there and around: host range oscillations during colonization of the Canary Islands by the parasitic nematode *Spauligodon*. Mol. Ecol. 27, 533–549.

Martin, L.J., Blossey, B., Ellis, E., 2012. Mapping where ecologists work: biases in the global distribution of terrestrial ecological observations. Front. Ecol. Environ. 10, 195–201.

Meier, R., 2008. DNA sequences in taxonomy: opportunities and challenges. In: Wheeler, Q.D. (Ed.), The New Taxonomy. Systematics Association. CRC Press, Boca Raton, Florida, pp. 95–127.

Nadler, S.A., Pérez-Ponce de León, G., 2011. Integrating molecular and morphological approaches for characterizing parasite cryptic species: implications for parasitology. Parasitology 138, 1688–1709.

Olson, P.D., Hughes, J., Cotton, J.A. (Eds.), 2016. Next Generation Systematics. Systematics Association. Cambridge University Press, Cambridge.

Olson, P.D., Tkach, V.V., 2005. Advances and trends in the molecular systematics of the parasitic Platyhelminthes. Adv. Parasitol. 60, 165–243.

Padial, J.M., Miralles, A., De la Riva, I., Vences, M., 2010. The integrative future of taxonomy. Front. Zool. 7, 16.

Pérez-Ponce de León, G., Poulin, R., 2018. An updated look at the uneven distribution of cryptic diversity among parasitic helminths. J. Helminthol. 92, 197–202.

Perkins, S.L., Martinsen, E.S., Falk, B.G., 2011. Do molecules matter more than morphology? Promises and pitfalls in parasites. Parasitology 138, 1664–1674.

Porter, T.M., Hajibabaei, M., 2018. Over 2.5 million COI sequences in GenBank and growing. PLoS One 13, e0200177.

Poulin, R., 2002. Qualitative and quantitative aspects of recent research on helminth parasites. J. Helminthol. 76, 373–376.

Poulin, R., 2011. Uneven distribution of cryptic diversity among higher taxa of parasitic worms. Biol. Lett. 7, 241–244.

Poulin, R., Jorge, F., 2018. The geography of parasite discovery across taxa and over time. Parasitology 146, 168–175. https://doi.org/10.1017/S003118201800118X.

Poulin, R., Pérez-Ponce de León, G., 2017. Global analysis reveals that cryptic diversity is linked with habitat but not mode of life. J. Evol. Biol. 30, 641–649.

Poulin, R., Presswell, B., 2016. Taxonomic quality of species descriptions varies over time and with the number of authors, but unevenly among parasite taxa. Syst. Biol. 65, 1107–1116.

R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/.

South, A., 2011. rworldmap: a new R package for mapping global data. R J. 3, 35–43.

Tahsin, T., Weissenbacher, D., Jones-Shargani, D., Magee, D., Vaiente, M., Gonzalez, G., Scotch, M., 2017. Named entity linking of geospatial and host metadata in GenBank for advancing biomedical research. Database 2017, bax093.

Winter, D.J., 2017. rentrez: an R package for the NCBI eUtils API. R J. 9, 520–526.