Cyberdiversity: Improving the Informatic Value of Diverse Tropical Arthropod Inventories

Jeremy A. Miller[1,2,3]*, **Joshua H. Miller**[4,5,6], **Dinh-Sac Pham**[7], **Kevin K. Beentjes**[8]

1. Department of Terrestrial Zoology, Naturalis Biodiversity Center, RA Leiden, The Netherlands, **2.** Department of Entomology, California Academy of Sciences, Golden Gate Park, San Francisco, California, United States of America, **3.** Plazi, Zinggstrasse 16, Bern, Switzerland, **4.** Department of Geology, University of Cincinnati, 500 Geology/Physics Building, Cincinnati, Ohio, United States of America, **5.** University of Alaska Museum, Fairbanks, Alaska, United States of America, **6.** Florida Museum of Natural History, University of Florida, Gainesville, Florida, United States of America, **7.** Institute of Ecology and Biological Resources, Vietnam Academy of Science and Technology, Cau Giay Dist, Ha Noi, Vietnam, **8.** Naturalis DNA Barcoding Facility, Naturalis Biodiversity Center, RA Leiden, The Netherlands

*jeremy.miller@naturalis.nl

🔓 OPEN ACCESS

Abstract

In an era of biodiversity crisis, arthropods have great potential to inform conservation assessment and test hypotheses about community assembly. This is because their relatively narrow geographic distributions and high diversity offer high-resolution data on landscape-scale patterns of biodiversity. However, a major impediment to the more widespread application of arthropod data to a range of scientific and policy questions is the poor state of modern arthropod taxonomy, especially in the tropics. Inventories of spiders and other megadiverse arthropods from tropical forests are dominated by undescribed species. Such studies typically organize their data using morphospecies codes, which make it difficult for data from independent inventories to be compared and combined. To combat this shortcoming, we offer cyberdiversity, an online community-based approach for reconciling results of independent inventory studies where current taxonomic knowledge is incomplete. Participating scientists can upload images and DNA barcode sequences to dedicated databases and submit occurrence data and links to a web site (www.digitalSpiders.org). Taxonomic determinations can be shared with a crowdsourcing comments feature, and researchers can discover specimens of interest available for loan and request aliquots of genomic DNA extract. To demonstrate the value of the cyberdiversity framework, we reconcile data from three rapid structured inventories of spiders conducted in Vietnam with an independent inventory (Doi Inthanon, Thailand) using online image libraries. Species richness and inventory completeness were assessed using non-parametric estimators. Community similarity was evaluated using a novel index

based on the Jaccard replacing observed with estimated values to correct for unobserved species. We use a distance-decay framework to demonstrate a rudimentary model of landscape-scale changes in community composition that will become increasingly informative as additional inventories participate. With broader adoption of the cyberdiversity approach, networks of information-sharing taxonomists can more efficiently and effectively address taxonomic impediments while elucidating landscape scale patterns of biodiversity.

Introduction

As biodiversity continues its unabated decline [1,2], taxonomic and geographic biases constrain our ability to understand and predict the consequences of these losses and devise effective mitigation strategies [3,4]. In terms of richness and abundance, arthropods dominate animal life, especially in the tropics. Yet vertebrates and vascular plants, both of which have comparatively low diversity, are the dominant study subjects for assessing the biological ramifications of anthropogenic perturbations and establishing conservation priorities [5–7]. While vertebrates have been shown to be poor surrogates for arthropod conservation priorities [5–7], the geographic distribution of arthropod species has been found to reliably predict the conservation priorities of vertebrates (i.e., optimizing networks of reserve areas to maximize the persistence of species [8]). This asymmetry arises largely because the geographic ranges and environmental tolerances of individual arthropod species tend to be more restricted than vertebrates or vascular plants, enabling megadiverse arthropod groups to track ecological gradients at finer spatial resolution. Because of their high richness and sensitivity to environmental variables, arthropods offer some of the finest-grained data with which to assess terrestrial biodiversity at individual localities (alpha [α] diversity) and changes across landscapes (beta [β] diversity) [5,9]. However, one of the greatest impediments to the broader use of arthropod communities for studying and maintaining global biodiversity is our current profoundly incomplete and geographically biased data on fundamental taxonomy. Substantial proportions of arthropods, especially those found in tropical regions, have yet to be formally described and lack scientific names [10]. Yet, names are the mechanism by which data about a species (including geographic distribution) are aggregated. The process of naming and describing species, however, is time-consuming. While online innovations are increasing the pace of species descriptions and improving accessibility to taxonomic data, these advances come at a time when investment and training in taxonomy is declining [11, 12]. If the fine-grained pattern of arthropod biodiversity is to be broadly integrated into conservation assessment, it is apparent that we need to diminish dependence on formal scientific names. To accommodate this need, an alternative model is emerging designed to share biodiversity data that is not yet ready for formal

taxonomic publication so researchers may efficiently and effectively evaluate and integrate that information with other data [13, 14].

Two technological advances are revolutionizing standard practice in inventories of arthropods (including spiders): (1) the increased accessibility of digital photomicroscopy, and (2) DNA barcoding. Perhaps the most precocious online collection of digital images of a diverse arthropod taxon is AntWeb (http://www.antweb.org/). At the time of this writing, this global effort offers photos of nearly 16,000 ant species freely available online along with specimen occurrence data. Unidentified ant species referred to only by morphospecies codes (rather than valid Linnaean binomials) in some published studies are also included [15]. AntWeb gives independent researchers working in the same region an opportunity to determine which species – including species not yet formally described – are shared between independent studies without the need to physically examine vouchers. Photographs and specimen occurrence records can provide much of the determination power we expect from formal taxonomic literature, even when these resolve to morphospecies codes rather than Linnaean names. DNA barcoding offers an independent method of species identification and classification. This approach involves building a library of sequences from a standard region of the genome to aid species identification and discovery [16]. For animals, the ~648 base pair region of the mitochondrial gene cytochrome oxidase I has become the dominant barcode marker [16, 17]. DNA barcoding as an enterprise has strengths and limitations, and these have been the subject of spirited debate [18–33]. Within this debate, some have argued convincingly that data from multiple independent sources (e.g., morphology, DNA sequences) should be considered (e.g., [19, 20, 22, 23, 26, 27, 31–33]). Given that classification based on either data source alone fails some of the time, disagreement between approaches indicates the need for focused study to resolve the conflict [19, 20, 34]. By extension, determining the number of species within and shared between ecological inventories based on a combination of morphological and molecular sequence data is preferable to relying on either method alone. Species-level taxa determined on the basis of combined morphological and DNA sequence data (i.e., without formal names) are referred to as integrated operational taxonomic units (IOTUs) [35].

Spiders are one of the richest orders of life on earth, and structured inventories (i.e., surveys that use replicable sampling protocols with multiple complimentary collecting methods) are fundamental sources of data about species richness [36–38]. However, inventories in temperate regions have significantly higher proportions of their species identified with scientific names than inventories in tropical regions (Fig. 1). This is perhaps not surprising given the relatively low number of species and high intensity of taxonomic research at temperate latitudes. The consequence of this taxonomic imbalance is that data from temperate inventories may be far more readily integrated with existing knowledge compared to tropical inventories. In regions with relatively undeveloped taxonomic literature, ecological studies typically categorize unidentified species using "morphospecies" concepts [39–41]. This means using the skills of a morphological
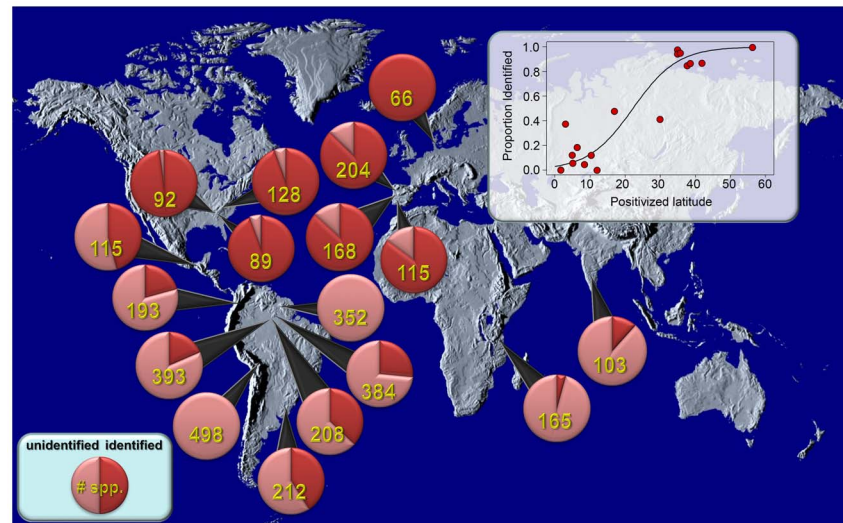
Fig. 1. Structured inventories of spiders from around the world. Red portion of each pie chart represents species identified according to formal nomenclature in the original publication; the remaining pink portion represents those identified to morphospecies. Total observed species richness is in yellow. Inset shows significant relationship between distance from the equator (expressed as positivized latitude) and the proportion of identified species (logistic regression; β=0.113 (β_{se}=0.0488, e^{β}=2.306), p=0.0211, McFadden's R^2=0.72). Studies with a high proportion of identified species (which are largely found in temperate regions) are relatively easy to evaluate for community similarity, while studies in regions with more poorly developed taxonomy (e.g., the tropics) may not be as easily reconciled. This roadblock can be bypassed using the cyberdiversity framework, which allows data on the whole community to be made publically available, and can foster reconciliation of independent inventories, including those with high proportions of undescribed species. Data from Colombia [62], Brazil [63–65], Denmark [54], Guyana [39], India [66], Mexico [67], Peru [56], Portugal [57–59], Tanzania [68], and United States [55, 60, 61]. In some cases, specimens representing morphospecies from these studies have been subsequently described in taxonomic publications [112, 113].

doi:10.1371/journal.pone.0115750.g001

taxonomist to classify individuals in the collection without depending on incomplete and fragmentary taxonomic literature. This approach is sufficient for elucidating biodiversity patterns within a particular study, but makes it cumbersome to compare results between independent studies. Conscientious investigators typically deposit voucher specimens in museum or university collections, which means that morphospecies from different studies can be reconciled, but doing so is often prohibitively time consuming. As a consequence, independent biodiversity studies on the same taxa in a single region have limited capacity to build on each other or to document biological patterns beyond the scope of each individual study.

The cyberdiversity approach

Cyberdiversity is an online approach to facilitate species recognition regardless of taxonomic determination status. Web-based tools may include collections of digital images, DNA sequences, or ideally both. The cyberdiversity approach engages several recognized impediments to the understanding of fundamental aspects of megadiversity and the wider adoption of arthropod data for

conservation assessment. These challenges include (a) the large number of species that remain undescribed, (b) the scientific ignorance concerning the geographic distribution, abundance, and environmental sensitivies of most species, and (c) the lack of awareness of invertebrate conservation issues in the social-political spheres [10]. Photographs and DNA barcodes, organized by persistent unique identifier, allow data on the geographic distribution and abundance of species to be compiled whether or not the species have scientific names [13,14]. The ready availability of these data through online databases makes the effort visible to a range of stakeholders. Following a central principal of biodiversity informatics, the primary data are formatted according to community standards and digitally exposed in ways that allow them to be aggregated, recombined, and repurposed [42–44].

Here, we introduce a new, expandable cyberdiversity resource: www. digitalSpiders.org, currently populated with data from our three 2009 inventories of Vietnamese spiders (Fig. 2). The web site layout features a three-panel design consisting of (1) a taxonomic navigation tree organized by IOTU within family, (2) species identifier (unique IOTU code, and where available taxonomic name with determination credit), images, records, links to DNA barcode sequences (where available), and user comments, and (3) a map indicating species presence for each inventory location. The three panel design permits simultaneous viewing of morphology and occurrence data, delivering with a single click the most critical information needed to reconcile future inventory data with that currently available on the web site. Unfortunately, the leading collaborative biodiversity data environment, Scratchpads (http://scratchpads.eu/), does not currently support such a layout. The digitalSpiders site also makes all records available through Google Earth with markers linked to collections of images on Morphbank. In the Google Earth environment, records can be filtered to display any combination of records or IOTUs. New content can be submitted using the digitalSpiders data template, subject to validation by the site administrator.

Ours is the first structured tropical spider inventory study to include DNA barcode data for most of the sampled community. Fortunately, it is not the first study from the region to make libraries of morphospecies images available online. Images of all morphospecies resulting from a 2003 survey of spiders from Doi Inthanon National Park, Chang Mai Province, Thailand have also been posted online (http://aracnologia.macn.gov.ar/ThaiPlot/). This inventory was conducted in October, the same time of year as our Vietnamese study, by an independent research team. The results of the Doi Inthanon study have not been published in the form of a scientific journal article. Nevertheless, because the leaders of this study chose to post images of their morphospecies online, we were able to rapidly assess characteristics of change in spider communities (β-diversity) across ca. 1,000 km.

The cyberdiversity approach is a call to researchers to share data that facilitate the reconciliation of inventory results across studies, multiplying the number of sites available for analysis. The ultimate goal is to accumulate enough reconciled inventory points to meaningfully model patterns of diversity on regional,

A.

B.

Fig. 2. Data visualization and schematic data relationships on the digitalSpiders cyberdiversity web site. (A) The digitalSpiders website is designed around a three-panel layout. The left panel is a taxonomic index to IOTUs. Once an IOTU is selected, the center panel shows the taxonomic name (where available), a collection of images, collection records, and a utility for submitting and viewing public comments. Individual images are linked to higher resolution versions on Morphbank annotated with extensive metadata. The image collection for the IOTU is also linked to a corresponding image collection on Morphbank. Records may be associated with a DNA barcode sequence and linked to the BOLD database. All records can be viewed through Google Earth with markers linked to collections of images on Morphbank. New content can be submitted using the digitalSpiders data template. (B) Data relationships are configured around an IOTU identified by a unique species code. Each IOTU is associated with one or more records. Each of those records is associated with a sample taken at a location. The map is derived from the set of unique locations. All records can be viewed through Google Earth with markers linked to collections of images on Morphbank. Records may be associated with a DNA barcode sequence and linked to the BOLD database. Each IOTU is associated with a collection of one or more images, linked to a corresponding image or image collection on Morphbank. Where an IOTU is associated with a valid taxonomic name, this determination is credited to one or more taxonomists along with their contact information.

continental, and even global scales. With data from only four sites currently available for comparison (three in Vietnam, one in Thailand), we cannot yet provide a definitive analysis of large-scale biodiversity patterns. Instead, we present a series of preliminary analyses to demonstrate the kinds of questions and diversity parameters that can be addressed once more inventories are reconciled.

Modeling horizon: landscape-scale change in megadiverse communities

Once we have assessed community similarity across sites, we can begin to focus on quantifying β-diversity and identifying its geographic and climatic drivers. We present a preliminary demonstration of this in a distance-decay framework [45, 46], where community change (i.e., pair-wise community similarity) is modeled as exponential decay functions of both the geographic and climatic (WorldClim, http://www.worldclim.org/bioclim [47]) distances between sites. Using this exponential decay framework, we can then estimate the species turnover rate; specifically, the geographic or climatic distance across which half the species are different from one point to another (halving distance; $d_{0.5}$, which is constant across space). We can also estimate the similarity of communities at small distances, specifically the initial similarity (s_0), when distance between community samples equals zero. With a sufficient number of sites, variance partitioning can be used to disentangle covariance between geographic and climatic change [48], thought we do not attempt this here. Our preliminary analyses are intended to demonstrate the kinds of questions that are possible to explore with reconciled inventory data.

Results

Our surveys of three study sites in Vietnam yielded 2,009 adult spiders comprising 240 species. Non-parametric species richness estimators (Chao 1, Chao 2, ACE, and ICE) indicate that the Cuc Phuong inventory was the closest to completion (range of mean estimated completeness [hereafter, estimated] 73–82%). The Cuc Phuong inventory yielded the lowest number of observed species (76) and the

Table 1. Observed and estimated species richness for the three rapid inventories of spiders in Vietnam.

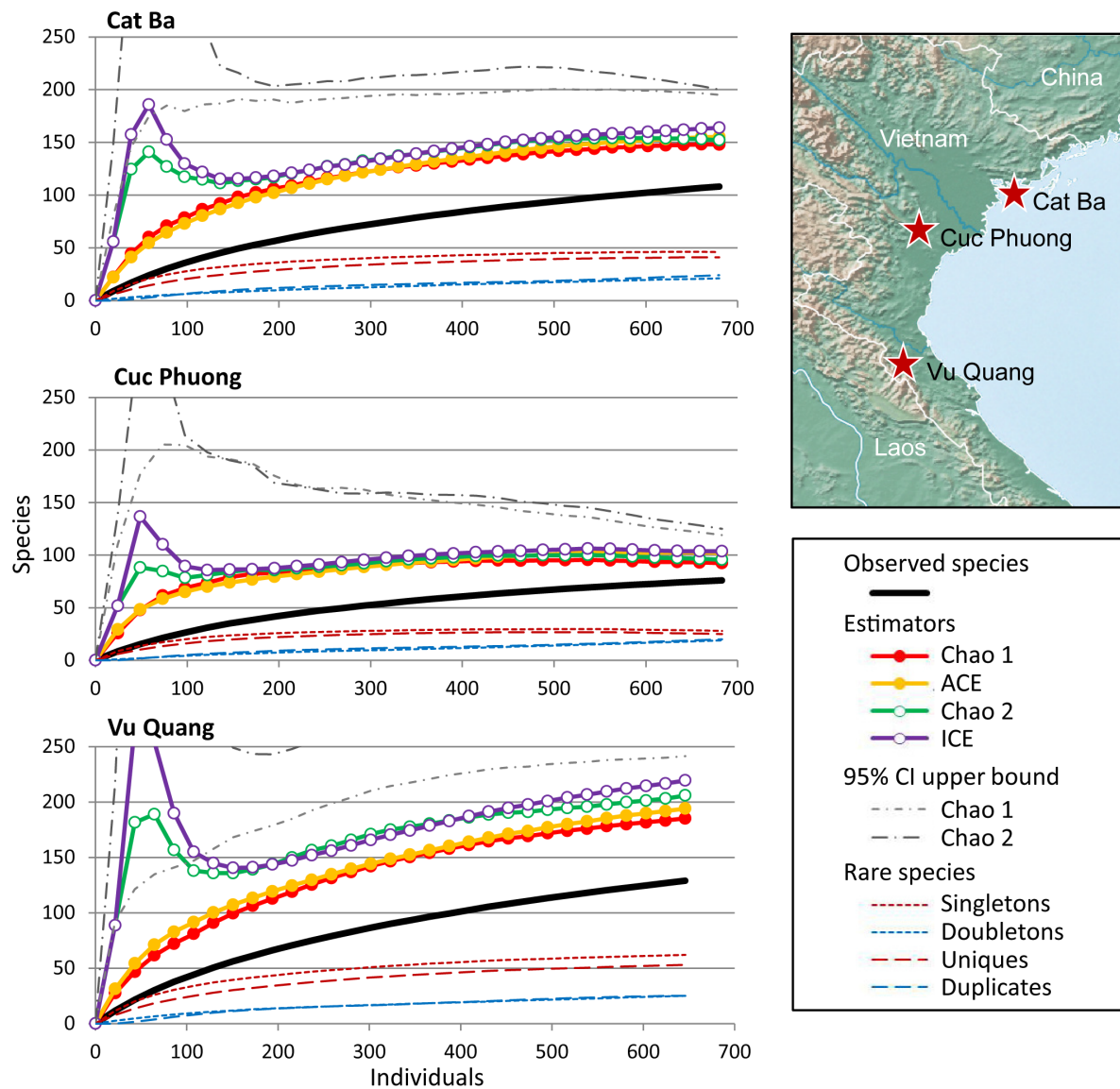| Site	Observed Richness	Estimated richness [estimated completeness of observed sample]			
		Chao 1 (upper 95% C.I.)	ACE	Chao 2 (upper 95% C.I.)	ICE
Cat Ba	108	148 (195), [73% (55%)]	155, [70%]	152 (200), [71% (54%)]	164, [66%]
Cuc Phuong	76	93 (119), [82% (64%)]	102, [74%]	96 (125), [80% (61%)]	104, [73%]
Vu Quang	128	178 (229), [72% (56%)]	188, [68%]	197 (262), [65% (49%)]	213, [60%]

doi:10.1371/journal.pone.0115750.t001

Fig. 3. Species richness estimation curves with map showing locations of Vietnamese study sites. Data include all adult spider specimens collected from one hectare plots in three Vietnamese National Parks: Cat Ba, Cuc Phuong, and Vu Quang. The number of species observed from each inventory, and the number of singletons, unique, doubletons, and duplicates in each inventory are given, as well as four non-parametric estimators of sample completeness: Chao 1, ACE, Chao 2, and ICE (upper 95% confidence intervals provided for Chao 1 and Chao 2). Estimated Inventory completeness is variable, but none are complete. Hence, assessment of community similarity based on these inventories should account for unobserved shared species.

doi:10.1371/journal.pone.0115750.g003

Table 2. Observed and estimated similarity of sampled spider communities in Southeast Asia.

Site pair	Jaccard similarity (shared/ combined species)	Estimated shared species (95% C.I.)	Estimated combined species	Chao's Estimated Proportional Similarity (95% C.I.)
Cat Ba - Cuc Phuong	0.20 (31/153)	45 (31, 65)	212	0.21 (0.14, 0.34)
Cat Ba - Vu Quang	0.14 (29/207)	42 (29, 73)	302	0.14 (0.10, 0.27)
Cat Ba - Doi Inthanon	0 (0/214)	0 (0, 0)	287	0 (0, 0)
Cuc Phuong - Vu Quang	0.17 (29/175)	44 (29, 93)	247	0.18 (0.11, 0.47)
Cuc Phuong - Doi Inthanon	0.0056 (1/181)	1 (1, 19)	233	0.005 (0.004, 0.090)
Vu Quang - Doi Inthanon	0.031 (7/227)	12 (7, 32)	309	0.037 (0.022, 0.109)

Estimated combined species is the sum of ACE for both sites minus the estimated shared species.

doi:10.1371/journal.pone.0115750.t002

highest number of individuals (683), so the sampling intensity (the number of individuals per species) was comparatively high (9.0). The Cat Ba inventory was only slightly less complete (estimated 66–73%). The observed species richness in Cat Ba was considerably higher than Cuc Phuong (108) for almost the same number of individuals (680), so the sampling intensity was correspondingly lower (6.3). The Vu Quang inventory was the richest of the three sites with 128 observed species from a sample of only 646 individuals (sampling intensity 5.1). Non-parametric estimators suggest that this inventory was 60–72% complete (Table 1, Fig. 3).

All three inventories recovered largely different faunas with pairs of sites sharing only 29–31 observed species; 17 species were observed at all three sites. The proportion of observed shared species (the Jaccard index) across site pairs ranges from 0.14–0.20. Chao's estimated proportional similarity (see Methods) ranges from 0.13–0.19 with fairly wide 95% confidence intervals (Table 2).

DNA barcode sequences were obtained from 176 of the 240 species in the study (73%). Sequencing was attempted on 531 specimens (26% of the collection), 372 of which (70%) yielded a barcode sequence. In total, DNA sequences were obtained for 19% of the collected specimens. The BIN (Barcode Index Number) algorithm [49], which partitions barcode sequences into species-like taxonomic units (independent of morphology), suggests the barcodes obtained for this study represent 188 species, a net increase of 12 species compared to the results based on the combination of morphological and molecular sequence data (i.e., IOTUs).

Intraspecific variation in the barcode sequence was assessed based on 73 species for which more than one individual was successfully sequenced. Of these, within-site variability was assessed using 192 conspecific pairwise comparisons from 63 species, and between-site variability was assessed using 121 conspecific pairwise comparisons from 33 species. Overall, conspecific distances between sites (based on the optimal Felsenstein 1984 [50] model) were considerably higher than within sites (Mann-Whitney test, U=3901, z=−9.908, p=0.0001), suggesting geographic population structure (Figs. 4A and S1 in S1 File).
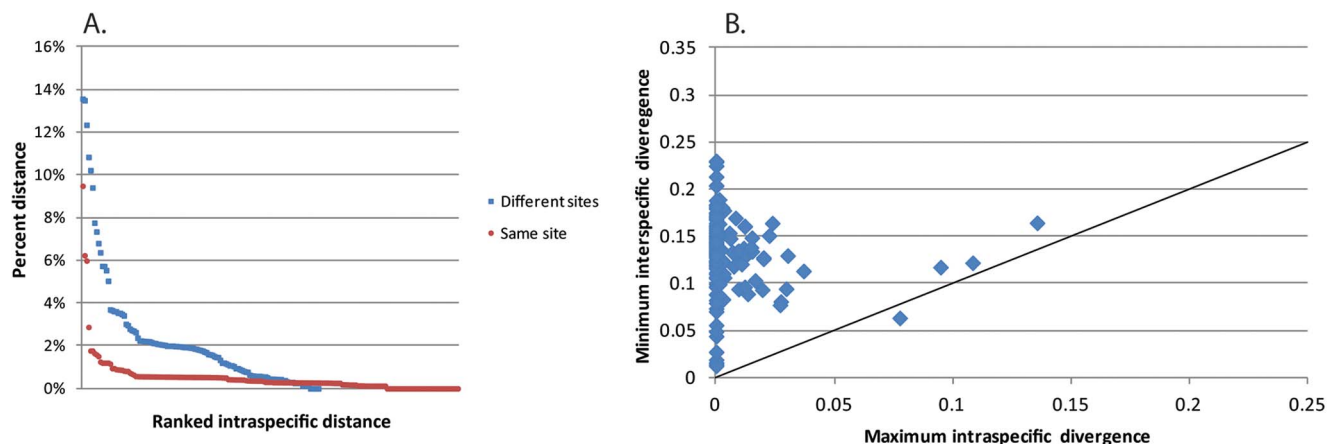
Fig. 4. Discriminatory power of DNA barcodes. (A) Within-species genetic distances (within-site [red], between-site [blue]) ranked by magnitude. (B) The barcode gap expressed as the maximum within-species distance compared to the minimum between-species distance; line shows equal interspecific-intraspecific distances. The magnitude of intraspecific genetic distance is variable across species, but maximum intraspecific distance is almost always less than the minimum interspecific genetic distances. Distance modeled using the Felsenstein 1984 model [50], which, using the IOTU classification, optimizes these data according to the Akaike Information Criterion. See supplementary documents for the same data modeled using Kimura 2-parameter and uncorrected p distances.

We find that community similarity (Chao's estimated proportional similarity) for Southeast Asian spiders is significantly correlated with geographic distance (Mantel test r=0.99, p=0.042), and that the community halving distance based on all available data is 171 km (Jackknifed mean (mean$_J$) and standard error (SE$_J$): 192±111 km) (Fig. 5). Initial similarity (i.e., similarity at 0 distance, s_0) was 0.43 (mean$_J$ ± SE$_J$: 0.48±0.17). Across the study area, the 19 climatic variables and altitude data are highly correlated, with the first two principal component axes characterizing 99% of the variance (Table S1 in S1 File). Based on a combined distance matrix of those first two principal component axes, we find that Chao's estimated proportional similarity is also significantly correlated with change in climate among sites (Mantel test r=0.95 p=0.044). Using all available data, initial similarity based on climate was close to that calculated for geographic distance: s_0=0.38, but the jackknifed mean and standard error was higher (mean$_J$ ± SE$_J$: 0.68±0.48). The halving distance for climate is $d_{0.5}$=2.25 (mean$_J$ ± SE$_J$ =3.00±3.77), which corresponds to a 93.2% (range across ± SE$_J$: 80.2–100%) decay in community similarity across the maximum distance of the sampled region. This is within error of the estimated change in community similarity based on the halving distance for geographic change: 96.5% (88–100%).

Cobra analysis (see Methods, Table 3) revealed both strengths and inefficiencies in our sampling regime. The largest number of sampling hours at all sites were devoted to AEN (searching for spiders in the aerial stratum at night), and this proved to be the most efficient method for species discovery in all cases. In Cat Ba (12 hours), AEN was relatively saturated; in Vu Quang (9 hours), more samples could have been allocated to this method; Cuc Phuong (10 hours) was intermediate. WIN (extraction of arthropods from sifted leaf litter using Winkler
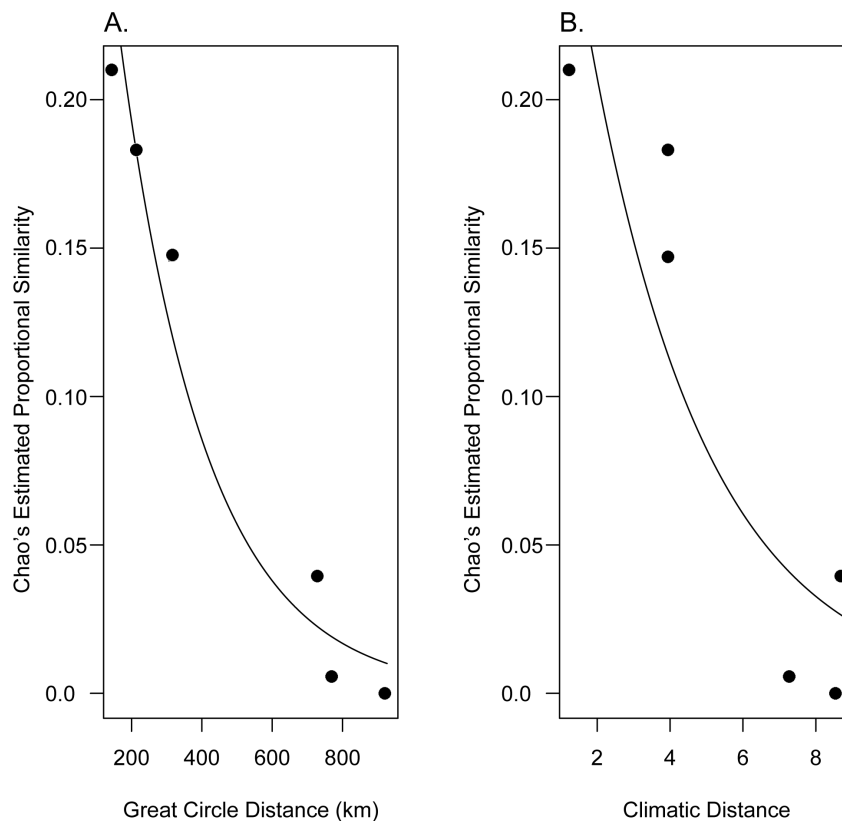
Fig. 5. Distance decay of community similarity. Pair-wise community similarity for Southeast Asian spider communities and their decay against (A) geographic and (B) climatic distances. Similarity is significantly correlated with both geographic and climatic distances (mantel tests; $p < 0.05$). For geographic distance, the estimated community halving distance is 171 km. With additional inventory sites, variance partitioning could be used to disentangle covariance between geographic and climatic change.

Table 3. Allocation of samples by method for the three rapid inventories of spiders in Vietnam with totals of adult specimens and species by method and site.

Site		Method					
		BED	AEN	LDD	LDN	WIN	Total
Cat Ba	Samples	6	12	4	8	5	35
	Adults	121	196	50	55	258	680
	Species	29	34	12	17	18	108
Cuc Phuong	Samples	4	10	4	6	4	28
	Adults	44	258	133	92	156	683
	Species	11	28	16	17	14	76
Vu Quang	Samples	5	9	6	6	4	30
	Adults	64	265	169	56	92	646
	Species	19	36	21	18	15	128

BED: beating vegetation during the day; AEN: searching for spiders in the aerial stratum at night; LDD: searching for spiders on the ground during the day; LDN: searching for spiders on the ground at night; WIN: extraction of arthropods from sifted leaf litter using Winkler traps.

traps) ranked second in field hours at all sites (two hours spent sifting leaf litter per sample), and was the second most efficient method in two sites (Cat Ba and Cuc Phuong). WIN at Cat Ba (10 hours) and Vu Quang (8 hours) approached relative saturation, but more WIN samples could have been allocated at Cuc Phuong (8 hours). BED (beating vegetation during the day) ranked third or fourth in efficiency and fourth (or tied for fourth) in effort. BED included moderately to highly efficient samples and more effort could have been usefully allocated here, especially at Cat Ba. LDD (searching for spiders on the ground during the day; 4–6 samples per site) contributed enough unique species to the inventory that it includes some high scores in the Cobra analysis, but most of the species contributed to the inventory using this method would have been collected with fewer samples. LDN (searching for spiders on the ground at night; 6–8 samples per site) was the least efficient method, and included several sample hours at each site that contributed little to the inventory; too much effort was devoted here.

Discussion

Cyberdiversity takes on the Taxonomic Impediment

Today's urgent need for better biodiversity knowledge is mismatched by the relatively slow pace of taxonomic progress for tropical spiders and other megadiverse groups. Using online databases to increase the recognizability of species regardless of whether they have a scientific name is one way to mitigate this asymmetry [13, 14]. For example, on the digitalSpiders web site, each IOTU page features a comments field to facilitate discussion and contribute to taxonomic identification (Fig. 2). Cyberdiversity resources can also serve to stimulate traditional taxonomy; taxonomic specialists can browse the online collection of images and data, find specimens relevant to their research, and request specimen loans and/or aliquots of extracted genomic DNA. With the participation of a broad network of contributors and taxonomists, the cyberdiversity approach can even improve the description rate of the undescribed portion of our global fauna [51]. In an era of biodiversity crisis, climate change, and other challenges, the scientific and public spheres have common interest in synergies that make research products more responsive to the questions of the day. Thus, practices that make it easier to compare and combine data across different inventory studies are highly desirable for deriving the maximum information value from our research investment. We encourage authors who include specimens and DNA extracts featured on cyberdiversity platforms in their research to follow an open access cybertaxonomic publication model [52, 53]. All data shown on digitalSpiders are protected by a creative commons license, meaning they can be used for third party research provided the original source is cited and derivative works are distributed according to a similar license. A healthy ethic of data sharing can advance research across the community. We ask authors to think carefully about the benefits they derive from shared data and

acknowledge accordingly those responsible if they hope to foster this incipient trend.

Sampling strategy and assessment

During structured inventory surveys, a selection of semi-quantitative field sampling methods is typically applied to estimate spider species richness [36, 39, 54–67]. Each sampling method in a structured inventory targets a different portion of the fauna, which may overlap to a greater or lesser extent with other methods. Given that inventories have limited resources, decisions about how to partition effort among the sampling methods can have a major impact on the ultimate completeness of the inventory. To assess and improve sampling design, Cardoso [38] developed a method for optimizing allocation of sampling effort for spider inventories based on data from three studies conducted in Portugal. His method is based on a *post hoc* randomization analysis of inventory data to determine the optimum allocation of sampling effort by method to maximize species encountered.

Based on results from previous tropical forest inventory studies [39, 68], our structured inventory sampling strategy allocated more time to AEN than any other single collecting method. Cobra analysis [38] confirmed that AEN was the most efficient method for species discovery (Fig. 6). WIN, which is used in standard inventories of some groups other than spiders [69], was also found to be an efficient method for spiders. These results are in contrast to those found for Portuguese spiders [38], where pitfall traps and sweeping were the most efficient methods for species accrual and AEN was less crucial. However, direct comparison between Cardoso [38] and our inventories is complicated due to mismatches in sampling methods: WIN was not included in the Portuguese inventories and neither sweeping nor pitfalls were included in our Vietnamese inventories. Sweeping would seem *a priori* to be a dubious investment in the generally thorny and herb-poor Vietnamese forests, and pitfalls were omitted because of our regrettable failure to locally obtain propylene glycol preservative, which facilitates DNA extraction from pitfall specimens [70]. Note that the protocol presented by Cardoso [38] is not intended as a global standard and it is acknowledged that efficient sampling in other regions, including tropical forests, probably requires a different sampling strategy.

DNA barcodes for species discrimination

Within the BOLD framework, DNA barcode sequences are assigned to species based on cluster analysis and empirically derived interspecific distance thresholds (BIN [49]). The distance threshold for animals is typically set at 2–3% sequence divergence [71, 72]. Barrett and Hebert [73] reported that a 2% divergence threshold was adequate for discriminating spider species. We found that 2–3% divergence was indeed sufficient to assign most barcodes to species. However, conspecific distances were considerably higher for a few species, especially species
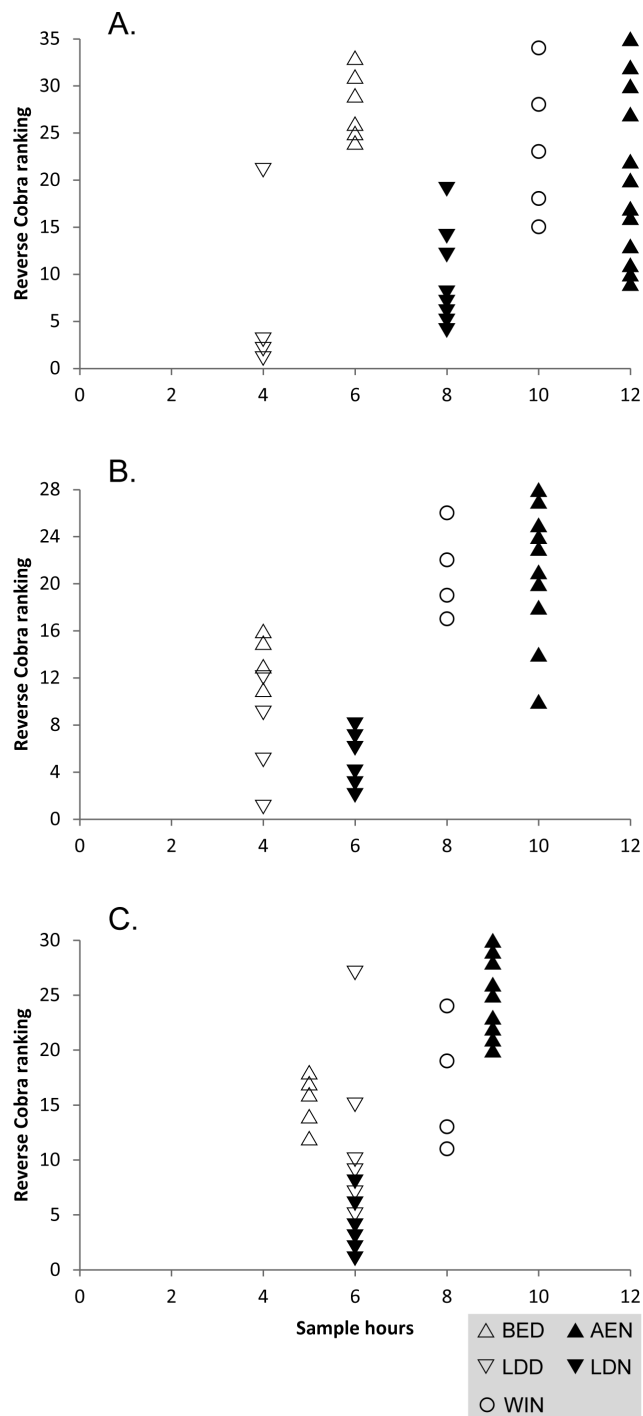
Fig. 6. Efficiency of sampling methods compared to hours of field time allocated. Sampling efficiency was scored using the reverse Cobra ranking (see text) for (A) Cat Ba, (B) Cuc Phuong, and (C) Vu Quang. Methods: BED: beating vegetation during the day; AEN: searching for spiders in the aerial stratum at night; LDD: searching for spiders on the ground during the day; LDN: searching for spiders on the ground at night; WIN: extraction of arthropods from sifted leaf litter using Winkler traps. Methods conducted during daylight hours are unfilled, methods conducted during night hours are black-filled. Methods targeting above-ground strata are up-pointing triangles, methods targeting ground strata are down-pointing triangles or a circle (WIN). Each WIN sample required two field hours; samples using all other samples were one hour each. High

reversed Cobra scores indicate maximum efficiency of a method for contributing species to the inventory; low scores indicate sampling saturation for that method.

shared between sites (Figs. 4A and S1A, C in S1 File). There are several possible explanations for this, including (a) taxonomic error (some IOTUs in this study may actually represent more than one species), (b) inadvertent amplification of a nuclear pseudogene of mitochondrial origin [18], and (c) bacterial infection (e.g., *Wolbachia*), which can distort patterns of mitochondrial variation and inheritance [74]. Alternatively, high conspecific genetic distances may simply reflect variable intra-species divergence, with most species characterized by small (e.g., <2%) divergences and a few species characterized by higher divergences (Figs. 4A and S1A, C in S1 File). Species sorting based on integrated analysis of DNA and morphology (IOTUs) found fewer species than the algorithmic approach [49] based on sequence data alone. Based on IOTU designations and barcode sequences, minimum genetic distances between species were almost always larger than distances within species (Figs. 4B and S1B, D in S1 File). The incongruence between DNA-only (BIN) and the integrated approach (IOTU) is almost certainly attributable to the many rare species that characterize our inventories; without a reasonable estimate of within-species variation, genetic gaps between species can be obscure [75]. The system of IOTUs presented here is subject to testing and refinement by future studies. Thus, an additional advantage of the cyberdiversity approach is the ability to readily integrate new data with legacy data to test and refine the findings of previous studies. Both IOTU designations and DNA-only BIN codes [49] are included in the supporting information to highlight incongruence between the two approaches and facilitate future re-assessment (S1 Appendix).

Modeling community change

The proportion of shared species is a useful and intuitive concept for comparing two communities. When inventories are nearly complete, the Jaccard index expresses this adequately. But when sampling is incomplete, especially in communities of megadiverse taxa with large proportions of rare species, the Jaccard can underestimate the proportion of shared species [76–78]. Shared rare species present the most significant challenge because they are most likely to be missed in one or both inventories. To account for unobserved shared species [76, 79], we replace the observed values in the Jaccard with estimated ones (see methods). We use the ACE [80] to estimate richness of each community because the shared species estimator is an extension of the ACE [77, 78]. We call this Chao's estimated proportional similarity. Note that combining estimated values in this way can inflate the variance of the resulting community similarity estimate [79, 81]. Chao's estimated proportional similarity should not be confused with Chao's abundance-based Jaccard index [76, 79], which does not report an estimated proportion of shared species (unless abundances of all species are

equal). One reason why we concentrate on presence-absence (as opposed to including relative abundance data) is that inventories of spiders and other diverse arthropods typically employ an assortment of field sampling methods, each targeted to a subset of the fauna. Thus, the relative abundances of species sampled in this way are not expected to be representative of their actual abundances within communities [82]; random sampling of individuals from ecologically and morphologically diverse species communities in structurally complex habitats is usually not realistic. While species rarity is useful in the context of non-parametric species richness and shared species estimators, diversity measures that rely on community species abundance distributions (e.g., Chao's abundance-based Jaccard index) may be problematic in this context.

Chao's estimated proportional similarity is significantly correlated with both geographic and climatic distances (mantel tests; $p < 0.05$). Nevertheless, it is clear that our understanding of regional and comparative β-diversity will greatly profit from an expanded dataset. In addition, future analyses involving more reconciled inventories should use variance partitioning to disentangle contributions of distance and climate to β-diversity pattern [48]. While acknowledging the preliminary nature of this analysis, to our knowledge this is the first quantification of the geographic rate of species turnover for tropical spider communities. Future analyses can also test for differences in the root causes of biodiversity structuring across landscapes. For example, data from angiosperms suggest stronger relationships between community similarity and geographic distance (as opposed to climate differences), suggesting a biodiversity pattern largely shaped by dispersal ability and climate history, particularly the Pleistocene glaciation of North America and Europe [48, 83, 84]. We are curious if spiders and other megadiverse arthropod groups with comparatively narrow environmental tolerances and fast generation times follow a similar pattern, or if historical relicts such as glaciations are more quickly obliterated in such communities.

Similarly, there is an enduring interest in distinguishing the root causes of the spatial structuring of biodiversity (β-diversity). Concordance between community composition and environmental conditions suggest biodiversity structuring driven by niche-sorting, while concordance with geographic distance invokes dispersal abilities, landscape characteristics, and neutrality [45, 46, 85–90]. Whether the β-diversity patterns of spiders and other megadiverse arthropods accord or contrast with other communities of organisms, increased availability and analysis of data based on diverse taxa with fine-grained climatic and spatial structuring will contribute to a richer understanding of these fundamental biological questions. In addition, a more complete understanding of the spatial structure of megadiverse communities, including their geographic, climatic, and latitudinal components, can improve the targeting of conservation priorities and provide quantitative estimates of biodiversity structuring across the globe [91]. The cyberdiversity approach will help us to realize the full complement of scientific and conservation benefits offered by structured inventories of megadiverse taxa in a rapid and rigorous manner.

Methods

Ethics statement

All necessary permits were obtained for the described study, which complied with all relevant regulations. A specimen collecting permit was granted by the Vietnam Administration of Forestry. All samples were collected in national parks. No protected species were involved in this study. An export permit to allow sample processing was granted by the Institute of Ecology and Biological Resources, Vietnam Academy of Science and Technology. Specimens have been divided between the Institute of Ecology and Biological Resources in Hanoi and the Naturalis Biodiversity Center in Leiden in accordance with an agreement made prior to the expedition (S1 Appendix).

Sampling and processing

Spiders were sampled from one-hectare plots in forest habitat [36, 39, 68] established in three Vietnamese national parks: Vu Quang, Cuc Phuong, and Cat Ba. Five sampling methods were used: (1) beating vegetation during the day (BED), (2) searching for spiders in the aerial stratum at night (AEN), (3) searching for spiders on the ground during the day (LDD), (4) searching for spiders on the ground at night (LDN), and (5) extraction of arthropods from sifted leaf litter using Winkler traps (WIN; www.entowinkler.at). Searching and beating methods were conducted in one-hour blocks; leaf litter sifting was done in two hour blocks plus a minimum drying time of 48 hours. Allocation of samples by method for each of the three inventories is reported in Table 3.

After field sampling, adult spiders were roughly sorted to morphospecies. When there was any question as to whether particular specimens belonged to one or more morphospecies, they were initially treated as different. These cases were later re-examined in light of DNA sequence data (it is easier to merge data from multiple putative morphospecies into one than to partition a hodgepodge). Morphological and barcode data were reconciled to create a collection of integrated operational taxonomic units (IOTUs) [35]. One or more specimens of every IOTU was photographed, both sexes when available. Photographs were made using a Nikon DS-Ri1 camera mounted on a Leica M165 C stereoscope operated using NIS Elements software. Images from multiple focus planes were combined and edited in Syncroscopy Auto-Montage software version 5.03 (http://www.syncroscopy.com). Images (1877 from 532 specimens representing all 240 species) and associated collection data were uploaded to Morphbank (www.morphbank.net; S2 Appendix).

DNA Barcoding

Tissues from 1–4 legs were sent to the Naturalis DNA barcoding facility. Specimens for DNA barcoding were selected to represent both sexes of all species from all sites, as available. All DNA voucher specimens were photographed. Extractions were performed using either the Qiagen DNEasy Blood and Tissue kit

or the Macherey-Nagel NucleoMag Tissue kit (http://www.mn-net.com/) on the Thermo Labsystems KingFisher extraction robot.

Following initial tests using a variety of primer combinations, PCR was performed using the primers LCO1490 (5′-GGTCAACAAATCATAAAGATATTGG-3′) [92] and Chelicerate Reverse 2 (5′-GGATGGCCAAAAAATCAAAATAAATG-3′) [73]. PCR reactions contained 18.75 µl mQ, 2.5 µ 10× PCR buffer CL, 1.0 µl 25 mM of each primer, 0.5 µl 2.5 mM dNTPs and 0.25 µl 5 U Qiagen Taq. PCR was performed using initial denaturation of 180 s at 94℃, followed by 40 cycles of 15 s at 94℃, 30 s at 50℃ and 40 s at 72℃, finished with a final extension of 300 s at 72℃ and pause at 12℃. Sequencing was performed by Macrogen (http://www.macrogen.com). For all barcoded specimens, sequences and collection data were uploaded to the Barcode of Life Database (BOLD; http://www.boldsystems.org/; S3 Appendix).

DNA barcode sequences were aligned using ClustalW [93] with default parameters as implemented in DAMBE [94, 95]. Neighbor joining distances were calculated in DAMBE with 10,000 replicates of random terminal input order. Three different models were applied: Kimura 2-parameter [96], because of its widespread use in the DNA barcoding literature [71], uncorrected p-distances [97], and the optimal model as determined using the Akaike information criterion to evaluate models implemented in jModelTest (version 0.1.1) [98].

For all three models, the barcode gap was expressed as the minimum interspecific distance against the maximum intraspecific distance [97] and also the pair-wise intraspecific sequence distances within sites and between sites.

Sampling strategy assessment

We used the sampling optimization method described by Cardoso [38, 99] to assess our sampling design efficiency. Inventory data from each of the three Vietnamese sites was analyzed using Cobra with 1000 randomizations. This program estimates the order of samples by method that will produce the greatest number of species with the least sampling effort. To compare the optimized sampling strategy to the actual allocation of field hours, we reversed the order of samples (so efficient samples received high values) and plotted this reverse Cobra ranking against the number of actual field hours devoted to each method at each site. High points on the reverse Cobra ranking score indicate the sampling of maximum efficiency for contributing new species to an inventory; low points indicate sampling saturation (i.e., inefficiency for discovering new species). Efficient, unsaturated methods warrant more resources during field sampling. Note that each Winkler sample was based on two hours of daylight field time spent sifting leaf litter, so each Winkler sample was counted as two hours.

Biodiversity analysis

Species richness for each site was estimated according to two abundance-based (Chao 1 and ACE) and two incidence-based (Chao 2 and ICE) non-parametric

estimators [80, 100–103] as implemented in EstimateS [104]. In all cases, the classic (not the bias corrected) formula was used following the *post hoc* recommendations given by EstimateS. Ninety-five percent confidence intervals for Chao 1 and Chao 2 were also calculated [100]. Community similarity between sites was calculated as:

$$\frac{V_{12(est)}}{(ACE_1 + ACE_2 - V_{12(est)})}$$

where $V_{12(est)}$ is Chao's estimated shared species between community 1 and community 2 (for communities where species discovery probabilities are heterogeneous; substituting Chao $1_{-shared}$ or Chao $1_{-shared-bias\ corrected}$ as appropriate) [77, 105], and ACE_1 and ACE_2 are the abundance-based coverage estimates of species richness in community 1 and community 2 [80] as implemented in SPADE [106]. This estimated proportional similarity is modeled after the Jaccard index:

$$\frac{S_{12}}{(S_1 + S_2 - S_{12})}$$

where S_1 and S_2 are the observed species richnesses in communities 1 and 2, and S_{12} is the observed number of shared species between communities. Thus, Chao's estimated proportional similarity is equivalent in composition to the Jaccard except it replaces empirical counts with non-parametric estimates of those values. Upper and lower bounds of the 95% confidence interval for $V_{12(est)}$ was used to estimate variability of the similarity measure.

We modeled β-diversity as an exponential decay of community similarity (species turnover) across (great circle or climatic) difference [45, 46, 48, 83, 107–109] using the distance-decay model:

$$s = s_0 e^{-\beta d}$$

where *s* is the community similarity (estimated here using Chao's estimated proportional similarity), *d* is the distance between sites, and s_0 (initial similarity, when *d*=0) and β (decay constant) are modeled parameters. While it is common for distance-decay studies to estimate s_0 and β using a linear regression of log(*s*) on distance, our dataset includes pairs of sites that have no species in common (*s*=0), which results in an undefined portion of the regression (log(0) is not defined). While it is possible to add a small value to these zero-points during the calculation (or remove such comparisons from the model) [45, 46, 48, 83], these manipulations result in meaningful changes to resulting estimates of s_0 and β [107]. Because community similarities range between 0 and 1, to include *s*=0 site pairs, we, instead, modeled distance-decay as binomial proportions using a generalized linear model with a log link function [107].

By parameterizing our exponential diversity-decay model, we gain two pieces of biological insight: (1) an estimate of the similarity of two samples taken from, essentially, the same locality (s_0; when *d*=0), which provides some indication of sample

completeness, underlying diversity, and habitat heterogeneity, and (2) the distance across which community similarity decays by half, (the "halving distance", $d_{0.5}$):

$$\frac{d_{0.5} = -\log(0.5)}{\beta}$$

Because of the nature of exponential curves, this halving distance is independent of placement along the curve. Thus, we gain an estimate of a fundamental characteristic of species turnover along any stretch of space or climate across the sampled region.

Because our calculations are conducted on paired-comparisons, traditional estimates of model fit and standard errors of parameter estimates are not valid. Thus, we use a jackknife approach [107] to calculate the standard error of parameter estimates. Our jackknife successively removes each of our n sites (not simply site-pairs) and re-runs our analyses n - 1 times. The variance of the parameters is then calculated as the total sum-of-squares divided by n jackknifed values multiplied by $(n$ - $1)/n$ [107, 110]. Finally, we use mantel tests (using 10,000 permutations) to calculate the significance of relationships between community similarity and distance (geographic or climatic).

Climate data were obtained from the 30 arc second rasters of 19 bioclimatic variables and a digital elevation model from WorldClim (http://www.worldclim. org/current, generic global grids, version 1.4 [47]). While we are fundamentally interested in testing how changes in climate and elevation influence patterns of β-diversity, our current sample size is too small to exhaustively explore their impacts. Additionally, the available 19 BioClim climatic variables are highly autocorrelated (86% of the 190 pair-wise climate and elevation comparisons from Southeast Asian sites have a Pearson correlation greater than 0.5, 60% have correlations greater than or equal to 0.8). To summarize the available climate data into fewer variables, we performed a principal component analysis (PCA; Table S1 in S1 File). Climate data were log-transformed, mean-centered, and scaled prior to analysis. To calculate climatic distances among sites, we then calculated a Euclidean distance matrix using the scores from the first j principal component axes that cumulatively summarize more than 95% of the variance. Because geographic and climatic distance are expressed in different units, direct comparisons of $d_{0.5}$ are difficult, although comparisons of s_0 between climate and geographic datasets are not affected. To compare community change, we calculate the estimated community change between the two furthest points. That is, we calculate how many $d_{0.5}$'s will have occurred across the sampled landscape and estimate that proportional change in overall spider community as:

$$1 - 0.5\char`\^[\frac{d_{max}}{d_{0.5}}]$$

where d_{max} is the maximum distance (geographic or climatic) observed by any paired site comparison. Community distance decay and associated analyses were scripted in R [111].

Supporting Information

S1 File. This file contains Table S1 and Figure S1. Figure S1, Discriminatory power of DNA barcodes under alternative models. A, C, within-species distances ranked by magnitude and partitioned into distances between individuals sampled from the same site (red) and distances between individuals sampled from different sites (blue). B, D, the barcode gap expressed as the maximum within-species distance against the minimum between-species distance. Distance models (based on IOTU classification): A, B, Kimura 2-parameter; C, D, uncorrected p. Table S1, Results of PCA analysis of environmental data (WorldClim) derived from the three Vietnamese and one Thai inventory sites. Variable loadings on the first two principal components (which cumulatively explain 99% of the variance) are also provided.

doi:10.1371/journal.pone.0115750.s001 (DOC)

S1 Appendix. Primary specimen occurrence data. Includes catalog numbers, IOTU codes (taxonID), DNA barcode identification numbers (BOLD BIN), specimen location (institutionID), and complete specimen-by-sample data for this study. Fields follow Darwin Core standards (http://rs.tdwg.org/dwc/) where applicable.

doi:10.1371/journal.pone.0115750.s002 (XLSX)

S2 Appendix. Images available on Morphbank (www.morphbank.net).
doi:10.1371/journal.pone.0115750.s003 (XLSX)

S3 Appendix. DNA barcode sequences available on BOLD (www.boldsystems.org/).
doi:10.1371/journal.pone.0115750.s004 (XLSX)

Acknowledgments

Sciences) for generously sharing their unpublished Doi Inthanon species abundance data, and for taking the initiative to post pictures of their morphospecies online. Thanks to Grace Woo (New Year Designs) for her major contributions to the web site, and to David Kavanaugh (California Academy of Sciences) for encouragement and critical discussions. We extend our appreciation to Martín Ramírez, Facundo Labarque, Peter Michalik (Universität Greifswald), an anonymous reviewer, and especially Pedro Cardoso (University of Helsinki) for constructive comments on manuscript drafts. Finally, we thank, and to Jonathan Coddington (Smithsonian Institution), Wayne Maddison (University of British Columbia), Anne Chao (National Tsing Hua University), Simon Rycroft and Laurence Livermore (The Natural History Museum, London) for helpful advice and discussion.

Author Contributions

References

1. **Butchart SHM, Walpole M, Collen B, van Strien A, Scharlemann JPW, et al.** (2010) Global biodiversity: Indicators of recent declines. Science 328: 1164–1168.

2. **Barnosky AD, Matzke N, Tomiya S, Wogan GOU, Swartz B, et al.** (2011) Has the Earth's sixth mass extinction already arrived? Nature 471: 51–57.

3. **Gaston KJ** (1994) Spatial patterns of species description: how is our knowledge of the global insect fauna growing? Biological Conservation 67: 37–40.

4. **Gaston KJ, May RM** (1992) Taxonomy of taxonomists. Nature 356: 281–282.

5. **Ferrier S, Gray MR, Cassis GA, Wilkie L** (1999) Spatial turnover in species composition of ground-dwelling arthropods, vertebrates and vascular plants in north-east New South Wales: implications for selection of forest reserves. In:, Ponder W, Lunney D, , editors., The Other 99% The Conservation and Biodiversity of Invertebrates. Mosman: Transactions of the Royal Zoological Society of New South Wales. pp. 68–76.

6. **Moritz C, Richardson KS, Ferrier S, Monteith GB, Stanisic J, et al.** (2001) Biogeographical concordance and efficiency of taxon indicators for establishing conservation priority in a tropical rainforest biota. Proceedings of the Royal Society of London, B 268: 1875–1881.

7. **D'Amen M, Bombi P, Campanaro A, Zapponi L, Bologna MA, et al.** (2013) Protected areas and insect conservation: questioning the effectiveness of Natura 2000 network for saproxylic beetles in Italy. Animal Conservation 16: 370–378.

8. **Margules CR, Pressey RL** (2000) Systematic conservation planning. Nature 405: 243–253.

9. **Ferrier S, Powell GVN, Richardson KS, Manion G, Overton JM, et al.** (2004) Mapping more of terrestrial biodiversity for global conservation assessment. BioScience 54: 1101–1109.

10. **Cardoso P, Erwin TL, Borges PAV, New TR** (2011) The seven impediments in invertebrate conservation and how to overcome them. Biological Conservation 144: 2647–2655.

11. **Riedel A, Sagata K, Ruhardjono YR, Tänzler R, Balke M** (2013) Integrative taxonomy on the fast track - towards more sustainablility in biodiversity research. Frontiers in Zoology 10: 15.

12. **Wheeler QD** (2008) The New Taxonomy. Boca Raton: The Systematics Association Special Volume Series 76, CRC Press. 237 p.

13. **Maddison DR, Guralnick R, Hill A, Reysenbach A-L, McDade LA** (2012) Ramping up biodiversity discovery via online quantum contributions. Trends in Ecology and Evolution 27: 72–77.

14. **Schindel DE, Miller SE** (2010) Provisional nomenclature: the on-ramp to taxonomic names. In: Polaszek A, editor. Systema Naturae 250: The Linnaean Ark. Boca Raton: CRC Press. pp. 109–115.

15. **Longino JT, Colwell RK** (2011) Density compensation, species composition, and richness of ants on a neotropical elevational gradient. Ecosphere 2: art29.

16. **Ratnasingham S, Hebert PDN** (2007) BOLD: the barcode of life data system (www.barcodinglife.org). Molecular Ecology Notes 7: 355–364.

17. **Hebert PDN, deWaard JR, Landry J-F** (2009) DNA barcodes for 1/1000 of the animal kingdom. Biology Letters.

18. **Song H, Buhay JE, Whiting MF, Crandall KA** (2008) Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. Proceedings of the National Academy of Sciences 105: 13486–13491.

19. **Will KW, Mishler BD, Wheeler QD** (2005) The perils of DNA barcoding and the need for integrative taxonomy. Systematic Biology 54: 844–851.

20. **Will KW, Rubinoff D** (2004) Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. Cladistics 20: 47–55.

21. **Sperling F** (2003) DNA barcoding. Deux et machina. Newsletter of the Biological Survey of Canada (Terrestiral Arthropods) 22: http://www.biology.ualberta.ca/bsc/news22_22/opinionpage.htm.

22. **Dayrat B** (2005) Towards integrative taxonomy. Biological Journal of the Linnean Society 85: 407–415.

23. **Meyer CP, Paulay G** (2005) DNA barcoding: error rates based on comprehensive sampling. PLOS Biology 3: e422.

24. **Rubinoff D** (2006) Utility of mitochondrial DNA barcodes in species conservation. Conservation Biology 20: 1026–1033.

25. **Rubinoff D, Holland BS** (2005) Between two extremes: mitochondrial DNA is neither the panacea nor the nemesis of phylogenetic and taxonomic inference. Systematic Biology 54: 952–961.

26. **Rubinoff D, Cameron S, Will K** (2006) A genomic perspective on the shortcomings of mitochondrial DNA for "barcoding" identification. Journal of Heredity 97: 581–594.

27. **Schindel DE, Miller SE** (2005) DNA barcoding a useful tool for taxonomists. Nature 435: 17.

28. **Frézal L, Leblois R** (2008) Four years of DNA barcoding: current advances and prospects. Infection, Genetics and Evolution 8: 727–736.

29. **Mitchell A** (2008) DNA barcoding demystified. Australian Journal of Entomology 47: 169–173.

30. **Moritz C, Cicero C** (2004) DNA barcoding: promise and pitfalls. PLOS Biology 2: e354.

31. **Paquin P, Hedin MC** (2004) The power and perils of "molecular taxonomy": A case study of eyeless and endangered *Cicurina* (Araneae: Dictynidae) from Texas caves. Molecular Ecology 13: 3239–3255.

32. **Goldstein PA, DeSalle R** (2011) Integrating DNA barcode data and taxonomic practice: determination, discovery, and description. Bioessays 33: 135–147.

33. **Riedel A, Sagata K, Surbakti S, Tänzler R, Balke M** (2013) One hundred and one new species of *Trigonopterus* weevils from New Guinea. ZooKeys 280: 1–150.

34. **Meier R** (2008) DNA sequences in taxonomy opportunities and challenges. In: Wheeler QD, editor. The New Taxonomy. Boca Raton: CRC Press. pp. 95–127.

35. **Galimberti A, Spada M, Russo D, Mucedda M, Agnelli P, et al.** (2012) Integrated operational taxonomic units (IOTUs) in echolocating bats: a bridge between molecular and traditional taxonomy. PLOS One 7: e40122.

36. **Coddington JA, Griswold CE, Silva Dávila D, Peñaranda E, Larcher SF** (1991) Designing and testing sampling protocols to estimate biodiversity in tropical ecosystems. In: Dudley EC, editor. The Unity of Evolutionary Biology: Proceedings of the Fourth International Congress of Systematic and Evolutionary Biology. Portland, OR.: Dioscorides Press. pp. 44–60.

37. **Kremen C, Colwell RK, Erwin TL, Murphy DD, Noss RF, et al.** (1995) Terrestrial arthropod assemblages: their use in conservation planning. Conservation Biology 7: 796–808.

38. **Cardoso P** (2009) Standardization and optimization of arthropod inventories—the case of Iberian spiders. Biodiversity and Conservation 18: 3949–3962.

39. **Coddington JA, Agnarsson I, Miller JA, Kuntner M, Hormiga G** (2009) Undersampling bias: the null hypothesis for singleton species in tropical arthropod surveys. Journal of Animal Ecology 78: 573–584.

40. **Longino JT, Coddington JA, Colwell RK** (2002) The ant fauna of a tropical rian forest: estimating species richness three different ways. Ecology 83: 689–702.

41. **Erwin TL, Pimienta MC, Murillo OE, Aschero V** (2005) Mapping patterns of beta diversity for beetler across the western Amazon Basin: a preliminary case for improving inventory methods and conservation strategies. Proceedings of the California Academy of Sciences 56: 72–85.

42. **Peterson AT, Knapp S, Guralnick R, Soberón J, Holder MT** (2010) The big questions in biodiversity informatics. Systematics and Biodiversity 8: 159–168.

43. **Smith VS** (2015) E-publishing descriptive taxonomy: the convergence of taxonomic journals and databases. In:, Watson M, Lyal C, , editors., Descriptive Taxonomy: The Foundation of Biodiversity Research Preprint Available: http://vsmithinfo/files/Smith-ePublishing-2010pdf: Cambridge University Press. (In press)

44. **Thessen AE, Patterson DJ** (2011) Data issues in the life sciences. ZooKeys 105: 15–51.

45. **Nekola JC, White PS** (1999) The distance decay of similarity in biogeography and ecology. Journal of Biogeography 26: 867–878.

46. **Soininen J, McDonald R, Hillebrand H** (2007) The distance decay of similarity in ecological communities. Ecogeography 30: 3–12.

47. **Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A** (2005) Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965–1978.

48. **Qian H, Ricklefs RE** (2007) A latitudinal gradient in large-scale beta diversity for vascular plants in North America. Ecology Letters 10: 737–744.

49. **Ratnasingham S, Hebert PDN** (2013) A DNA-based registry for all animal species: the barcode index number (BIN) system. PLOS One 8: e66213.

50. **Felsenstein J** (1984) Distance methods for inferring phylogenies: a justification. Evolution 38: 16–24.

51. **Fontaine B, Perrard A, Bouchet P** (2012) 21 years of shelf life between discovery and description of new species. Current Biology 22: R943–R944.

52. **Miller JA, Griswold CE, Yin C-M** (2009) The symphytognathoid spiders of the Gaoligongshan, Yunnan, China (Araneae: Araneoidea): Systematics and diversity of micro-orbweavers. ZooKeys 11: 9–195.

53. **Miller JA, Griswold CE, Haddad CR** (2010) Taxonomic revision of the spider family Penestomidae (Araneae, Entelegynae). Zootaxa 2534: 1–36.

54. **Scharff N, Coddington JA, Griswold CE, Hormiga G, Bjørn PdP** (2003) When to Quit? Estimating species richness in a northern European deciduous forest. Journal of Arachnology 31: 246–273.

55. **Coddington JA, Young LH, Coyle FA** (1996) Estimating spider species richness in a southern Appalachian cove hardwood forest. Journal of Arachnology 24: 111–128.

56. **Silva Dávila D, Coddington JA** (1996) Spiders of Pakitza (Madre de Dios, Perú): Species richness and notes on community structure. In: Wilson DE, Sandoval A, editors. The Biodiversity of Southeastern Perú: Smithsonian Institution. pp. 253–311.

57. **Cardoso P, Gaspar C, Pereira LC, Silva I, Henriques SS, et al.** (2008) Assessing spider species richness and composition in Mediterranean cork oak forests. Acta Oecologica 33: 114–127.

58. **Cardoso P, Scharff N, Gaspar C, Henriques SS, Carvalho R, et al.** (2008) Rapid biodiversity assessment of spiders (Araneae) using semi-quantitative sampling: a case study in a Mediterranean forest. Insect Conservation and Diversity 1: 71–84.

59. **Cardoso P, Henriques SS, Gaspar C, Crespo LC, Carvalho R, et al.** (2008) Species richness and composition assessment of spiders in a Mediterranean scrubland. Journal of Insect Conservation.

60. **Toti DS, Coyle FA, Miller JA** (2000) A structured inventory of Appalachian grass bald and heath bald spider assemblages and a test of species richness estimator performance. Journal of Arachnology 28: 329–345.

61. **Dobyns JR** (1997) Effects of sampling intensity on the collection of spider (Araneae) species and the estimation of species richness. Environmental Entomology 26: 150–162.

62. **Cabra-García J, Chacón P, Valderrama-Ardila C** (2010) Additive partitioning of spider diversity in a fragmented tropical dry forest (Valle del Cauca, Colombia). Journal of Arachnology 38: 192–205.

63. **Rego FNAA, Venticinque EM, Brescovit AD, Rheims CA, Albernaz ALKM** (2009) A contribution to the knowledge of the spider fauna (Arachnida: Araneae) of the floodplain forests of the main Amazon River channel. Revista Ibérica de Aracnología 17: 85–96.

64. **Höfer H** (1990) The spider community (Araneae) of a Central Amazonian blackwater inundation forest (igapó). Acta Zoologica Fennica 190: 173–179.

65. **Bonaldo AB, Dias SC** (2010) A structured inventory of spiders (Arachnida, Araneae) in natural and artificial forest gaps at Porto Urucu, Western Brazilian Amazonia. Acta Amazonica 40: 357–372.

66. **Kapoor V** (2006) An assessment of spider sampling methods in tropical rainforest fragments of the Anamalai Hills, Western Ghats, India. Zoos' Print Journal 21: 2483–2488.

67. **Pinkus-Rendón M, León-Cortés JL, Ibarra-Núñez G** (2006) Spider diversity in a tropical habitat gradient in Chiapas, Mexico. Diversity and Distributions 12: 61–69.

68. **Sørensen LL, Coddington JA, Scharff N** (2002) Inventorying and estimating subcanopy spider diversity using semiquantitative sampling methods in an afromontane forest. Pest Management and Sampling 31: 319–330.

69. **Agosti D, Alonso LE** (2000) The ALL protocol—a standard protocol for the collection of ground-dwelling ants. In: Agosti D, Majer JD, Schultz TR, editors. Ants–standard methods for measuring and monitoring biodiversity. Washington: Smithsonian Institution Press. pp. 204–206.

70. **Vink CJ, Thomas SM, Paquin P, Hayashi CY, Hedin M** (2005) The effects of preservatives and temperatures on arachnid DNA. Invertebrate Systematics 19: 99–104.

71. **Hebert PDN, Cywinska A, Ball SL, deWaard JR** (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society of London, B 270: 313–321.

72. **Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM** (2004) Identification of birds through DNA barcodes. PLOS Biology 2: 1657–1663.

73. **Barrett RDH, Hebert PDN** (2005) Identifying spiders through DNA barcodes. Canadian Journal of Zoology 83: 481–491.

74. **Smith MA, Bertrand C, Crosby K, Eveleigh ES, Fernandez-Triana J, et al.** (2012) *Wolbachia* and DNA barcoding in insects: patterns, potential, and problems. PLOS One 7: e36514.

75. **Lim GS, Balke M, Meier R** (2012) Determining species boundaries in a world full of rarity: singletons, species delimitation methods. Systematic Biology 61: 165–169.

76. **Chao A, Chazdon RL, Colwell RK, Shen T-J** (2006) Abundance-based similarity indices and their estimation when there are unseen species in samples. Biometrics 62: 361–371.

77. **Chao A, Hwang W-H, Chen YC, Kuo CY** (2000) Estimating the number of shared species in two communities. Statistica Sinica 10: 227–246.

78. **Magurran AE** (2004) Measuring Biological Diversity: Blackwell Publishing. 256 p.

79. **Chao A, Chazdon RL, Colwell RK, Shen T-J** (2005) A new statistical approach for assessing similarity of species composition with incidence and abundance data. Ecology Letters 8: 148–159.

80. **Chao A, Lee S-M** (1992) Estimating the number of classes via sample coverage. Journal of the American Statistical Association 87: 210–217.

81. **Gotelli NJ, Chao A** (2013) Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. In: Levin SA, editor. Encyclopedia of Biodiversity, second edition, Volume 5. Waltham, MA: Academimc Press. pp. 195–211.

82. **Longino JT, Colwell RK** (1997) Biodiversity assessment using structured inventory: Capturing the ant fauna of a tropical rain forest. Ecological Applications 7: 1263–1277.

83. **Qian H, Ricklefs RE, White PS** (2005) Beta diversity of angiosperms in temperate floras of eastern Asia and eastern North America. Ecology Letters 8: 15–22.

84. **Fitzpatrick MC, Sanders NJ, Normand S, Svenning J-C, Ferrier S, et al.** (2013) Environmental and historical imprints on beta diversity: insights from variation in rates of species turnover along gradients. Proceedings of the Royal Society of London, B 280: 20131201.

85. **Hubbell SP** (2001) The Unified Neutral Theory of Biodiversity and Biography: Princeton University Press. 375 p.

86. **Condit R, Pitman N, Leigh EG, Chave J, Terborgh J, et al.** (2002) Beta-diversity in tropical forest trees. Science 295: 666–669.

87. **Gilbert B, Lechowicz MJ** (2004) Neutrality, niches, and dispersal in a temperate forest understory. Proceedings of the National Academy of Sciences 101: 7651–7656.

88. **Rominger AJ, Miller TEX, Collins SL** (2009) Relative contributions of neutral and niche-based processes to the structure of a desert grassland grasshopper community. Oecologia 161: 791–800.

89. **Novotny V, Weiblen GD** (2005) From communities to continents: beta diversity of herbivorous insects. Annales Zoologici Fennici 42: 463–475.

90. **Novotny V, Miller SE, Hulcr J, Drew RAI, Basset Y, et al.** (2007) Low beta diversity of herbivorous insects in tropical forests. Nature 448: 692–695.

91. **McKnight MW, White PS, McDonald RI, Lamoreux JF, Sechrest W, et al.** (2007) Putting beta-diversity on the map: broad-scale congruence and coincidence in the extremes. PLOS Biology 5: e272.

92. **Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R** (1994) DNA primers for the amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Molecular Marine Biology and Biotechnology 3: 294–299.

93. **Thompson JD, Higgins DG, Gibson TJ** (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Research 22: 4673–4680.

94. **Xia X, Xie Z** (2001) DAMBE: Data analysis in molecular biology and evolution. Journal of Heredity 92: 371–373.

95. **Xia X** (2000) Data Analysis in Molecular Biology and Evolution. Boston: Kluwer Academic Publishers. 276 p.

96. **Kimura M** (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of mucleotide sequences. Journal of Molecular Evolution 16: 111–120.

97. **Srivathsan A, Meier R** (2011) In the inappropriate use of Kimura-2-parameter (K2P) divergences in the DNA-barcoding literature. Cladistics 27: 1–5.

98. **Posada D** (2008) jModelTest: phylogenetic model averaging. Molecular Biology and Evolution 25: 1253–1256.

99. **Cardoso P** (2009) Cobra 1.0: Available: http://www.ennor.org/pro_software.html.

100. **Chao A** (1987) Estimating the population size for capture-recapture data with unequal catchability. Biometrics 43: 783–791.

101. **Chao A** (1984) Non-parametric estimation of the number of classes in a population. Scandinavian Journal of Statistics 11: 265–270.

102. **Colwell RK, Coddington JA** (1994) Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society, London 345.

103. **Lee S-M, Chao A** (1994) Estimating population size via sample coverage for closed capture-recapture models. Biometrics 50: 88–97.

104. **Colwell RK** (2009) EstimateS: Statistical estimation of species richness and shared species from samples. Version 8.2. Computer program and documentation. Available: http://viceroy.eeb.uconn.edu/EstimateS.

105. **Chao A, Shen T-J, Hwang W-H** (2006) Application of Laplace's boundary-mode approximations to estimate species and shared species richness. Australian & New Zealand Journal of Statistics 48: 117–128.

106. **Chao A, Shen T-J** (2009) Spade. Available: http://chao.stat.nthu.edu.tw/.

107. **Millar RB, Anderson MJ, Tolimieri N** (2011) Much ado about nothings: using zero similarity points in distance-decay curves. Ecology 92: 1717–1722.

108. **Whittaker RH** (1972) Evolution and measurement of species diversity. Taxon 21: 213–251.

109. **Whittaker RH** (1960) Vegetation of the Siskiyou Mountains, Oregon and California. Ecological Monographs 30: 270–338.

110. **Chernik MR** (2008) Bootstrap methods: a guide for practitioners and researchers. Second edition. Hoboken, New Jersey: Wiley.

111. **R_Development_Core_Team** (2011) R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

112. **Agnarsson I** (2003) The phylogenetic placement and circumscription of the genus *Synotaxus* (Araneae: Synotaxidae), a new species from Guyana, and notes on theridioid phylogeny. Invertebrate Systematics 17: 719–734.

113. **Grismado CJ** (2002) Palpiomanid spiders from Guyana: new species of the genera Fernandezina and Otiothops (Araneae, Plapimanidae, Otiothopinae). Iheringia Série Zoologia 92: 13–16.