A combined field survey and molecular identification protocol for comparing forest arthropod biodiversity across spatial scales

Brent C. Emerson[1,2]*, Juliane Casquet[3], Heriberto López[1], Pedro Cardoso[4,5], Paulo A. V. Borges[5], Noémy Mollaret[6], Pedro Oromí[7], Dominique Strasberg[8], and Christophe Thébaud[3]

[1]Island Ecology and Evolution Research Group, Instituto de Productos Naturales y Agrobiología, C/Astrofísico Francisco Sánchez 3, La Laguna, Tenerife, Canary Islands, 38206, Spain

[2]School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK.

[3]Laboratoire Evolution & Diversité Biologique, UMR 5174 CNRS-Université Paul Sabatier-ENFA, 31062 Toulouse Cedex 9, France

[4]Finnish Museum of Natural History, University of Helsinki, P.O. Box 17, 00014 Helsinki, Finland

[5]CE3C – Centre for Ecology, Evolution and Environmental Changes / Azorean Biodiversity Group and Universidade dos Açores –Departamento de Ciências e Engenharia do Ambiente, Rua Capitão João d'Ávila s/n, 9700-042, Angra do Heroísmo, Açores, Portugal

[6]Direction des collections, Invertébrés marins, Muséum National d'Histoire Naturelle, CP53, 61 rue Buffon, 75231 Paris cedex 05, France.

[7]Departamento de Biología Animal (Zoología), Universidad de La Laguna, C/ Astrofísico Francisco Sánchez, 38206 La Laguna, Tenerife, Canary Islands, Spain.

[8]UMR PVBMT, Peuplements Végétaux et Bio-agresseurs en Milieu Tropical, Université de La Réunion, 15 avenue René Cassin, CS 93002, 97 744 Saint Denis, Cedex 9, Reunion Island, France.

* Contact author: bemerson@ipna.csic.es

Abstract

Obtaining fundamental biodiversity metrics such as alpha, beta and gamma diversity for arthropods is often complicated by a lack of prior taxonomic information and/or taxonomic expertise, which can result in unreliable morphologically based estimates. We provide a set of standardized ecological and molecular sampling protocols that can be employed by researchers whose taxonomic skills may be limited, and where there may be a lack of robust *a priori* information regarding the regional pool of species. These protocols combine mass sampling of arthropods, classification of samples into parataxonomic units (PUs), and selective sampling of individuals for mtDNA sequencing to infer biological species. We sampled ten lowland rainforest plots located on the volcanic oceanic island of Réunion (Mascarene archipelago) for spiders, a group with limited taxonomic and distributional data for this region. We classified adults and juveniles into PUs and then demonstrated the reconciliation of these units with presumed biological species using mtDNA sequence data, ecological data and distributional data. Because our species assignment protocol is not reliant upon prior taxonomic information, or taxonomic expertise, it minimises the problem of the Linnean shortfall to yield diversity estimates that can be directly compared across independent studies. Field sampling can be extended to other arthropod groups and habitats by adapting our field sampling protocol accordingly.

Introduction

Understanding the drivers of arthropod biodiversity in most terrestrial ecosystems remains an important frontier and challenge in ecological and evolutionary research (Cardoso *et al.* 2011). Despite the fact that most terrestrial organisms are arthropods, fundamental biodiversity metrics such as alpha, beta and gamma diversity remain elusive at the level of higher-order communities for several key reasons. Perhaps the primary obstacle is simply one of scale - there are so many arthropod species, and at any given point on Earth they may encompass a complex diversity of form and function. The logistical challenges of total species richness estimation is exemplified by a recent and comprehensive sampling program undertaken in a tropical forest,

where twelve 20 x 20 m plots within an area of several km^2 were sampled. Using structured protocols, and a total of 24,354 person days sampling in the San Lorenzo forest of Panama, Basset *et al.* (2012) collected an estimated 6144 arthropod species.

Exhaustive sampling programs such as that of Basset *et al.* (2012) represent a tremendous step forward in filling the knowledge void of arthropod biodiversity. However, resource requirements in terms of taxonomic expertise, field and lab person hours, and financial backing limit the extent to which such sampling can be expanded across larger spatial scales. A lack of standardised and accessible approaches to inventory arthropod species within an area is an obstacle to reliable comparisons of arthropod species richness and diversity in an era of major global changes and potential species loss. In response to this, Rapid Biodiversity Assessment (RBA) programs that use minimal time and resources have been developed to gather as much information as possible from an area (e.g. Duelli 1997; Jones & Eggleton 2000; Oliver & Beattie 1996a) usually requiring a combination of methods (e.g. Scharff *et al.* 2003; Snyder *et al.* 2006) that require standardisation for comparison among sites (e.g. Borges *et al.* 2005; Duelli *et al.* 1999). The challenge for effective RBA programs is to seek to optimise data return and reliability against investment of time and resources.

The spatially explicit quantification of local arthropod species richness and diversity remains a formidable challenge, especially when sufficiently complete regional species lists and taxonomy are lacking, and when taxonomic expertise is not readily available. This taxonomic impediment (the so-called "Linnean shortfall") presents itself in most studies that aim to investigate arthropod species richness and diversity. The level of impediment will vary depending on geographic region and taxonomic groups, and becomes more problematic with apparently high levels of cryptic species diversity, characteristic of some arthropod groups (e.g. Cicconardi *et al.* 2013; Cicconardi *et al.* 2010). Overcoming the taxonomic impediment can thus be crucial to minimize the broader biodiversity impediment that affects the accuracy and precision, and ultimately the validity, of any biodiversity assessment. In the absence of a meaningful and reliable taxonomic framework, species identification can be incorporated into the inventory process by the designation of morphospecies (e.g. Basset *et al.* 2012; Novotny *et al.* 2002). Parataxonomists can reduce the taxonomic workload by collecting specimens and building reference collections, the units of

which have been referred to as morphospecies (e.g. Basset *et al.* 2000; Oliver & Beattie 1996b), but may best be referred to as parataxonomic units (PUs, Krell 2004). PUs are estimations of species as discrete morphological entities, based on external morphology only without considering taxonomy. Such PUs then require subsequent species confirmation by a taxonomist, a task that can be challenging to accomplish.

It has been shown that parataxonomic accuracy can deviate substantially with regard to taxonomic identity within the Arthropoda, with an average of 76 percent accuracy (which is a consequence of both splitting and lumping processes in PU assignment) across eight published comparisons (Krell 2004). Thus, while a parataxonomic approach helps to alleviate the taxonomic impediment, taxonomic expertise is still necessary to revise the taxonomic significance of PUs, which may still amount to a substantial task. If one accepts that access to a good taxonomist is a bottleneck, but at the same time important, then anything that optimizes a taxonomist's workload, or reduces their time investment within a biodiversity assessment, will be an advantage. To this end we propose an approach using a molecular step to deal, in an integrative way, with large scale arthropod biodiversity assessment to further address the taxonomic impediment especially in species-rich or poorly documented regions. We highlight four important advantages that this may offer:

(1) Validation or refinement of parataxonomic units as presumed species
While there are valid criticisms against reliance on simplistic single sequence approaches to species discovery, such as the "barcoding gene" mitochondrial cytochrome c oxidase subunit I (COI), sensu Hebert *et al.* (2003) (e.g. Goldstein & DeSalle 2011; Krishnamurthy & Francis 2012), used judiciously such approaches can take us further for the comparative quantification of biodiversity over different spatial scales (e.g. Emerson *et al.* 2011; Ji *et al.* 2013; Yu *et al.* 2012), and provide a useful complement for the validation of species boundaries (e.g. Riedel *et al.* 2013). In the case of PUs, DNA sequence variation may segregate with a PU, consistent with the taxonomic distinctiveness of that PU from other such units. Alternatively two or more PUs may be undiagnosable at the DNA sequence level, suggesting they may belong to the same taxonomic entity (e.g. phenotypic plasticity). However, in case of two or

more species being lumped within a single PU (e.g. cryptic species), a strategy involving the sequencing of multiple individuals within a PU will be required.

(2) Detection of cryptic species hypotheses

The description of cryptic species has grown exponentially over the past two decades (Scheffers *et al.* 2012), largely due to increased population level sampling of DNA sequence data. Subsequently it has been suggested that undescribed cryptic species may account for more species than are currently described (Ceballos & Ehrlich 2009). Cryptic species are morphologically difficult or impossible to distinguish among, and in some cases the number of undescribed cryptic species could be an order of magnitude higher than described species richness, as has been recently proposed for the arthropod class Collembola (Cicconardi *et al.* 2013). Although the problem of cryptic biodiversity in the context of biodiversity assessments is usually not considered (Ceballos & Ehrlich 2009), it would lead to inaccurate measures of alpha, beta and gamma diversity because cryptic species contribute to unobserved heterogeneity that cannot be easily accounted for. Such cryptic diversity can however be revealed by complementing the sorting of PUs with representative sampling of their DNA sequence data, thus improving estimation and inference about the underlying components of biodiversity patterns.

(3) Improved comparability

Incorporating DNA sequence data in biodiversity assessment increases the potential for biodiversity data to be more immediately available for comparative purposes. Even if a subset of inventoried species await formal taxonomic description or identification, their reference DNA sequence data can be made publicly available via databases such as GenBank (Benson *et al.* 2015) and the Barcode of Life Database (BOLD) (Ratnasingham & Hebert 2007) for subsequent inventory work and specimen identification. Related to this, such databases can also be used to augment taxonomic identification within a biodiversity assessment if sampled species are already represented in a database.

(4) Association of different arthropod life history stages and genders

One of the successes of combining DNA sequence data with traditional taxonomy has been the association of life stages or genders where not all may be described because of either large or limited morphological variation (Krishnamurthy & Francis 2012). DNA sequence data has been used successfully to identify cryptic larval or juvenile stages in arthropod species including mayflies (Ball *et al.* 2005), butterflies (Gossner & Hausmann 2009), beetles (Garcia-Robledo *et al.* 2013), flies (Pramual & Wongpakam 2014) and spiders (Raso *et al.* 2014). As such, where PUs are comprised of juvenile or larval forms, they can be associated with adult forms by reference to their DNA sequences. Differences among or within gender among adult forms may also confound species identity, particularly if samples sizes are small. However, PUs resulting from splitting and thus comprised of different forms within the same species will be united as a single presumed species based on DNA sequence similarity (e.g. Glowska *et al.* 2014; Jusoh *et al.* 2014; Przybylowicz & Tarcz 2015).

In this paper, we extend and develop a semi-quantitative protocol for the analysis of Mediterranean spiders (Cardoso 2009), incorporating a DNA sequence-based assessment of PUs for the inference of presumed biological species (Mayr 1942). We provide a set of field and laboratory protocols that can be employed by researchers whose taxonomic skills may be limited, where landscapes may be challenging for the purpose of sampling, and where there may be a lack of robust *a priori* information regarding the regional pool of species. Our broad aim is to combine standardized ecological and molecular sampling protocols to address the taxonomic impediment within biodiversity assessment. We demonstrate the value of our approach with a biodiversity assessment of the spider fauna from the volcanic oceanic island of Réunion (Mascarene archipelago; Southwest Indian Ocean), sampled from ten biodiversity plots established within lowland rainforests along the eastern coast of the island. This is a challenging focal group that has been the subject of only limited prior investigation in this island (Casquet *et al.* 2015) and for which both the species pool and taxonomy are not well known. In general, spider biodiversity assessments are forced to ignore juvenile forms because they usually cannot be assigned to species, meaning up to 75 percent of spider samples are discarded (e.g. Cardoso *et al.* 2009; Cardoso *et al.* 2008b). Our protocol allows us to include both adult and juvenile

spiders to estimate alpha, beta and gamma diversity. We compare these values to similar estimates based only on adult forms to illustrate the benefits of our approach relative to biodiversity assessment based on morphospecies only.

Materials and methods

Protocol overview

Our protocol integrates standardised plot-based field sampling with subsequent parataxonomic sample handling and DNA sequencing for the inference of biological species boundaries (Fig. 1). We provide a template for data entry and curation to facilitate the cross referencing of individual specimens with field sampling data, parataxonomic labelling and DNA sequence data for downstream biodiversity analyses (Appendix S1). We sequence the mitochondrial DNA (mtDNA) COI barcode region (Hebert *et al.* 2003) to facilitate taxonomic assignment using the specimen identification database within the Barcode of Life Database (BOLD) (Ratnasingham & Hebert 2007). It should be noted that the barcode region could be substituted for an alternative, but similarly informative gene region, and this may be preferable for some taxonomic groups, should amplification be more successful. However, this would preclude the use of BOLD to facilitate taxonomic assignment.

Sampling procedures

Field sampling methods are adapted from Cardoso *et al.* (2008a; 2009; 2008b), with the incorporation of additional sampling techniques designed to optimise the capture of both spider and beetle species. Full details of sampling methods are presented in Appendix S2. An important difference between the sampling design of Cardoso *et al.* (2008a; 2009; 2008b) and the present study is sampling area. Here we define a sampling area of 0.25 hectare (50m x 50m), which is intermediate between the 0.04 ha sampling area of Basset *et al.* (2012) and the 1 ha sampling area of Cardoso *et al.* (2008a; 2009; 2008b). Our choice of sampling area is a compromise among several constraints. Larger sampling areas such as those of Cardoso *et al.* (2008a; 2009; 2008b) may be limiting if topography is complex and habitat is disturbed, as is often

the case within rugged mountain landscapes. Smaller sampling areas such as those of Basset *et al.* (2012), are likely to be too small if the sampling effort is reasonable as the sites may get easily trampled and fauna depleted. Ten sampling areas (plots) were established within lowland rainforest within an altitudinal range of 570 - 790 m (see Strasberg *et al.* 2005 for a description of the habitat) along the eastern coast of the island of Réunion (Fig. 2). Plot locations were chosen based on accessible forest habitat that was deemed to be relatively undisturbed in terms of human impact and invasive plant species. Sampling was conducted between the 23/10/2012 and 14/01/2013.

Specimen sorting to parataxonomic units and genetic analyses

Sorting to spider PUs using morphological features, including epigynum and copulatory bulb shape, was carried out by one of us (JC) using a binocular microscope. Provisional taxonomic identifications were made to species level where possible, otherwise to genus or family level, using available taxonomic literature. The number of individuals to be sequenced per PU per plot is a decision to be based upon financial resources, but we recommend four individuals as a useful minimum. Choice of samples within each PU will also be based upon focal taxon and project objectives. In our case spiders may present problems for species identification due to gender and life history differences, with the majority of individuals sampled expected to be juveniles. To account for this, where possible, one adult male, one adult female and two juveniles for each PU within each plot were selected for DNA sequencing. In the absence of a sufficient number of individuals within one or more of these three classes, or for those PUs that were exclusively composed of either juvenile or adult forms, sampling was complemented with the available classes to obtain four individuals per PU per plot.

DNA extractions were carried out using a fast and cost effective chelex protocol using 96-well PCR plates (Casquet *et al.* 2012). In brief, single legs were incubated in PCR plate wells with 150 μl of a 10% Chelex® solution and 10 μl of a 10 mg ml^{-1} solution of Proteinase K. Plates were then sealed and incubated at 55 °C overnight, and then stored at -20 °C prior to DNA amplification. PCRs were conducted with the primers LCO1490 and HCO2198 (Folmer *et al.* 1994) using an Eppendorf Mastercycler ep thermocycler system in a 25 μl total PCR volume containing 16.3 μl of

purified water, 5 µl of Promega green 5x buffer, 0.5 µl of a 20 µM aliquot of each primer, 0.5 µl of 10 mM dNTP, 0.2 µl of 5 U/µl Taq polymerase and 2 µl of DNA extract. PCR was carried out under the following conditions: hot-start at 95 C for 2 min; repeated 38–40 cycles (denaturation at 95 C for 30 s, primer annealing at 55 C for 30 s and primer extension at 72 C for 40 s); and final extension at 72 C for 7 min. PCR success was then checked by running products on a 1% TAE agarose gel, and successfully amplified products were then outsourced for DNA sequencing to the French National Sequencing Centre, Genoscope (http://www.genoscope.cns.fr). All PCR products were Sanger sequenced with both the forward and reverse primers. All sequences were then checked and ambiguous bases called before constructing alignments for each PU using Geneious Pro, version 5.4 (http//www.geneious.com).

Reconciling parataxonomic units with presumed biological species

As a first step after DNA sequencing, DNA sequence alignments were constructed for each PU and checked for outlier sequences that exhibit high genetic divergence relative to other samples within the PU. Outlier sequences were first blasted locally against other PUs. In the case of a positive match, the specimen associated with the DNA sequence was assessed against the PU that the DNA sequence assigns to, to determine if the outlier DNA sequence was due to (i) misclassification of the sample to the wrong PU, or (ii) sample or laboratory procedural contamination resulting in the amplification and sequencing of non-target DNA. If no positive match was found with a local blast, the outlier sequence was submitted to the BOLD specimen identification database to determine the taxonomic origin of the sequence. In light of misclassifications identified by DNA sequencing, unsequenced samples within PUs that presented misclassified samples were reassessed. Subsequent to the analysis of, and correction for, outlier DNA sequences, representative DNA sequences sampled from within each PU were subject to a neighbour joining analysis using MEGA6 (Tamura *et al.* 2013) for the characterisation of genetic lineages (GL). GLs are defined as monophyletic groups of DNA sequences which aim to include all individuals of all PUs contained within a GL, while minimising the number of PUs within a GL. PUs represented by a single specimen are classified as a GL if they do not cluster within another GL. Samples within each GL were then assessed by one of us (PC) for

taxonomic assignment following the nomenclature provided by the World Spider Catalogue (http://wsc.nmbe.ch), and representative sequences were submitted to the BOLD Identification System.

We adopt the biological species concept (Mayr 1942) as our working species definition, with the simplest case being that all individuals within a GL belong to a single presumed biological species. We consider that divergent mtDNA lineages within a GL may belong to different biological species when they exceed the average of 6.77% Kimura 2-parameter (K2P) corrected divergence between species and their closest sampled related species, derived from an analysis of 361 spider species from 19 genera (Robinson *et al.* 2009). We evaluate the ecological or geographical independence of such lineages for the inference of presumed biological species using ecological and geographical sampling data for individuals, and their maximum likelihood haplotype relationships estimated using the K2P model with MEGA6. Our approach can be considered conservative, as it should lead to the underestimation, rather than overestimation, of biological species.

Alpha and beta diversity across plots

We used rarefaction analysis to compare observed species richness across all plots for: (i) PUs; (ii) presumed biological species with juveniles included, and; (iii) presumed biological species when only adult forms were included. A number of individuals smaller than the minimum abundance across plots (300) was used and confidence limits calculated as the 0.025 and 0.975 percentiles among 1000 resamples without replacement using R package BAT (Cardoso *et al.* 2015). The ranking of plots according to their observed richness was compared between the three approaches with Spearman correlation using the R package Hmisc (http://biostat.mc.vanderbilt.edu/wiki/Main/Hmisc). To understand if estimating species richness from each of the three datasets would improve their correlation we used the Chao abundance estimator (Chao 1984) in BAT and used Spearman correlation as above. Finally, we calculated the pairwise beta diversity components (βtotal, βrepl and βrich, see (Carvalho *et al.* 2012)) among all plots with the three datasets using BAT and their Spearman correlations with Mantel tests for significance using R package vegan (Oksanen *et al.* 2015).

Results

Specimen sorting and assignment to parataxonomic units

Across all ten plots a total of 5558 individual spiders were collected, of which 780 were adult male, 824 were adult female and 3954 (71%) were juvenile. The number of individuals per plot ranged between 372 and 935 (Table 1). Specimens were sorted into 148 PUs, of which 77 were comprised of both adult and juvenile forms, 38 were comprised only of adults, and 33 were comprised of only juveniles. For samples belonging to families for which there is published information about species diversity on Réunion (Clubionidae, Hahniidae, Miturgidae, Nephilidae, Salticidae, Sparassidae, Thomisidae), we compared PUs to genus and species descriptions, assigning PUs to taxonomic species when possible. For PUs assigned to families for which there is no published information for species diversity on Réunion, we liaised with local taxonomists who were able to taxonomically place PUs from seven families (Araneidae, Gnaphosidae, Linyphiidae, Pholcidae, Theridiidae, Uloboridae, Theridiosomatidae). No reliable information was available for the further taxonomic placement of PUs assigned to the following families: Anapidae, Corinnidae, Cyatholipidae, Lycosidae, Mysmenidae, Nesticidae, Oonopidae, Symphognathidae, Tetrablemmidae.

DNA sequencing

A total of 1588 individuals were chosen for DNA extraction and PCR amplification, representing 604 adult individuals and 984 juveniles, and an average of 2.6 individuals per PU per plot. A first round of PCR resulted in the successful amplification of 800 individuals (50,4%), with a further 596 (37,5%) amplifying with a second PCR attempt, and another 126 individuals with a third attempt. We performed a second extraction on the 66 individuals that did not successfully amplify after three PCR attempts. A first round of PCR (following re-extraction) resulted in the successful amplification of 30 individuals. Thus a total of 1554 out of 1588 individuals (97,8%) were submitted for DNA sequencing, of which 1171 individuals (75,3%) yielded good DNA sequences for both the forward and reverse primers, 84 sequenced well only for the forward primer, 66 sequenced well only for the reverse primer, and 233 yielded no sequence. Overall, DNA sequences were obtained for 1350 (85%) of the 1588 individuals chosen for DNA extraction, representing 123 (83%) of the 148

PUs. After re-examining all specimens with outlier sequences within each PU, and reconciling these with either sequencing error or incorrect PU assignment, 61 individuals were excluded due to sequences attributed to contamination or laboratory error. A further 121 individuals were reassigned to a different PU, resulting in the removal of 8 PUs (i.e. all individuals within these PUs reassigned to alternative PUs), yielding a total of 1289 individuals with DNA sequences, assigned to 115 PUs.

Reconciling parataxonomic units with presumed biological species, and taxonomic assignment

A total of 54 GLs were delineated from the 115 PUs, each comprised of from one to eight PUs (Table 2). Of the 115 PUs, all but one (PU-93) were assigned to a single GL. PU-93 comprises 13 sequenced specimens, from which nine specimens form GL-46, while the remaining four specimens grouped together with PUs 25, 27, 48, 70, 94 and 100 in GL-44. Across the 54 GLs, three presented genetic divergences in excess of Robinson *et al.*'s (2009) 6.77% K2P corrected average divergence between closest congeneric species. The first of these, GL-01, comprises three divergent genetic lineages (GL-01a, GL-01b and GL-01c) that show no segregation with the 7 PUs that comprise GL-01. However individuals of GL-01a were obtained from a broad range of sampling methods (FBN, AAS, FSN, BLM, TWS, LIT and PIT), while samples of GL-01b and GL-01c were almost exclusively obtained with only the LIT and PIT sampling methods. This difference suggests GL-01a represents a lineage of broader ecological niche than GL-01b and GL-01c, and we consider it to be a probable biological species. While not differing in terms of sampling techniques, GL-01b and GL01c are geographically discrete, with all 61 individuals of GL-01b sampled from plots 4-10, and all 14 individuals of GL-01c sampled from plots 1-3 (Fig. 3), and we treat each as a probable biological species. The second GL with divergences exceeding the threshold is GL-47, comprised of two divergent lineages (GL-47a and GL-47b) that show no segregation with the 6 PUs that comprise GL-47, and no difference in terms of ecological sampling. While they do not exhibit obviously different distributions, they do present different juvenile to adult ratios (80% of the GL-47a are adults, compared to only 38% for GL-47b), and we therefore infer both lineages to represent probable biological species. GL-12 is the third GL exceeding the threshold, with three divergent

lineages. However, in the absence of any clear ecological or geographical differentiation among these lineages, we treat GL-12 as a single presumed biological species. Representative sequences from each of the 57 presumed biological species are available on GenBank (KX762052-KX762107), and full alignments for each GL are available on Dryad (doi:10.5061/dryad.8g5d2).

Taxonomic assignment to species level was possible for 18 of the 57 inferred biological species, with a further two tentative species level assignments, and the remaining 36 assigned to genus or family level (Table 2). Sequences from seven species submitted to the BOLD specimen identification database yielded matches of 99% or higher. The lack of matching for many species is not surprising, given the limited number of studies that include spider species from Indian Ocean islands. Six of the seven matches were consistent with our taxonomic assignment, four being sequences previously obtained by one of us (JC) in a separate study (*Tetragnatha* sp., *Argyrodes* sp., *A. borbonicus* and *A. minax*), and a further two from published studies including *Arachnura scorpionoides* from Madagascar (Cheng & Kuntner 2014) and *Nephila inaurata* from several Indian Ocean islands (Kuntner & Agnarsson 2011). The final match reveals GL-27 to be *Nesticella mogera* (Zhang & Li 2013), a species of Asian origin that is known to have been introduced to Europe, but until now unrecorded from Réunion. Three families (Symphognathidae, Tetrablemmidae and Trachelidae), seven genera (*Nesticella*, *Parasteatoda*, *Patu*, *Prasonicella*, *Tetrablema*, *Theridiosoma*, *Trachela*), and at least 11 species (all but one unidentified) are recorded for the first time in Réunion.

Alpha and beta diversity across plots

At the individual level, it was not possible to assign unsequenced individuals to the presumed cryptic biological species GL-01a, GL-01b, GL-01c and GL-47a, GL47b. We thus estimated the number of a given presumed cryptic species in a given sampling plot as the proportion of unsequenced individuals equal to the proportion of sequenced individuals for that cryptic species within the plot. For each cryptic species this was done for all individuals, and for adult forms only.

The richness values obtained by the three approaches are significantly different for all ten plots (Fig. 4). Nevertheless, values are highly and significantly correlated, indicating that the ranking of plots is similar ($0.74 < r < 0.75$; $0.013 < p <$

0.014 in all cases). Estimating species richness for each approach worsened all correlations ($0.44 < r < 0.67$; $0.033 < p < 0.200$ in all cases). Beta diversity values were usually correlated between PUs and species using all individuals or the latter and species using adults only, but not between PUs and species using adults only (Table 3).

Discussion

Our sampling yielded 5558 spiders that were assigned to 151 parataxonomic units, for which DNA sequence data was obtained for 116, which were then collapsed to 54 genetic lineages, inferred to represent 57 presumed biological species. While it is clear that our PUs represent a poor approximation of biological species, for the most part they do represent collections of individuals that are ultimately inferred to belong to the same biological species. With the exception of GL-01 and GL-53, each of which is inferred to comprise more than one biological species, individuals from all but one PU (PU-93, see details above), were grouped together in the same presumed biological species. Thus, for the most part, our PUs represent partitions of life history and gender variation within species, and morphological variation within these. The danger of extrapolating from PUs to morphospecies has previously been noted (Krell 2004), pointing out that taxonomic accuracy is compromised. In the case of the spiders from Réunion we know that our PUs are an overestimate of the number of biological species, but in the absence of a robust a priori taxonomic framework, the challenge is to identify by how much. Our use of a DNA sequence step to identify the relatedness of PUs overcomes this challenge, providing an objective criterion to extrapolate presumed biological species from PUs. It is worth pointing out that this approach is insensitive to the definition of individual PUs, as long as they do not contain individuals from more than one species. Thus a conservative approach to defining PUs is advocated.

By not restricting our sampling to only adult forms we were able to increase our sample size from 1604 to 5558 individuals. While this only resulted in a marginal increase in the number of presumed biological species sampled across all plots (only 3 of 57 were comprised only of juveniles), it resulted in substantially higher species richness estimates within plots (Fig. 4). Additionally, while it is perhaps unsurprising

that beta diversity estimates derived from only adults samples are correlated with estimates from adult and juvenile samples, the correlation is generally weak (Table 3), indicating that expanding sampling to include juveniles will yield more realistic beta diversity estimates and may prove critical for elucidating ecological mechanisms that can shape arthropod community structure.

Cryptic species and cryptic diversity

A useful consequence of incorporating a DNA step into the delimitation of presumed biological species is that it has the potential to also reveal otherwise undetectable variation, such as the geographic structuring of genetic variation within species, or cryptic species. Much has been debated about the merit of inferring boundaries among species based on mtDNA divergence values (e.g. Taylor & Harris 2012), however, support for the existence of cryptic species is improved when additional sources of variation co-segregate with divergent mtDNA lineages (Kress *et al.* 2015), and our sampling protocol enables this to be explored. Sampling across the ten biodiversity plots revealed evidence for cryptic species, based on the application of Robinson *et al.*'s (2009) average of 6.77% K2P corrected divergence between closest congeneric species as a threshold. In the case of GL-01, differences in ecological sampling data are consistent with mtDNA divergences. While 92% of the 61 sequenced individuals assigned to either GL-01b or GL-01c were captured in pitfall traps and leaf litter, 61% of the 59 sequenced GL-01a individuals were captured with other sampling techniques, consistent with a broader ecological niche for GL-01a that includes subcanopy habitat.

While not differentiated by ecological sampling, GL-01b and GL-01c are geographically vicariant. GL-01b was sampled in the seven plots north of Piton de la Fournaise, while GL-01c is restricted to the three plots further south, suggesting that the extensive open lava fields that extend from Piton de la Fournaise to the coast may be a barrier to dispersal (Fig. 3). This is consistent with the high βtotal and βrich values between plots 3 and 4 (Fig. 4). In addition to this vicariant pattern, mtDNA variation within GL-01b is marked by deep coalescent divergences of up to 4.8% (Fig. 3), indicating an absence of movement of mtDNA haplotypes from north to south of Piton de la Fournaise over a substantial period of evolutionary time. The mtDNA variation within GL-01b also exhibits geographic structuring among the seven

northern sampling plots (Fig. 3), supporting a hypothesis of dispersal limitation within this taxon. The geographical patterns of mtDNA structure among and within GL-01b and GL-01c could be argued to be the product of female dispersal limitation within a single species, as there is evidence for sex-biased dispersal in some spider species (Croucher *et al.* 2011; Gillespie & Oxford 1998). Thus our data is indicative of two species and dispersal limitation, but confirmation of this requires a more comprehensive population genetic analysis including nuclear markers.

It is becoming routine to associate immature life stages with their adult forms (Kress *et al.* 2015), and the incorporation of mtDNA sequencing into our protocol enabled us to associate PUs comprised solely of only adult or juvenile forms. An unexpected outcome from including juvenile life stages was that they can provide support for inferences of cryptic species. The two lineages within GL-53, while exceeding the proposed 6.77% K2P divergence threshold of Robinson *et al.* (2009), do not present any discernable ecological differentiation based on sampling techniques. However, support for the hypothesis that these divergent lineages represent different biological species is provided by their very different adult to juvenile ratios (80% adults for GL-53a vs 38% for GL-53b), indicating divergent phenologies associated with the two lineages.

The ability of our protocol to reveal presumed cryptic species diversity is demonstrated by the cases above. However, it should be noted that inferred species number may still be an underestimate, due to further undetected cryptic species diversity. Such undetected cryptic diversity is unlikely to be consequential for biodiversity analyses at the community scale, but may be of conservation and management interest for the identification of diversified lineages and endemic species. The mtDNA sequence data from our protocol provides added value by signalling GLs with high genetic divergence values (e.g. Table 2: GL-01b, GL-03, GL-11, GL-12, GL-17), that may be of interest for subsequent more detailed investigation.

New species records, and the diversity of Réunion spider fauna

The primary purpose of plot-based inventories, like the one we describe in this study, is not to simply improve taxonomic knowledge within a region. However, such inventories, when focused on target taxa with a strong taxonomic impediment, such as spiders, hold great potential to disclose new species records and undescribed taxa. In this study, using a combination of standardized ecological and molecular sampling protocols applied to samples collected over a total of 2.5 ha (i.e. 1/100,000 of the total area of Réunion), we were able to identify 11 new species records, five of which belonging to families that had not been recorded previously on the island, and many of which are likely to be undescribed species endemic to Réunion or to the Mascarenes (J. Casquet, pers. obs.). Improving coverage of taxonomic diversity is obviously important to describe accurately the biological diversity and the ecological complexity present within a region. More importantly perhaps, it also provides new opportunities to investigate the processes that produce biodiversity patterns, by providing the first step through which evolutionary history can be taken into account for comparing samples.

Conclusions

By applying the protocols developed in this study of Réunion spider communities we have been able to demonstrate that it is possible to reduce the taxonomic impediment (Linnean shortfall) to quantify arthropod species richness and diversity when information on regional species pools and taxonomic expertise is lacking. Combining standardised field sampling and parataxonomic sample sorting with objective estimation of species boundaries with mtDNA sequencing reduces the biodiversity impediment that otherwise impedes the accuracy, and ultimately the validity, of arthropod biodiversity assessments. Thus, we expect that our combined field survey and molecular identification protocol can be extended to other arthropod groups and habitats by adapting the field sampling protocol accordingly.

References

Ball SL, Hebert PDN, Burian SK, Webb JM (2005) Biological identifications of mayflies (Ephemeroptera) using DNA barcodes. *Journal of the North American Benthological Society* **24**, 508-524.

Basset Y, Cizek L, Cuenoud P*, et al.* (2012) Arthropod Diversity in a Tropical Forest. *Science* **338**, 1481-1484.

Basset Y, Novotny V, Miller SE, Pyle R (2000) Quantifying biodiversity: experience with parataxonomists and digital photography in Papua New Guinea and Guyana. *Bioscience* **50**, 899-908.

Benson DA, Clark C, Karsch-Mizrachi I*, et al.* (2015) GenBank. *Nucleic Acids Research* **43**, D30-D35.

Borges PAV, Aguiar C, Amaral J*, et al.* (2005) Ranking protected areas in the Azores using standardized sampling of soil epigean arthropods. *Biodiversity and Conservation* **14**, 2029-2060.

Cardoso P (2009) Standardization and optimization of arthropod inventories - the case of Iberian spiders. *Biodiversity and Conservation* **18**, 3949-3962.

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.

Cardoso P, Gaspar C, Pereira LC*, et al.* (2008a) Assessing spider species richness and composition in Mediterranean cork oak forests. *Acta Oecologica* **33**, 114-127.

Cardoso P, Henriques S, Gaspar C*, et al.* (2009) Species richness and composition assessment of spiders in a Mediterranean scrubland. *Journal of Insect Conservation* **13**, 45-55.

Cardoso P, Rigal F, Carvalho JC (2015) BAT - Biodiversity Assessment Tools, an R package for the measurement and estimation of alpha and beta taxon, phylogenetic and functional diversity. *Methods in Ecology and Evolution* **6**, 232-236.

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

Carvalho JC, Cardoso P, Gomes P (2012) Determining the relative roles of species replacement and species richness differences in generating beta-diversity estimates. *Global Ecology and Biogeography* **21**, 760-771.

Casquet J, Bourgeois YXC, Cruaud C*, et al.* (2015) Community assembly on remote islands: a comparison of Hawaiian and Mascarene spiders. *Journal of Biogeography* **42**, 39-50.

Casquet J, Thebaud C, Gillespie RG (2012) Chelex without boiling, a rapid and easy technique to obtain stable amplifiableDNA from small amounts of ethanol-stored spiders. *Molecular Ecology Resources* **12**, 136-141.

Ceballos G, Ehrlich PR (2009) Discoveries of new mammal species and their implications for conservation and ecosystem services. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 3841-3846.

Chao A (1984) Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics, Theory and Applications* **11**, 265-270.

Cheng RC, Kuntner M (2014) Phylogeny suggests nondirectional and isometric evolution of sexual size dimorphism in argiopine spiders. *Evolution* **68**, 2861-2872.

Cicconardi F, Fanciulli PP, Emerson BC (2013) Collembola, the biological species concept, and the underestimation of global species richness. *Molecular Ecology* **22**, 5382-5396.

Cicconardi F, Nardi F, Emerson BC, Frati F, Fanciulli PP (2010) Deep phylogeographic divisions and long-term persistence of forest invertebrates (Hexapoda: Collembola) in the North-Western Mediterranean basin. *Molecular Ecology* **19**, 386-400.

Croucher PJP, Oxford GS, Gillespie RG (2011) Population structure and dispersal in a patchy landscape: nuclear and mitochondrial markers reveal area effects in the spider *Theridion californicum* (Araneae: Theridiidae). *Biological Journal of the Linnean Society* **104**, 600-620.

Duelli P (1997) Biodiversity evaluation in agricultural landscapes: an approach at two different scales. *Agriculture, Ecosystems & Environment* **62**, 81-91.

Duelli P, Obrist MK, Schmatz DR (1999) Biodiversity evaluation in agricultural landscapes: above-ground insects. *Agriculture, Ecosystems & Environment* **74**, 33-64.

Emerson BC, Cicconardi F, Fanciulli PP, Shaw PJA (2011) Phylogeny, phylogeography, phylobetadiversity and the molecular analysis of biological communities. *Philosophical Transactions of the Royal Society B* **336**, 2391-2404.

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

Garcia-Robledo C, Kuprewicz EK, Staines CL, Kress WJ, Erwin TL (2013) Using a comprehensive DNA barcode library to detect novel egg and larval host plant associations in a Cephaloleia rolled-leaf beetle (Coleoptera: Chrysomelidae). *Biological Journal of the Linnean Society* **110**, 189-198.

Gillespie RG, Oxford GS (1998) Selection on the color polymorphism in Haiaiian happy-face spiders: evidence from genetic structure and temporal fluctuations. *Evolution* **52**, 775-783.

Glowska E, Dragun-Damian A, Brida L, Dabert J, Dabert M (2014) DNA barcodes reveal female dimorphism in syringophilid mites (Actinotrichida: Prostigmata:

Cheyletoidea): Stibarokris phoeniconaias and Ciconichenophilus phoeniconaias are conspecific. *Folia Parasitologica* **61**, 272-276.

Goldstein PZ, DeSalle R (2011) Integrating DNA barcode data and taxonomic practice: Determination, discovery, and description. *Bioessays* **33**, 135-147.

Gossner MM, Hausmann A (2009) DNA barcoding enables the identification of caterpillars feeding onnative and alien oak (Lepidoptera: geometridae). *Mitteilungen Muenchener Entomologischen Gesellschaf* **99**, 135-140.

Hebert PDN, Cywinska A, Ball SL, DeWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B* **270**.

Ji Y, Ashton L, Pedley S*, et al.* (2013) Reliable, verifiable, and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters* **16**, 1245-1257.

Jones DT, Eggleton P (2000) Sampling termite assemblages in tropical forests: testing a rapid biodiversity assessment protocol. *Journal of Applied Ecology* **37**, 191-203.

Jusoh WFA, Hashim NR, Saaksjarvi IE, Adam NA, Wahlberg N (2014) Species delineation of Malaysian mangrove fireflies (Coleoptera: Lampyridae) using DNA barcodes. *Coleopterists Bulletin* **68**, 703-711.

Krell F-T (2004) Parataxonomy vs. taxonomy in biodiversity studies – pitfalls and applicability of 'morphospecies' sorting. *Biodiversity and Conservation* **13**, 795-812.

Kress WJ, García-Robledo C, Uriarte M, Erikson DL (2015) DNA barcodes for ecology, evolution, and conservation. *Trends in Ecology & Evolution* **30**, 25-35.

Krishnamurthy PK, Francis RA (2012) A critical review on the utility of DNA barcoding in biodiversity conservation. *Biodiversity and Conservation* **21**, 1901-1919.

Kuntner M, Agnarsson I (2011) Phylogeography of a successful aerial disperser: the golden orb spider *Nephila* on Indian Ocean islands. *BMC Evolutionary Biology* **11**, 119.

Mayr E (1942) *Systematics and the origin of species from the viewpoint of a zoologist.* Harvard University Press, Cambridge, Massachusetts.

Novotny V, Basset Y, Miller SE*, et al.* (2002) Low host specificity of herbivorous insects in a tropical forest. *Nature* **416**, 841-844.

Oksanen J, Blanchet FG, Kindt R*, et al.* (2015) Vegan: Community Ecology Package. R package version 2.0.

Oliver I, Beattie AJ (1996a) Designing a cost-effective invertebrate survey: a test of methods for rapid assessment of biodiversity. *Ecological Applications* **6**, 594-607.

Oliver I, Beattie AJ (1996b) Invertebrate morphospecies as surrogates for species: a case study. *Conservation Biology* **10**, 99-109.

Pramual P, Wongpakam K (2014) Association of black fly (Diptera: Simuliidae) life stages using DNA barcode. *Journal of Asia-Pacific Entomology* **17**, 549-554.

Przybylowicz L, Tarcz S (2015) Strong sexual dimorphism unraveled by DNA analysis - towards a better understanding of Pseudothyretes classification (Lepidoptera: Erebidae: Arctiinae). *Zoological Journal of the Linnean Society* **173**, 22-54.

Raso L, Sint D, Rief A, Kaufmann R, Traugott M (2014) Molecular identification of adult and juvenile linyphiid and theridiid spiders in alpine glacier foreland communities. *PLoS ONE* **9(7)**, e101755.

Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System. *Molecular Ecology Notes* **7**, 355-364.

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

Robinson EA, Blagoev GA, Hebert PDN, Adamowicz SJ (2009) Prospects for using DNA barcoding to identify spiders in species rich genera. *ZooKeys* **16**, 27-46.

Scharff N, Coddington JA, Griswold Ce, Hormiga G, De Place Bjorn P (2003) When to quit? Estimating spider species richness in a northern European deciduous forest. *Journal of Arachnology* **31**, 246-273.

Scheffers BR, Joppa LN, Pimm SL, Laurance WF (2012) What we know and don't know about Earth's missing biodiversity. *Trends in Ecology & Evolution* **27**, 501-510.

Snyder BA, Draney ML, Sierwald P (2006) Development of an optimal sampling protocol for millipedes (Diplopoda). *Journal of Insect Conservation* **10**, 277-288.

Strasberg D, Rouget M, Richardson DM, *et al.* (2005) An assessment of habitat diversity and transformation on La Réunion Island (Mascarene Islands, Indian Ocean) as a basis for identifying broad-scale conservation priorities. *Biodiversity and Conservation* **14**, 3015-3032.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology and Evolution* **30**, 2725-2729.

Taylor HR, Harris WE (2012) An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Molecular Ecology Resources* **12**, 377-388.

Yu DW, Ji YQ, Emerson BC, *et al.* (2012) Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution* **3**, 613-623.

Zhang Y, Li S (2013) Ancient lineage, your troglobites: recent colonization of caves by *Nesticella* spiders. *BMC Evolutionary Biology* **13**, 183.

Author contributions: B.C.E, P.A.V.B, D.S., C.T. and P.O. designed the study. J.C. led the field collecting, sample sorting and classification of samples to PU. N.M. assisted J.C. with fieldwork and sample sorting. J.C. led the lab work and H.L led sequence editing and alignment. P.C. conducted alpha and beta diversity analyses, and together with J.C. undertook the taxonomic assignment of GLs. B.C.E analysed the data and wrote the paper, and all authors discussed the results and commented on the manuscript.

Table 1. Spider samples obtained across ten lowland rainforest plots on the island of Réunion.

Plot	Adult male	Adult female	Juvenile	Total
1	61	77	308	448
2	42	42	283	372
3	51	55	402	517
4	72	64	280	425
5	82	63	342	492
6	57	78	388	524
7	55	89	372	517
8	104	110	392	609
9	146	140	648	935
10	110	106	539	765

Table 2. Genetic lineages (GL) inferred for adult and juvenile spider samples after sample assignment to parataxonomic unit (PU) and mtDNA sequencing, and their taxonomic assignment. Indiv is the number of individuals assigned to each GL and PU indicates the number of PUs that comprise each GL. K2P indicates the maximum K2P genetic divergence within each GL.

GL	1	2	3	4	5	6	7	8	9	10	Taxonomic ID	Indiv.	PUs	K2P
1a	X	X	X	X		X	X	X	X	X	*Trachelas* (?) sp.	155	7	2.0
1b				X	X	X	X	X	X	X	*Trachelas* (?) sp.	109	7	4.8
1c	X	X	X								*Trachelas* (?) sp.	32	7	0.2
2				X			X				Lycosidae Gen. sp.	2	1	NA
3	X	X	X	X	X	X	X	X	X	X	Lycosidae Gen. sp.	200	2	6.0
4	X	X		X	X	X			X	X	*Ledouxia allauadi*	12	1	3.7
5					X						*Prepotelus curtis*	1	1	NA

#	1	2	3	4	5	6	7	8	9	10	Species			
6	X	X	X	X	X	X	X	X	X	X	*Prepotelus lanceolatus*	1042	2	0.8
7							X				*Xeropheaus* (?) sp.	2	1	NA
8	X	X	X	X	X	X	X	X	X	X	*Clubiona nemorum*	471	4	1.6
9	X	X	X	X	X	X	X	X	X	X	*Cheiracanthium insulare*	228	1	1.3
10		X					X				*Alistra mendenai* (?)	2	1	2.7
11			X	X	X	X		X	X		Theridiidae Gen. sp.	69	1	5.6
12	X	X			X	X	X	X	X	X	*Ulwembua* sp.	38	2	8.0
13			X								Theridiidae Gen. sp.	1	1	NA
14			X							X	Pholcidae Gen. sp.	3	1	0.0
15			X			X	X	X			*Tetrablema* sp.	8	1	NA
16	X	X	X	X	X	X	X	X	X	X	*Tylorida* sp.	374	3	4.3
17	X	X	X	X	X	X	X	X	X	X	*Tetragnatha* sp.	204	2	5.3
18				X		X		X	X		*Afroneta longipalpis*	19	2	0.2
19			X						X	X	*Zosis geniculata*	6	2	0.2
20			X			X	X			X	*Hahnia* sp.	6	2	0.9
21	X	X	X	X	X	X	X	X	X	X	*Lophostica nova*	387	8	1.1
22	X	X	X	X	X	X	X	X	X	X	*Pseudemathis trifida*	263	2	1.1
23	X	X	X	X			X		X	X	*Lophostica mauritiana*	65	1	0.5
24	X	X	X		X		X			X	*Miagrammopes* sp.	17	1	0.3
25			X		X		X				*Xerophaeus* sp.	4	2	1.1
26									X		Linyphiidae Gen. sp.	3	2	0.9
27	X	X		X			X		X	X	*Nesticella mogera*	27	3	0.7
28	X	X		X	X	X	X		X	X	Linyphiidae Gen. sp.	62	3	4.1
29	X										*Helsdingenia* sp.	3	1	0.2
30			X			X	X	X	X	X	*Theridiosoma* sp.	29	1	2.6

#	1	2	3	4	5	6	7	8	9	10	11	Species			
31	X	X	X	X	X	X	X	X	X	X	X	*Theridiosoma* sp.	343	1	2.7
32	X	X	X	X	X	X	X	X	X	X	X	*Neoscona triangula*	72	6	1.1
33						X	X	X	X		X	*Cyrtophora citricola*	10	1	1.1
34	X	X	X	X	X	X	X	X	X	X	X	*Prasonicella* sp.	84	3	1.7
35	X	X	X	X	X						X	*Poltys kochi* (?)	18	2	0.5
36	X		X		X			X				*Arachnura scorpionoides*	5	2	0.3
37		X										*Leucauge* sp.	7	1	0.3
38											X	Clubionidae Gen. sp.	9	1	0.2
39									X		X	*Patu* (?) sp.	4	1	0.3
40	X	X	X			X	X		X		X	*Nephilengys borbonica*	26	1	0.2
41	X	X	X			X	X	X	X		X	*Nephila inaurata*	28	1	0.6
42		X									X	*Cyrtarachne* sp.	6	2	1.3
43							X					*Platnicka* (?) sp.	1	1	NA
44	X	X	X	X	X	X	X	X	X	X	X	*Theridion* (?) sp.	121	7	2.7
45				X	X				X		X	*Theridion* (?) sp.	16	2	3.6
46						X		X	X			*Platnicka* (?) sp.	9	1	0.2
47a		X	X	X	X			X	X	X		*Chrysso* sp.	54	6	0.3
47b	X	X	X			X	X		X		X	*Chrysso* sp.	123	6	1.1
48	X		X				X	X	X		X	Theridiidae Gen. sp.	10	1	0.0
49	X	X	X	X	X	X	X	X	X	X	X	*Parasteatoda* sp.	210	3	1.4
50	X	X	X			X	X	X			X	*Thwaitesia* (?) sp.	65	4	1.0
51	X	X	X	X	X	X	X	X	X	X	X	*Rhomphaea* sp.	35	3	0.2
52		X	X	X	X	X	X	X	X	X	X	*Argyrodes* sp.	128	4	0.9
53	X	X	X		X		X					*Argyrodes borbonicus*	33	1	0.9
54	X	X	X	X	X	X	X	X	X	X	X	*Argyrodes minax*	239	1	0.8

Table 3. Spearman correlations between beta diversity values calculated using parataxonomic units (PUs) and species (including juveniles or adults only). βtotal = total beta diversity, βrepl = beta diversity due to species replacement, βrich = beta diversity due to differences in species richness.

Comparison	βtotal	βrepl	βrich
PUs x Species (all)	0.487**	0.528**	0.370
PUs x Species (adults)	0.093	0.243	0.370*
Species (all) x Species (adults)	0.490**	0.582***	0.444**

Significance (from Mantel tests): * p < 0.05, ** p < 0.01, *** p < 0.001

Figure legends

Figure 1. Flow chart summarising the main steps for the application of parataxonomic classification and mtDNA sequencing for the inference of biological species from large field collections of arthropod samples. See main text for details.

Figure 2. Locations of 50m x 50m sampling plots within lowland rainforest along the eastern coast of Réunion.

Figure 3. MtDNA sequence variation within GL-01b and GL-01c, presented as a maximum likelihood tree using the Kimura 2-parameter model of DNA sequence evolution. Bootstrap support values above 80 are shown. The scale bar represents 1% of mutational change. All individuals of GL-01b were sampled from the seven locations to the north of Piton de la Fournaise, while all individuals of GL-01c were sampled from the three locations to the south. GL-01b also exhibits geographical structuring of mtDNA sequence variation within its range, as represented by the differently coloured groupings. Taxon labels indicate sampling plot number and sample ID.

Figure 4. Species richness estimated using rarefaction across all plots for: (A) PUs; (B) presumed biological species with juveniles included, and; (C) presumed biological species when only adult forms were included. Confidence limits calculated as the 0.025 and 0.975 percentiles among 1000 resamples (see text). Pairwise beta diversity between neighbouring sampling plots estimated for presumed biological species with juveniles included (D).

```
                    ┌─────────────────┐
                    │ Field sampling  │
                    └────────┬────────┘
                             ▼
                  ┌─────────────────────┐
                  │ Sample sorting to   │
                  │ parataxonomic unit  │
                  │ (PU)                │
                  └──────────┬──────────┘
                             ▼
                  ┌─────────────────────┐
                  │ MtDNA sequencing of │
                  │ 4 individuals for   │
                  │ each PU             │
                  │ In each plot        │
                  └──────────┬──────────┘
                             ▼
                  ┌─────────────────────┐
                  │ Construct sequence  │
                  │ alignments for each │
                  │ PU                  │
                  └──────────┬──────────┘
```

Field sampling

Sample sorting to parataxonomic unit (PU)

MtDNA sequencing of 4 individuals for each PU In each plot

Construct sequence alignments for each PU

No outlier sequences

Outlier sequences

Local blast of outlier sequences → No match → Discard

match → Non-target sequence → Discard

Sample misidentification → Revise unsequenced samples of both PUs

New alignment

NJ analysis of samples from each PU

Identification of GLs

Assess divergence threshold

Single mtDNA lineage

Multiple mtDNA lineages

Evaluate mtDNA lineages for ecological or geographical consistency

Define presumed biological species

Taxonomic assignment of morphological material

BOLD specimen identification analysis of DNA sequences

A PUs

B Species (total data)

Plots

C Species (adults only)

D

	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10
βtotal	0.19	0.40	0.49	0.38	0.41	0.40	0.43	0.38	0.33
βrepl	0.17	0.27	0.18	0.16	0.34	0.29	0.36	0.33	0.22
βrich	0.03	0.13	0.30	0.21	0.07	0.12	0.07	0.05	0.11