SUPPLEMENTARY MATERIAL

Poor geographical match between the distributions of host diversity and parasite discovery effort

Fátima Jorge and Robert Poulin

1. Extended methods

(a) Host and parasite data

We compiled data on the spatial distribution of parasite species discovery from species description records by conducting a detailed search on the ISI Web of Science™ for the period of 1970-2017, as numbers of parasite species discovered and described annually have been higher in the past 50 years than ever before [1]. The search was restricted to acanthocephalan, cestode, trematode and nematodes parasites of vertebrates. Specifically the search keywords used were: (("new species" *OR* "n sp" *OR* "nov sp" *OR* "new gen*" *OR* "sp n" *OR* "sp nov" *OR* "n gen*" *OR* "gen n" *OR* redescript*) *AND* (nematod* *OR* roundworm *OR* trematod* *OR* fluke *OR* digenea* *OR* cestod* *OR* tapeworm *OR* acanthocephal*)). A total of 7,724 entries were retrieved by November 29th 2017 (last day of search). Species redescriptions were also considered whenever the original description was made prior to 1970 and if based on new material. In such cases, only studies where a redescription was actually made were considered (e.g. amendments were excluded). Further exclusion criteria included: i) descriptions of parasites of domestic animals (including pets) or captive animals (unless information from their wild location of origin was given); ii) new species where only a name was given but lacking a proper formal description; iii) studies where the definitive host was unknown and no experimental approaches were used to infer the identity of the definitive host. In the latter case, whenever two different host groups, e.g. mammal and birds, were found suitable hosts, the parasite was included in both vertebrate datasets (only 1 case). We also included records of parasites described in vertebrate hosts from fish or crocodile farms. We examined all retrieved publications individually and recorded from all genuine species descriptions: (i) parasite species name, (ii) higher taxon, (iii) description type (i.e. new or redescription), (iv) host species, (v) host higher taxon, (vi) host order, (vii) habitat, i.e. terrestrial, freshwater or marine, (viii) locality where the parasite was discovered, (ix) its latitude and longitude, and (x) and the full reference. Whenever geographical coordinates were not given in the original article, locality coordinates were obtained from Google Earth v. 7.3.0 [2]. Whenever multiple nearby localities were given, only one was selected at random, and in cases where a longitudinal or latitudinal range was given we determined the mid-point coordinates. The final dataset included 4889 articles, from which descriptions of 4943 parasite species were collected (Table 1).

For data on host species richness, we downloaded from the IUCN online data base ([3]; http://www.iucnredlist.org/technical-documents/spatial-data) data on species' geographic distributions of amphibians, reptiles, terrestrial mammals, freshwater and marine fishes (including both Osteichthyes and Chondrichthyes). Note that IUCN data on reptiles, marine fish and freshwater fish are considered "*not comprehensive*". For birds, data were obtained from BirdLife International [4] with permission for their non-commercial use. The original providers of the vertebrate host data remain the owners of the data.

(b) Spatial analysis

Prior to analysis, parasite point location data was converted to a spatial points data frame using the *sp* package (function *SpatialPointsDataFrame*) [5]. Host distribution data was also edited prior to analysis: i) reptile and marine fish ("Chondrichthyes" and "Marine Fish") distribution data were cropped to remove points from polygons that fell outside longitude and latitude range values (i.e. -180, 180, -90, 90) using the function *crop* from the R package *raster* [6] with extent of -180,180,-90,90; ii) the shapefiles of marine fish data ("Chondrichthyes" and "Marine Fish") were then joined in QGIS v. 2.14.3-Essen [7].

To generate global maps of both parasite discoveries and host species richness, species' geographic distribution data were transformed into two presence-absence matrices, one with a global grid of 1º of resolution and the other with 2º resolution, using the function *lets.presab* of the R package *letsR* [8]. To explore similarity (or dissimilarity) in patterns of spatial distribution between parasite species discoveries and host species richness, we computed correlation coefficients among grid cells, separately for all the six vertebrate groups, at each of the two resolutions. Given the sample sizes for each of the four parasite groups, calculations were performed only for the pooled parasite data. Prior to statistical analysis, joint absences (double zeros, i.e. grid cells where hosts do not occur and no parasite has been found) were excluded from the dataset, since they artificially contribute to similarity between variables [9,10]. We first computed Spearman's correlations (R function *cor*.test), ignoring spatial autocorrelation. However, species distributional data often display spatial autocorrelation, i.e. locations close to each other are more likely to have comparable values than expected by chance [11]. Both host and parasite species distribution data are likely to be spatially autocorrelated, which to some degree can result from sampling biases especially in the case of parasite species discovery. Statistically, this lack of independence means that each sampling location does not represent a full degree of freedom, and adjusted degrees of freedom should be used to account for the intensity of spatial autocorrelation in each variable. Given the scale of our study, to control for spatial non-independence we used a modified *t*-test [12] to calculate the statistical significance of the correlation coefficient (a corrected Pearson's correlation) based on geographically effective degrees of freedom [13] as implemented in the *SpatialPack* package (function *modified.ttest*) [14]. Since the reliability of this correction is directly related to the estimated degree of spatial autocorrelation, which in turn varies

according to the number of distance classes [15,16], the correlation was calculated for 5, 13 (default) and 20 classes. To examine patterns of autocorrelation of each variable, the estimated Moran's indices [17] of each variable (also an output from the function *modified.ttest*) were plotted as a correlogram.

To more explicitly consider spatial information when determining the degree of association between the distributions of parasite discoveries and host species richness, we calculated the Tjøstheim's coefficient [18] with the function *cor.spatial* (*SpatialPack* package) [14]. The codispersion coefficient (also known as Matheron's coefficient) [19] which quantifies the coefficient of association between two spatial variables that are separated by a distance *h* (lags) was also estimated using the function *codisp* of the *SpatialPack* package for 13 distance classes. The above measurements tackle different aspects of spatial correlation, with codispersion and the corrected Pearson's correlation coefficient being more similar [see 20-22 for further discussion].

To visually represent the mismatch between the global distribution of parasite discoveries and that of host species richness while accounting for differences in study effort, we first obtained relative values by dividing the raster containing numbers of species per cell by the total number of species of either parasites found or known hosts, for each of the two resolutions. Then, we subtracted the relative value for hosts from that for parasites of the same cell, across all cells, and produced global maps with the resulting values. A predominance of values very close to zero, either negative or positive, would indicate strong proportionality between local host species richness and how many parasite species have been found. The higher the resulting value in a cell (the more positive it is), the greater the relative discovery of parasites relative to the local host species richness. Conversely, cells with low resulting values (i.e. strongly negative values) represent areas where disproportionately few parasites have been discovered relative to local host richness.

Also, we examined whether differences in sampling effort among host groups shape patterns of parasite species discovery. We calculated the percentage of total known host species richness (from IUCN and BirdLife International data) represented by the host species in our database, i.e. hosts from which new parasites have been discovered between 1970 and 2017 (species described from experimental procedures were not considered). We also calculated the percentage of host species in the database from which more than one parasite was described (i.e. host sharing). Typically, from a sample of individual hosts taken from one population (one grid cell), only one new parasite species is described. However, sometimes two or more parasites are described from the same host sample. To test whether the relationship between the number of parasite species described per grid cell and the number of host species from which parasites were found varies among vertebrate host groups, we used spatial Generalized Linear Mixed Models (GLMM). We fitted the structure of the variance-covariance- matrix to the data as described in [11]. These spatial models can eliminate or at least decrease spatial autocorrelation [11], allowing for a more reliable estimate of the degree of association. Analyses were performed separately for the six vertebrate host groups (amphibians, reptiles, birds, terrestrial mammals, freshwater fish, marine fish) by pooling all parasite species, and separately for the four

parasite groups (Acanthocephala, Cestoda, Nematoda and Trematoda) pooling all host taxa. Data were transformed into two presence-absence matrices as described above, but only for a global grid of 2º of resolution. The GLMM models were performed with the function *glmmPQL* (MASS package; [23]) implementing a spherical correlation structure, and fitting a quasi-Poisson distribution to account for overdispersion in the data (i.e. variance greater than the mean). Amphibian and acanthocephalan data were not overdispersed, so a Poisson distribution was used. Coefficients were computed as odds ratio, such that a value of 1 indicates that for each host species sampled in a grid cell, one parasite species was described. Confidence intervals could not be estimated for the coefficients, since glmmPQL does not return such values due to its penalizing behaviour. Spatial autocorrelation in model residuals was evaluated using a variogram (package gstat [24]), and with Moran'*s* I correlograms (ncf package [25]) and subsequently plotting Moran's *I* for 100 distance classes.

Finally, to visualize how parasite species discovered accumulate as a function of the number of sampled grid cells for each vertebrate and parasite taxon, we calculated the cumulative sum of parasite species (excluding cells where zero parasites were found; R function *cumsum*) with each additional cell sampled of a global grid of 2º of resolution, as described above. We computed 999 random permutations of the order of cells to obtain a 95% confidence interval.All analyses were performed in R statistical computing environment [26].

References

1. Poulin R. 2014 Parasite biodiversity revisited: frontiers and constraints. *Int. J. Parasitol.* **44,** 581-589. (doi:10.1016/j.ijpara.2014.02.003)
2. Google Earth. 2017 Google Earth Version 7.3.0. Retrieved from https://www.google.com/earth/.
3. IUCN. 2017 The IUCN Red List of Threatened Species, Version 5.2. Available at http://www.iucnredlist.org.
4. BirdLife International and Handbook of the Birds of the World. 2016 Bird species distribution maps of the world. Version 6.0. Available at http://datazone.birdlife.org/species/requestdis.
5. Pebesma EJ, Bivand RS. 2005 Classes and methods for spatial data in R. R News 5 (2), https://cran.r-project.org/doc/Rnews/.
6. Hijmans RJ. 2017 raster: Geographic Data Analysis and Modeling. R package version 2.6-7. https://CRAN.R-project.org/package=raster
7. QGIS Development Team. 2016 QGIS Geographic Information System. Open Source Geospatial Foundation. URL http://qgis.osgeo.org
8. Vilela B, Villalobos F. 2015 letsR: a new R package for data handling and analysis in macroecology. *Methods Ecol. Evol.* **6,** 1229-1234. (doi:10.1111/2041-210X.12401)

9. Legendre P, Legendre L. 1998 *Numerical ecology*, second English Edition. Amsterdam: Elsevier.
10. Zuur AF, Leno, EN, Elphick CS. 2010 A protocol for data exploration to avoid common statistical problems. *Methods Ecol. Evol.* **1,** 3-14. (10.1111/j.2041-210X.2009.00001.x)
11. Dormann CF**,** McPherson JM**,** Araujo MB**,** Bivand R**,** Bolliger J**,** Carl G, Davies RG, Hirzel A, Jetz W, Kissling WD *et al*. 2007 Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* **30,** 609-628. (doi:10.1111/j.2007.0906-7590.05171.x)
12. Clifford P, Richardson S, Hémon D. 1989 Assessing the significance of the correlation between two spatial processes. *Biometrics* **45,** 123-144. (doi:10.2307/2532039)
13. Dutilleul P. 1993 Modifying the t test for assessing the correlation between two spatial processes. *Biometrics* **49,** 305-312. (doi:10.2307/2532625)
14. Osorio F, Vallejos R. 2014 SpatialPack: Package for analysis of spatial data. R package version 0.2-3. http://cran.r-project.org/package=SpatialPack
15. Fortin M-J. 1999 Effects of sampling unit resolution on the estimation of spatial autocorrelation. *Ecoscience* **6,** 636-641. (doi:10.1080/11956860.1999.11682547)
16. Fortin M-J, Payette S. 2002 How to test the significance of the relation between spatially autocorrelated data at the landscape scale: a case study using fire and forest maps. *Ecoscience* **9,** 213-218. (doi:10.1080/11956860.2002.11682707)
17. Moran PAP. 1950 Notes on continuous stochastic phenomena. *Biometrika* **37,** 17-23. (doi:10.2307/2332142)
18. Tjøstheim D. 1978 A measure of association for spatial variables. *Biometrika* **56,** 109-114. (doi:10.2307/2335284)
19. Matheron C. 1965 *Les variables régionalisées et leur estimation*. Paris: Masson.
20. Glick BJ. 1982 A spatial rank-order correlation measure. *Geogr. Anal.* **14,** 177-181. (doi:10.1111/j.1538-4632.1982.tb00066.x)
21. Vallejos R. 2008 Assessing the association between two spatial or temporal sequences. *J. Appl. Stat.* **35,** 1323-1343. (doi:10.1080/02664760802382418)
22. Vallejos R. 2012 Testing for the absence of correlation between two spatial or temporal sequences. *Pattern Recogn. Lett.* 33, 1741-1748. (doi:10.1016/j.patrec.2012.05.013)
23. Venables WN, Ripley BD. 2002 *Modern Applied Statistics with S*, fourth edition. Springer, New York.
24. Pebesma EJ. 2004 Multivariable geostatistics in S: the gstat package. Computers & Geosciences **30**, 683-691. (doi:10.1016/j.cageo.2004.03.012)
25. Bjornstad ON. 2013 ncf: Spatial nonparametric covariance functions. R package version 1.1-5, http://CRAN.R-project.org/package=ncf.

26. R Core Team. 2015 R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Figure S1. Correlogram of Moran's index *I* for each host and parasite dataset at a resolution of 1º for 13 distance classes.

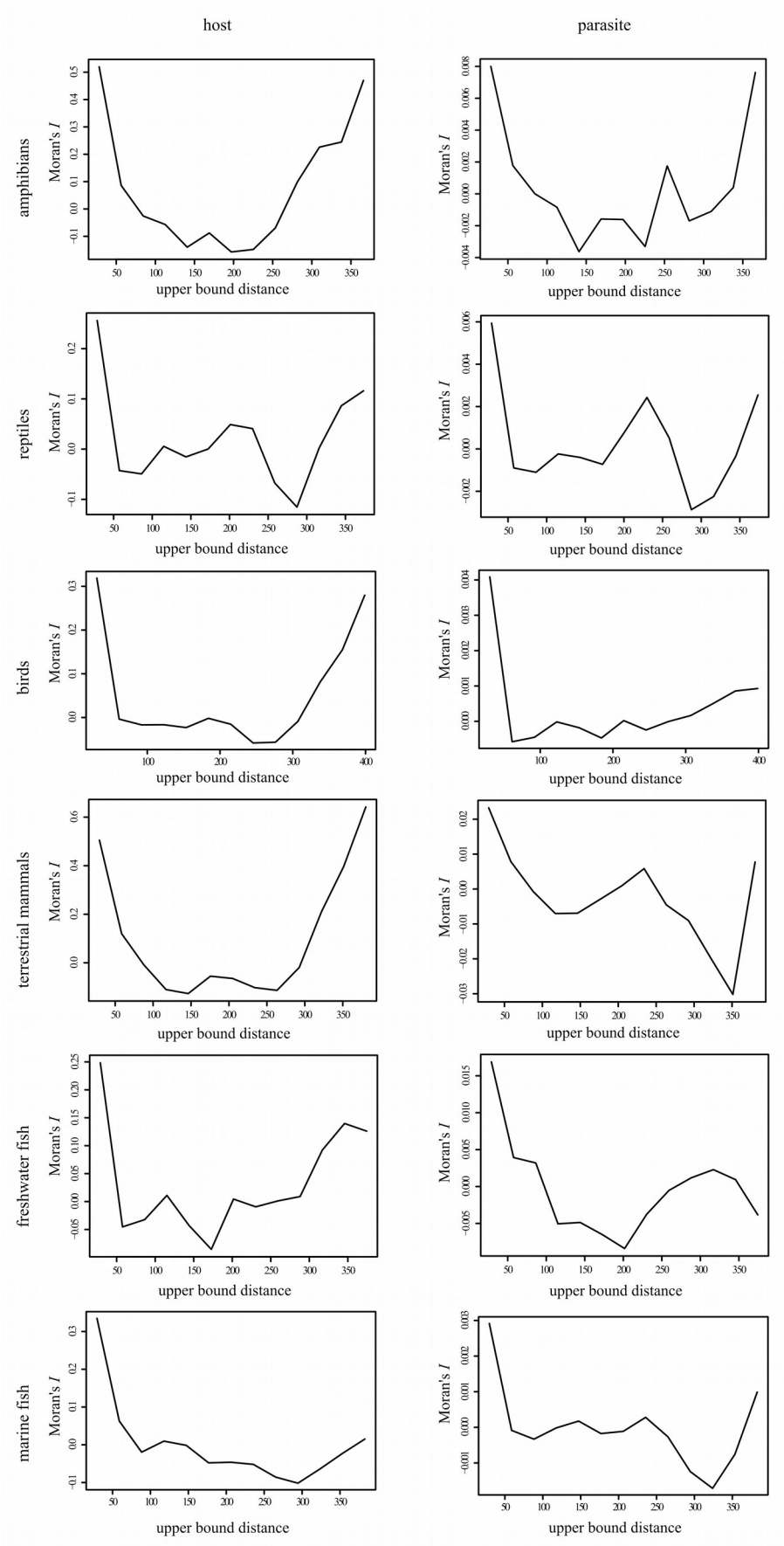


Figure S2. Correlogram of Moran's index *I* for each host and parasite dataset at a resolution of 2º for 13 distance classes.

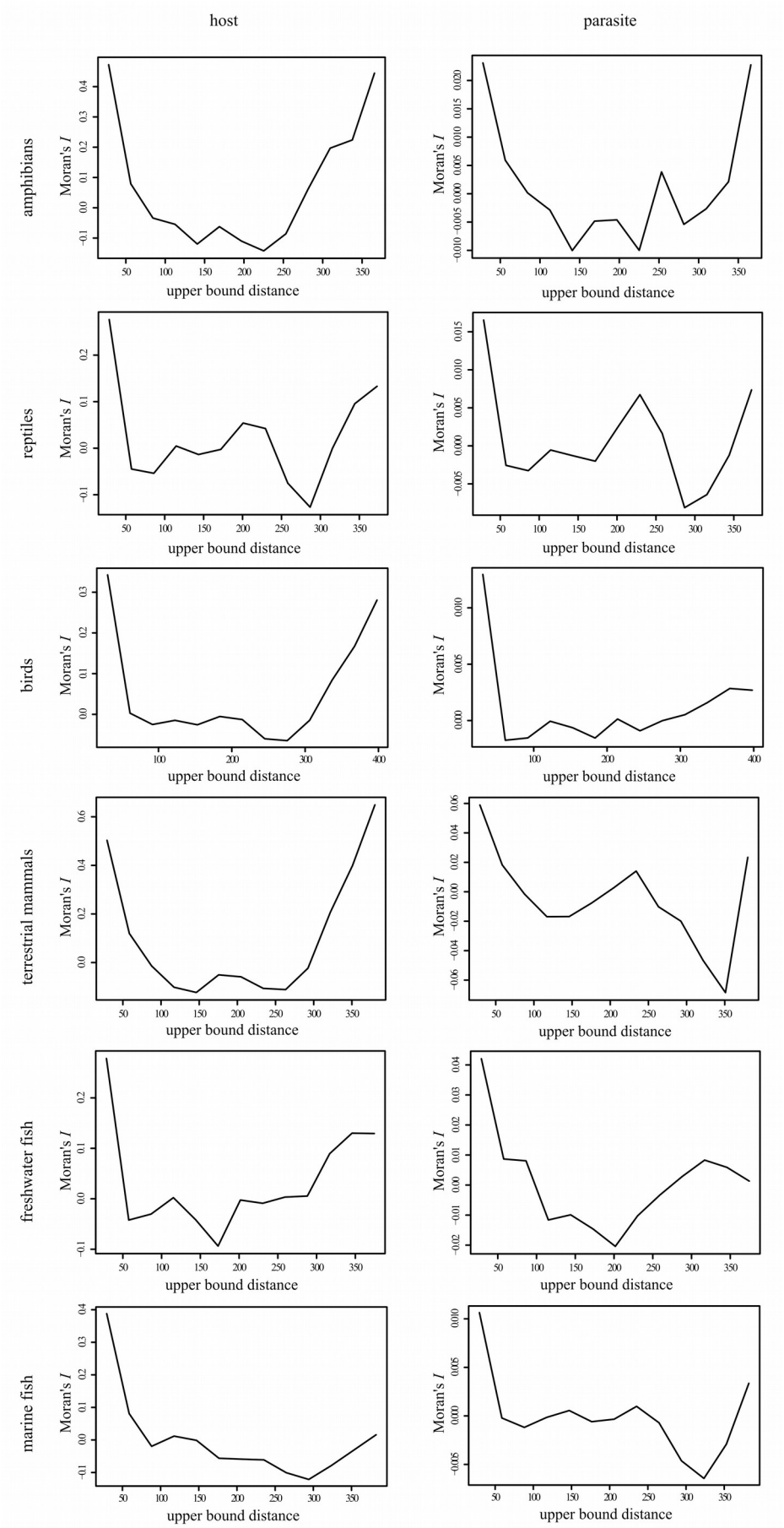


Figure S3. Codispersion coefficient between parasite and host data for each vertebrate taxon for 13 distance classes (lags), at resolutions of 1º and 2º.

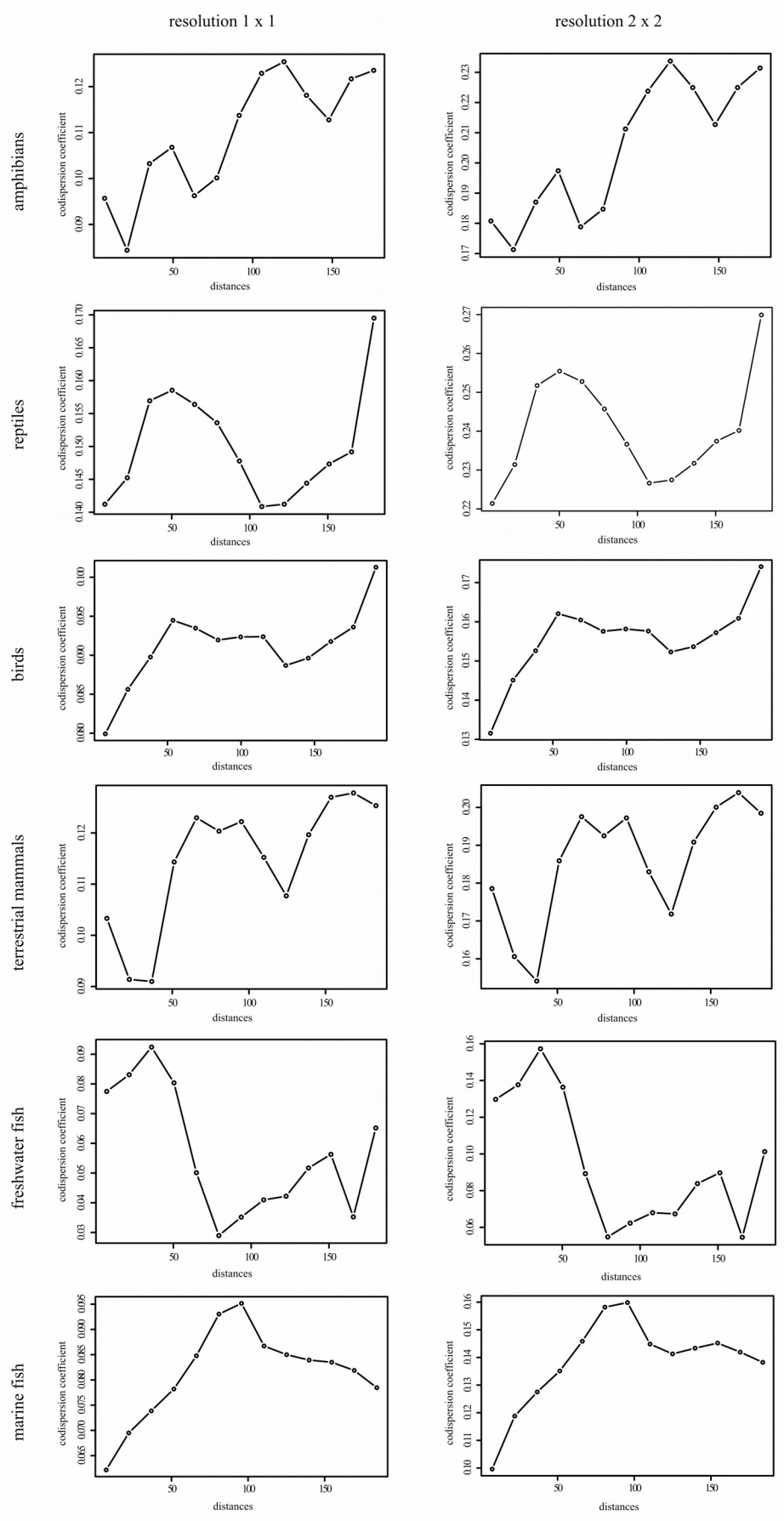
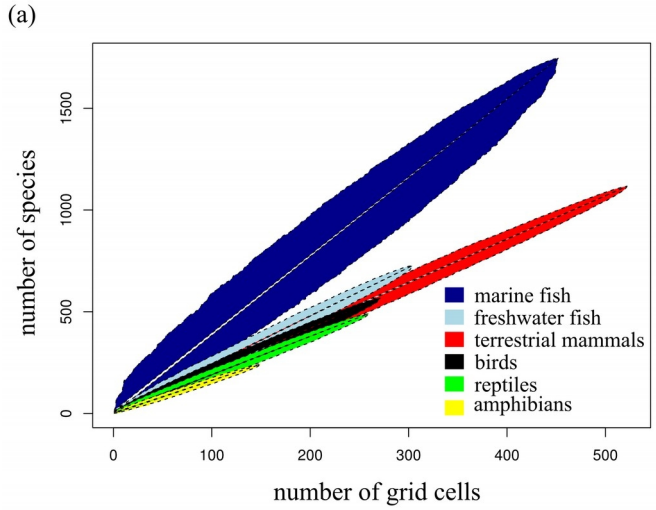


Figure S4. Cumulative parasite species discovery as a function of the number of sampled grid cells, for (a) each vertebrate host taxon and (b) each parasite taxon. Shaded polygons represent 95% confidence intervals from 999 permutations.

(a)



(b)