European Journal of Taxonomy 272: 1–13
http://dx.doi.org/10.5852/ejt.2017.272

ISSN 2118-9773
www.europeanjournaloftaxonomy.eu
2017 · Ramage T. *et al.*

DNA Library of Life, research article

urn:lsid:zoobank.org:pub:8EC5F9FB-5A94-4FA6-8274-86326E8404B3

A DNA barcode-based survey of terrestrial arthropods in the Society Islands of French Polynesia: host diversity within the SymbioCode Project

Thibault RAMAGE [1], Patricia MARTINS-SIMOES [2], Gladys MIALDEA [3], Roland ALLEMAND [4,†], Anne DUPLOUY [5], Pascal ROUSSE [6], Neil DAVIES [7], George K. RODERICK [8] & Sylvain CHARLAT [9,*]

[1] 9 Quartier de la Glacière, 29900 Concarneau, France.
[2,3,4,9] Laboratoire de Biométrie & Biologie Evolutive, CNRS – Université Lyon 1, Bat. Mendel, 43 Boulevard du 11 November, 69622 Villeurbanne, France.
[2] CIRI, International Center for Infectiology Research, Lyon, France.
[5] University of Helsinki, Metapopulation Research Centre, Department of Biosciences, P.O. Box 65, Viikinkaari 1, 00014 Helsinki, Finland.
[6] 38 Rue des Primevère, 35160 Le Verger, France.
[7,8] Richard B. Gump South Pacific Research Station, University of California Berkeley, BP 244 Maharepa, 98728 Moorea, French Polynesia.
[8] Environmental Science, Policy and Management, University of California, Berkeley, California 94720, USA

[*] Corresponding author: sylvain.charlat@univ-lyon1.fr
[1] Email: thibault.ramage@hotmail.fr
[2] Email: patmsimoes@gmail.com
[3] Email: gmialdea@gmail.com
[5] Email: duplouyanne@yahoo.fr
[6] Email: rousse.pascal@wanadoo.fr
[7] Email: neiltahiti@gmail.com
[8] Email: roderick@berkeley.edu
[†] Deceased

[1] urn:lsid:zoobank.org:author:8DE31F66-13BF-4516-A205-60F2EA39E3DD
[2] urn:lsid:zoobank.org:author:BB4451A0-44AC-46FD-A60D-F019AA87D62E
[3] urn:lsid:zoobank.org:author:E18E3969-C359-4474-AEAC-0D53027672C7
[4] urn:lsid:zoobank.org:author:E1E1055D-D791-4882-AB67-EF29F1392F6A
[5] urn:lsid:zoobank.org:author:AF956717-D0E6-44A3-8F4B-8E312993EDEE
[6] urn:lsid:zoobank.org:author:B06C2640-700A-429B-AA2F-1BE09251C845
[7] urn:lsid:zoobank.org:author:FB092614-B8DC-40F6-BAD5-454A0C880799
[8] urn:lsid:zoobank.org:author:16FFE533-CECC-44BE-AE1A-3E4B543BF48A
[9] urn:lsid:zoobank.org:author:A9AE69C2-039D-47FD-9DD2-B34C4363CB71

Abstract. We report here on the taxonomic and molecular diversity of 10 929 terrestrial arthropod specimens, collected on four islands of the Society Archipelago, French Polynesia. The survey was part of the 'SymbioCode Project' that aims to establish the Society Islands as a natural laboratory in which to investigate the flux of bacterial symbionts (e.g., *Wolbachia*) and other genetic material among branches of the arthropod tree. The sample includes an estimated 1127 species, of which 1098 included at least one DNA-barcoded specimen and 29 were identified to species level using morphological traits only. Species counts based on molecular data emphasize that some groups have been understudied in this region and deserve more focused taxonomic effort, notably Diptera, Lepidoptera and Hymenoptera. Some taxa that were also subjected to morphological scrutiny reveal a consistent match between DNA and morphology-based species boundaries in 90% of the cases, with a larger than expected genetic diversity in the remaining 10%. Many species from this sample are new to this region or are undescribed. Some are under description, but many await inspection by motivated experts, who can use the online images or request access to ethanol-stored specimens.

Keywords. Arthropods, DNA barcoding, French Polynesia, Moorea BioCode, SymbioCode.

Ramage T., Martins-Simoes P., Mialdea G., Allemand R., Duplouy A., Rousse P., Davies N., Roderick G.K. & Charlat S. 2017. A DNA barcode-based survey of terrestrial arthropods in the Society Islands of French Polynesia: host diversity within the SymbioCode Project. *European Journal of Taxonomy* 272: 1–13. http://dx.doi.org/10.5852/ejt.2017.272

Introduction

In this paper, we report on the diversity of a large and non-taxonomically focused sample of terrestrial arthropods, collected on four islands of the Society Archipelago (French Polynesia) from 2005 to 2007. This sample was obtained as part of the SymbioCode initiative (coordinated by SC) in collaboration with the Moorea Biocode Project (http://biocode.berkeley.edu) (Check 2006). Although primarily focused on symbiotic bacteria and their flux across host species (Bailly-Bechet *et al.* in press), this dataset also offers an opportunity to complement biodiversity records from this region, which is the subject of the present paper.

This study focused on four of the five main islands of the Society, situated along a 200 km northwest/southeast axis, corresponding to the movement of the Pacific plate over a unique hot spot during the last three million years (Guillou *et al.* 2005). Typical of young, small and isolated volcanic islands, their fauna is generally characterised by a low diversity of species and a high level of endemism, with an increasing proportion of introduced and invasive species (Whittaker & Fernández-Palacios 2009). Much of the South Pacific's terrestrial biota originates from the west (Australasia), colonizing the oceanic islands via stepping-stone dispersal (Miller 1996; Gillespie & Roderick 2002; Gressitt 1956).

Fairmaire (1849, 1850) was the first to focus on the insect fauna of French Polynesia. This fauna was again intensively studied between 1926 and 1940 as a result of the collections of Saint-George and the Pacific Entomological Survey (reviewed in Ramage in press). Among the 2972 valid arthropod species names reported from French Polynesia, 61% are considered endemic, but the level of introduced species is also high, representing 10% of the overall species count (Ramage in press).

In the present study, we report on the occurrence of 1127 arthropod species (Supplementary file, sheet S1): 228 of them were assigned to previously described species, and of the rest, 105 to genus, 567 to family and 227 to order. DNA barcodes were the main source of taxonomic information, but some taxa (listed in Table 1) were also thoroughly characterised based on morphology. For groups that were analysed by taxonomic experts, combined genetic and morphological evidence suggests the occurrence

Table 1. Taxa that have been identified on the basis of morphology. For these groups, names of previously described species were used and new species are being described.

Taxon	Expert
Araneae	Michaël Dierkens
Blattodea, Phasmida, Psocodea, Scolopendromorpha, Mantodea	Thibault Ramage
Coleoptera (Anthribidae, Buprestidae, Cerambycidae, Chrysomelidae, Coccinellidae, Curculionidae, Dryophthoridae, Elateridae, Endomychidae, Lucanidae, Monotomidae, Mycetophagidae, Nitidulidae, Oedemeridae, Scarabaeidae, Silvanidae)	Thibault Ramage
Coleoptera (Carabidae)	Alexander Anichtchenko
Diptera (Neriidae, Platystomatidae, Stratiomyidae, Syrphidae, Tephritidae)	Thibault Ramage
Hemiptera (Aphrophoridae, Aradidae, Cicadellidae, Coreidae, Geocoridae, Gerridae, Miridae, Oxycarenidae, Pentatomidae, Plataspidae, Scutelleridae, Tingidae)	Thibault Ramage
Hemiptera (Cixiidae)	Hannelore Hoch
Hemiptera (Delphacidae)	Manfred Asche
Hymenoptera (Apidae, Chrysididae, Crabronidae, Evaniidae, Formicidae, Sphecidae, Vespidae)	Thibault Ramage
Hymenoptera (Ichneumonidae)	Pascal Rousse
Lepidoptera (Lycaenidae, Nymphalidae, Sphingidae)	Thibault Ramage
Odonata	Daniel Grand
Orthoptera (Gryllidae, Mogoplistidae)	Thibault Ramage

of many species new to this region or new to science. Some of these species are in the process of formal description by various colleagues (Dierkens & Ramage 2016; Rousse *et al.* in prep.), but many more await inspection by motivated experts, on the basis of online photos or ethanol-stored material, available upon request.

Material and methods

Sampling

Arthropods were collected in the Society Islands of Tahiti, Moorea, Huahine and Raiatea, with most effort, for logistical reasons, on the island of Moorea and to a lesser extent Tahiti. Collecting took place from 2005 to 2007 using sweep nets, Malaise traps (only in Moorea and Tahiti) and light traps, or by hand, without focusing on any particular habitat or taxonomic group. Details of the 317 collecting events, defined as a unique combination of location and date, are given in the Supplementary file, sheet S2. Raw material from the field was stored at ambient temperature in 95% ethanol in 50 ml centrifuge tubes for further processing.

Photographs and morpho-species assignments

Specimens showing no apparent morphological differences to a non-expert eye were attributed the same morpho-species numbers. Up to ten specimens per morpho-species were individually photographed (under binocular microscope or with a reflex camera depending on their size) and stored in 95% ethanol at -20°C before further processing. In total, 10 929 specimens were processed in this way (listed in the Supplementary file, sheet S3). All photographs can be accessed through the BOLD database (http://www.boldsystems.org) (SYC project, DS-SYMC dataset, http://dx.doi.org/10.5883/DS-SYMC) and the

Moorea Biocode database (http://biocode.berkeley.edu). Additional specimens from overly represented morpho-species were mass stored in 95% ethanol at -20°C for potential future analysis (additional specimens, see the Supplementary file, sheet S1).

DNA extraction

In order to maximize the phylogenetic and geographic diversity of the study, we selected specimens for DNA extraction and further molecular analysis using the following criteria when possible: up to three specimens per morpho-species per island and from different locations. We thus selected 4837 specimens, from which DNA was extracted using Nucleospin 96 Tissue kits (MACHEREY-NAGEL GmbH & Co. KG, Düren, Germany), according to the manufacturer's instructions, with the following modifications: (a) the pre-lysis step consisted of incubation with proteinase K overnight and (b) the elution step included two sub-steps (each with 50 µl of water). If specimens were smaller than 5 mm long, the whole body was used for DNA extraction; otherwise, we used tissue from the mid-lower abdomen, including the gonads but leaving the genitalia intact for potential subsequent taxonomic work. We stored 20 µl of each genomic extract at -20°C for further analysis; the remaining 80 µl were placed at -80°C for long-term storage.

DNA barcoding

A 643–664 bp (most commonly 658 bp) region of the mitochondrial gene *Cytochrome Oxidase 1* (*CO1*) was amplified with standard LCO1490 and HCO2198 primers (Folmer *et al.* 1994). DNA amplification (PCR) was performed in a total volume of 30 µl with 1.5 mM of $MgCl_2$, 2 mM of all four dNTPs, 0.2 µM of each primer, 0.02 Units/µl of EuroTaq R DNA polymerase (EUROBIO, Les Ulis, France) and 2 µl of template using the following temperature profile: initial denaturation at 95°C for 120 seconds (s); 35 cycles of 94°C for 30 s, 47°C for 30 s and 72°C for 90 s; and a final extension at 72°C for 600 s. All reactions took place in a Tetrad R Thermocycler (BIO-RAD, Hercules, CA, USA). PCR products were purified and Sanger-sequenced on an ABI 3730 using both the forward and reverse PCR primers.

Trace files were imported in GENEIOUS v. 5.4.0 (Biomatters) (Kearse *et al.* 2012). The first 45 bp of the 5' end of each read were trimmed, as well as further low-quality end regions with error probability larger than 0.1%. Heteroplasmic positions were identified based on peak similarity (more than 50% overlap between two conflicting peaks at a given position in one read); ambiguous base calls were assigned to heteroplasmic positions using the IUPAC code. Forward and reverse reads were then assembled and a consensus sequence integrating Phred quality scores was produced. In the consensus sequence, the Phred score used for a given position was the Phred score from a single read if only one read covered this position, or the sum of the forward and reverse scores in case of agreement between the forward and reverse reads. In case of conflict between the reads at a given position, the base used and the associated Phred score in the consensus were those of the highest quality read at the position. An ambiguous base call (N) was assigned to poorly supported positions (Phred scores below 20) in the consensus. High quality reads (i.e., those carrying more than 80% nucleotides with Phred scores larger than 40), but not assembled to their reverse complement because of failure of one reaction, were also recovered. Consensus sequences and recovered single reads were then all aligned to an arbitrarily chosen reference *CO1* sequence (GenBank KF226475, from the butterfly *Hypolimnas bolina*) in order to unambiguously identify and trim primer regions.

Reads longer than 200 bp were then selected and the longest open reading frame was determined for all reads, yielding only 12 sequences with stop codons, which were discarded. In addition, 18 sequences were identified as experimental contaminants (precisely matching another sequence in the dataset) and thus were excluded. All such contaminations occurred locally in DNA extraction or PCR plates: the

contaminant well was never more than two wells away from the contaminated well. Notably, specimens were randomly distributed across 54 plates, so that specimens from the same morpho-species were not clustered, allowing cross contaminations to be detected when it occurred. We also excluded five natural contaminants, harbouring an insect parasitoid instead of host barcode. Sequences were aligned in Geneious using the predicted protein sequences.

Clustering of sequences within OTUs

Of 3627 *CO1* sequences, 3180 passed the BOLD database quality filters and were clustered using the Refined Single Linkage (RESL) algorithm into 1026 species-level groups or OTUs (operational taxonomic units), each identified with a unique Barcode Index Number (BIN) (Ratnasingham & Hebert 2013). In brief, RESL employs a single linkage clustering procedure (producing clusters including all sequences connected by a genetic distance below a certain threshold) followed by a refinement step in which clusters including at least three sequences can be split if they show significant discontinuity in genetic distances. Of the 447 sequences that did not pass the BOLD quality filters, 314 were assigned to one of the BINs identified by RESL. The remaining 133 specimens were clustered in groups of sequences diverging by no more than 3% substitutions per site, resulting in an additional 84 OTUs also included as they represent significant information, although they should be considered more cautiously. The procedure used for each specimen is indicated in the Supplementary file, sheet S3.

Taxonomic assignments

With the exception of spiders, which were classified to the lowest possible taxonomic level based on morphology prior to DNA extraction (Dierkens & Charlat 2009), all specimens were only assigned the taxonomic rank of order or class upon morpho-species assignments. For barcoded specimens, the consistency of this assignment with the 1st *blast* hit was verified (using *blastn* with default options against the NCBI nr database). Closer inspections of photographs or specimens allowed correction of any remaining inconsistencies. Such errors were either due to contaminations (as noted above) or, more often, to an incorrect initial assignment. A more accurate DNA barcode-based taxonomic assignment was then proposed based on the quality of the best *blast* hit, with the following categories: (1) species level assignment for at least 98% sequence identity over at least 80% of the barcode; (2) genus level for 95% sequence identity and 80% overlap; (3) family level for 90% sequence identity and 80% overlap.

In combination with *blast*, we developed a phylogenetic approach to propose a taxonomic assignment based on DNA sequence data. We developed R scripts to extract one representative *CO1* sequence from each arthropod genus from GenBank (resulting in 20 350 sequences). For each arthropod order, we added these sequences to our own alignment using profile alignment in MAFFT (Katoh & Standley 2013) and constructed a tree for each order using FastTree (Price *et al.* 2010). Poorly supported nodes (less than 80% aLRT support) (Anisimova & Gascuel 2006) were collapsed in Archaeopteryx (Han & Zmasek 2009). We then inspected these trees in Dendroscope (Huson & Scornavacca 2012), with colour coding of tips based on family names, to refine and verify *blast*-based assignments. An improved classification of our specimen was then proposed using the "strict" criteria defined by Wilson *et al.* (2011), which results in a very low level (2%) of false positives. In brief, a taxon name is assigned to a specimen if (1) its position within a clade is strongly supported (at least 80% aLRT support) and (2) the clade where it is placed is taxonomically homogeneous. For groups listed in Table 1, we found no inconsistencies between morphology and barcode-based assignments. For the other groups, most taxonomic assignments relied on barcodes only, as detailed in the Supplementary file, sheet S1. Some morpho-species that had not produced a good quality barcode were also identified to the species level when this was feasible without ambiguity.

Results

Characteristics of the sample

In total, the SymbioCode sample includes 10 929 arthropod specimens that were assigned to morpho-species, individually photographed, and deposited in the BOLD database (http://www.boldsystems.org) (SYC project, DS-SYMC dataset, http://dx.doi.org/10.5883/DS-SYMC) and the Moorea Biocode databases (http://biocode.berkeley.edu). The Supplementary file (sheet S3) provides a complete list of the specimens with a unique URL for each specimen. More than half (52%) of the specimens were collected at low elevations or near sea level (below 100 m elevation, Fig. 1) in potentially disturbed habitats, which are more likely to harbour non-native species than locations at higher elevations. Malaise traps were used only on the islands of Moorea and Tahiti, and mostly in remote locations at high altitude.

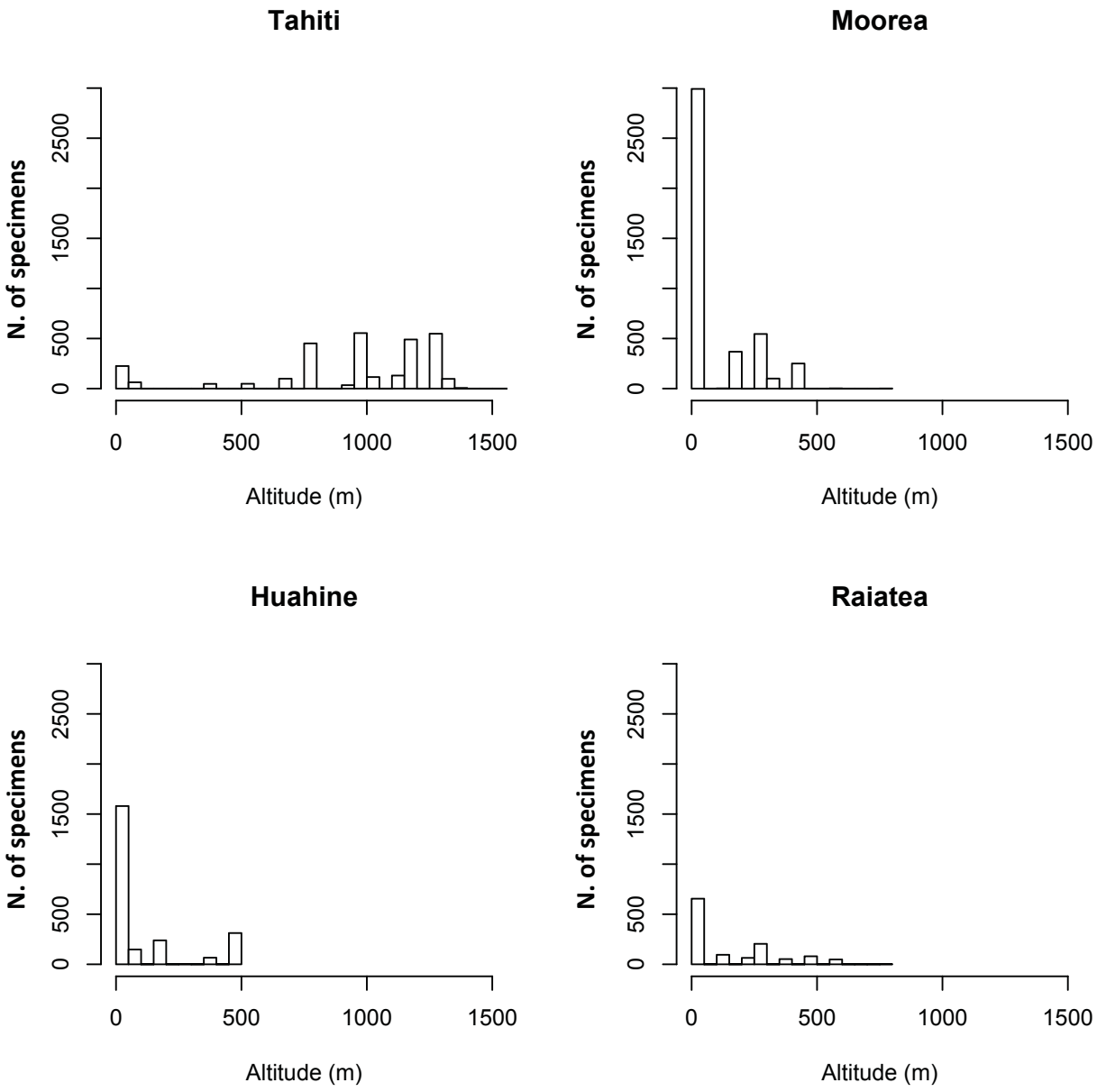

Fig. 1. Distributions of the elevations at which specimens were collected on each island.

Table 2. Rate of DNA barcoding success in the main taxa under study. Groups with fewer than 10 specimens tested are not included. The barcoding success rate is defined as the proportion of specimens tested that yielded a quality DNA barcode (see Material and methods for quality criteria).

Taxa	Barcoding success rate	N specimens	N species
Amphipoda	61.50%	13	4
Araneae	59.50%	412	50
Blattodea	68.20%	44	11
Coleoptera	62.70%	413	119
Collembola	26.70%	60	8
Dermaptera	5.90%	17	1
Diptera	90.80%	1109	305
Hemiptera	65.70%	696	133
Hymenoptera	74.60%	689	172
Isopoda	5.00%	20	1
Isoptera	30.80%	13	1
Ixodida	13.00%	23	1
Lepidoptera	91.50%	885	223
Neuroptera	100.00%	18	5
Odonata	63.60%	44	8
Orthoptera	65.10%	172	16
Polydesmida	9.10%	11	1
Psocodea	70.20%	104	24
Thysanoptera	86.40%	22	5

Hence, island, collecting method, and elevation are not independent (Fig. 1). Considering the four islands together, 30% of the specimens from low elevation were collected with Malaise traps, compared to 90% of those from high altitude. This sampling effort should be kept in mind when quantifying diversity, because Malaise traps tend to collect relatively more Diptera and Hymenoptera compared to other orders, especially Coleoptera (Gressitt & Gressitt 1962; Lamarre *et al.* 2012). Among the 317 collecting events, 80% took place at low elevation and contained one or a few samples (thanks, notably, to the contribution of middle-school students from the Paopao School of Moorea). In contrast, high elevation collecting events were generally more intense and contributed more specimens per event.

Barcoding success rate

DNA was extracted from 4837 specimens. As detailed in Table 2 (by order), the Supplementary file, sheet S2 (by species) and the Supplementary file, sheet S3 (by specimen), amplification and sequencing of the *CO1* locus was successful in 3627 specimens, that is, 75% of the genomic extracts, but the success rate varied widely among arthropod orders (Table 2). Phylogeny explained a large part of this variation, with some clades being clearly less efficiently sequenced than others, possibly because of divergence in the primer target regions. Most of the barcode data were of high quality (92% of the *CO1* sequences were longer than 400 bp and included fewer than 1% ambiguous positions); 8% were of lower quality (between 200 and 400 bp long, including up to 20 ambiguous positions) but carried enough information to be maintained in the analysis. Sequences are available in BOLD, and also in GenBank (BankIt1909431: KX051578–KX055204).

Biodiversity

The *CO1* sequences clustered into 1693 different haplotypes and into 1110 species-like groups or OTUs defined on the sole basis of molecular data. Of the 1110 OTUs, 1026 (92%) were defined using the RESL algorithm implemented in BOLD and thus received a unique identifier in this database (BIN); 642 of the BINs were new to the BOLD database. The remaining 84 OTUs comprised sequences that did not pass the BOLD quality filters, but were clustered in groups of sequences diverging by no more than 3% substitutions per site. Some specimens, from which DNA was not extracted, or for which *CO1* sequencing failed, were nevertheless assigned on the basis of morphology (when possible to do so with confidence) to OTUs from this study (n = 4295 specimens) or to a named species not represented in the other OTUs (n = 411 specimens).

In summary, the biodiversity reported here is based on 8248 specimens, representing an estimated 1127 species designated as follows: (i) 199 named species (2924 specimens) including at least one sequenced specimen, (ii) 899 OTUs (5177 specimens), including at least one sequenced specimen, not named at the species level but taken as a proxy for species identity, and (iii) 29 named species (147 specimens) without DNA sequence data. Each species used in these counts was assigned a species ID, as listed in the Supplementary file, sheet S1.

The 1127 species are distributed among major arthropod groups as shown in Table 3. Most of the biodiversity collected during this survey comprises the five major insect orders (Hemiptera, Coleoptera, Diptera, Hymenoptera, Lepidoptera). Importantly, the species richness in the three orders most frequently represented in the sample (Diptera, Lepidoptera and Hymenoptera) is higher than documented in all earlier reports combined for the Society Islands (Ramage in press). For Hymenoptera, the diversity reported here is greater than known previously for the entirety of French Polynesia, including all of its five archipelagos. Considering groups that have been well characterised in our study (listed in Table 1), we observe that only 60% of the species found in our sample have previously been reported.

Our sample includes 107 species from which we can assess the level of correlation between OTUs and morphology-based species identity, because they were distinguished on the basis of morphology and include at least two barcoded specimens (Supplementary file, sheet S1). No specimens from the same OTU received different species names based on morphology, suggesting the DNA sequence-based approach was at least as efficient as morphology in distinguishing species. In contrast, 11 out of 107 proposed species included more than one BIN, resulting in a 10% error rate, which is in the lower range of what has been found in other studies (Meier *et al.* 2006; Ratnasingham & Hebert 2013) (see the Supplementary file, sheet S4 for a list of these species and summary statistics on their mitochondrial diversity).

New records and other notable occurrences

In this section we consider the 227 species that were named and highlight notable occurrences. Although Coleoptera has been the most intensively collected order in this region, we note new records (Supplementary file, sheet S5). For Hymenoptera, the present sample generally shows equal or higher diversity than all previous reports combined. Our sample also contributes two new family records for French Polynesia, the Sphecidae and the Chrysididae, as reported previously (Ramage *et al.* 2015; Ramage & Kimsey 2015). Among species from the Aculeata infra-order identified to genus or species, new records for French Polynesia include two species of Apidae (genera *Xylocopa* Latreille, 1802 and *Braunsapis* Michener, 1969), three species of Crabronidae (genera *Liris* Fabricius, 1804 and *Polemistus* Saussure, 1892) and two species of Vespidae (genera *Pachodynerus* Saussure, 1875 and *Vespula* Thomson, 1869). These collections will be the subject of upcoming, more specific publications (Ramage in press). The examination of the family Ichneumonidae led to a similar conclusion: among the 19 species belonging to this family, 14 were identified to genus or species level and five were already known in French Polynesia, but four are new records for the country and five represent new species (Rousse *et al.* in prep.).

Table 3. Species counts in this study and in previous records from the Society Archipelago (based on Ramage submitted).

Taxon	N species earlier surveys (Society)	N species earlier surveys (French Polynesia)	N species SymbioCode	Named sp. SymbioCode: previous record/ new record/new sp.
Blattodea	12	28	14	5/0/0
Chelicerata	238	337	77	24/11/0
Coleoptera	558	772	119	42/3/0
Collembola	10	33	8	0/0/0
Crustacea	37	78	6	1/0/0
Dermaptera	8	9	1	0/0/0
Diptera	201	328	307	20/8/0
Embioptera	1	1	0	0/0/0
Hemiptera	224	401	135	17/0/0
Hymenoptera	126	172	178	32/5/5
Lepidoptera	151	494	224	30/4/0
Mantodea	1	1	1	1/0/0
Myriapoda	15	19	5	2/0/0
Neuroptera	11	14	5	0/0/0
Odonata	12	19	8	7/0/0
Orthoptera	17	32	16	2/0/0
Phasmida	2	2	1	1/0/0
Psocodea	42	56	25	4/0/0
Siphonaptera	2	3	1	1/0/0
Thysanoptera	43	59	5	1/0/0
Zygentoma	1	5	0	0/0/0

The other insect orders, which to date have been less thoroughly studied taxonomically in this project, are also likely to harbour many species previously undocumeted in French Polynesia. For example, the family Stratiomyidae (Diptera) was only known for two species in French Polynesia, *Hermetia illucens* (Linnaeus, 1758) and *Chromatopoda annulipes* (Walker, 1849), while our data show that four other species are likely present. Within the order Hemiptera, new records for French Polynesia are likely in the families Plataspidae (genus *Coptosoma* Laporte de Castelnau, 1832), Scutelleridae (genus *Calliphara* Germar, 1839; Ramage & Gourvès 2016), Tingidae (genus *Corythucha* Stål, 1873), Oxycarenidae (genus *Oxycarenus* Fieber, 1837) and Membracidae (genus *Spissistilus* Caldwell, 1949). Among the 227 named species in our sample, 88 are indigenous or with unknown status, 122 are introduced (including new records), and 17 are endemic (Supplementary file, sheet S1).

Discussion

The SymbioCode collection, described in the present study, provides an overview of the terrestrial arthropod diversity of the Society Archipelago. Approximately one fifth of the species present in this sample have been identified at the species level, extending the known range of some species to French Polynesia and enriching knowledge of other groups of taxa.

Although not exhaustive, and constrained by the collecting and taxonomic methods used, the study reveals that many arthropod groups are far more diverse than suggested by previous reports. Focusing on groups that have been morphologically well characterised in our study, we observe that only 60% of the species collected have previously been reported. In other words, many of the collected species are new to this region or to science (Dierkens & Ramage 2016; Rousse *et al.* in prep.), emphasising the scale of the challenge to characterize tropical arthropod biodiversity, even for small and isolated islands.

The correlation between molecular-based OTUs and morphology-based species identity was assessed in 107 species that were distinguished on the basis of morphology and included at least two barcoded specimens. No specimens from the same OTU received different species names based on morphology, but 10% of the proposed species based on morphology included more than one OTU. Three explanations can be proposed to explain such high sequence diversity within these groups. Some species may have very large effective population sizes and/or high natural levels of genetic diversity (for example, invasive species introduced from multiple distant sources). Alternatively, some proposed species names may correspond to more than one true species that may or may not be distinguishable morphologically following deeper scrutiny. Finally, some species may have been subject to mitochondrial introgression, possibly fuelled by cytoplasmic symbionts (Hurst & Jiggins 2005) and thus carry mitochondrial DNA lineages from distinct species. Further work is needed, including more thorough morphological inspection and sequencing of nuclear DNA markers, to determine which explanations hold.

While this study confirms that DNA barcoding offers a powerful means to quantify the diversity of complex samples, it also illustrates that molecular data are more valuable when associated with that from morphological characters. We thus hope this report will stimulate further synergistic projects to characterize the biodiversity of this region and encourage experts to complement our analysis, using the on-line photographs and metadata, or the many remaining specimens, available upon request.

Acknowledgments

We are grateful to the CNRS for the ATIP grant "SymbioCode" to SC, the Genoscope for DNA sequencing through "Bibliothèque du Vivant" and "Speed ID" projects, and the Moorea Biocode team and the Gordon and Betty Moore Foundation for financial and logistic support, including access to Malaise traps. We also thank Julien Varaldi for handy R scripts; Alexander Anichtchenko, Daniel Grand, Hannelore Hoch and Manfred Asche for invaluable taxonomic expertise; and finally Marjorie Saillan and the "entomology club" of the Paopao School in Moorea for their unique contribution to the collecting effort. This article is dedicated to Roland Allemand for his early encouragement, and more generally for his major contributions to the field of entomology.

References

Anisimova M. & Gascuel O. 2006. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Systematic Biology* 55: 539–552.

Bailly-Bechet M., Simoes P., Szöllősi G., Mialdea G., Sagot M.-F. & Charlat S. In press. How long does *Wolbachia* remain on board? *Molecular Biology and Evolution*.

Check E. 2006. Treasure island: pinning down a model ecosystem. *Nature* 439: 378–379. http://dx.doi.org/10.1038/439378a

Dierkens M. & Charlat S. 2009. Contribution à la connaissance des araignées des îles de la Société (Polynésie française). *Revue Arachnologique* 17: 63–81.

Dierkens M. & Ramage T. 2016. Deuxième contribution à la connaissance des araignées de Polynésie française. *Bulletin mensuel de la Société Linnéenne de Lyon* 85: 134–172.

Fairmaire L.M.H. 1849. Essai sur les Coléoptères de la Polynésie. *Revue et Magasin de Zoologie* 2: 277–291, 352–365, 410–422, 445–460, 504–516, 550–559.

Fairmaire L.M.H. 1850. Essai sur les Coléoptères de la Polynésie. *Revue et Magasin de Zoologie* 2: 50–64, 115–122, 181–185.

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.

Gillespie R.G. & Roderick G.K. 2002. Arthropods on islands: colonization, speciation, and conservation. *Annual Review of Entomology* 47: 595–632.

Gressitt J.L. 1956. Some distribtion patterns of Pacific island faunae. *Systematic Zoology* 5: 11–47.

Gressitt J.L. & Gressitt M.K. 1962. An improved Malaise trap. *Pacific Insects* 4: 87–90.

Han M.V. & Zmasek C.M. 2009. phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics* 10: e356. http://dx.doi.org/10.1186/1471-2105-10-356

Hurst G.D. & Jiggins F.M. 2005. Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. *Proceedings of the Royal Society B* 272: 1525–1534.

Huson D.H. & Scornavacca C. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Systematic Biology* 61: 1061–1067. http://dx.doi.org/10.1093/sysbio/sys062

Katoh K. & Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* 30: 772–780. http://dx.doi.org/10.1093/molbev/mst010

Kearse M., Moir R., Wilson A., Stones-Havas S., Cheung M., Sturrock S., Buxton S., Cooper A., Markowitz S., Duran C., Thierer T., Ashton B., Meintjes P. & Drummond A. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28: 1647–1649. http://dx.doi.org/10.1093/bioinformatics/bts199

Lamarre G.P.A., Molto Q., Fine P.V.A. & Baraloto C. 2012. A comparison of two common flight interception traps to survey tropical arthropods. *ZooKeys* 216: 43–55. http://dx.doi.org/10.3897/zookeys.216.3332

Meier R., Shiyang K., Vaidya G. & Ng P.K.L. 2006. DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic Biology* 55: 715–728. http://dx.doi.org/10.1080/10635150600969864

Miller S. 1996. Biogeography of Pacific insects and other terrestrial invertebrates: a status report. *In*: Keast A. & Miller S. (eds) *The Origin and Evolution of Pacific Island Biotas, New Guinea to Eastern Polynesia: Patterns and Processes*: 463–475. Academic Publishers, Amsterdam.

Price M.N., Dehal P.S. & Arkin A.P. 2010. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5 (3): e9490. http://dx.doi.org/10.1371/journal.pone.0009490

Ramage T. In press. Checklist of the terrestrial and freshwater arthropods of French Polynesia (Chelicerata; Myriapoda; Crustacea; Hexapoda). *Zoosystema*.

Ramage T. & Gourvès J. 2016. Un nouveau Scutelleridae pour la Polynésie française, *Calliphara bifasciata* White, 1839 (Hemiptera). *Bulletin de la Société Entomologique de France*, in press.

Ramage T. & Kimsey L.S. 2015. The Aculeata of French Polynesia. IV. First record of *Chrysis angolensis* (Hymenoptera, Chrysididae). *Bulletin de la Société Entomologique de France* 120: 209–211.

Ramage T., Charlat S. & Jacq F. 2015. The Aculeata of French Polynesia. III. Sphecidae, with the record of three new species for the Society Islands (Hymenoptera). *Bulletin de la Société Entomologique de France* 120: 157–163.

Ratnasingham S. & Hebert P.D.N. 2013. A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. *PLoS ONE* 8 (7): e66213. http://dx.doi.org/10.1371/journal.pone.0066213

Rousse O., Ramage T. & Charlat S. In prep. Ichneumonid wasps of French Polynesia: synopsis, illustrated key, and description of 10 new species.

Whittaker R. & Fernández-Palacios J. 2009. *Island Biogeography. Ecology, Evolution, and Conservation*. Second Edition. Oxford University Press, Oxford.

Wilson J.J., Rougerie R., Schonfeld J., Janzen D.H., Hallwachs W., Hajibabaei M., Kitching I.J., Haxaire J. & Hebert P.D. 2011. When species matches are unavailable are DNA barcodes correctly assigned to higher taxa? An assessment using sphingid moths. *BMC Ecology* 11: e18. http://dx.doi.org/10.1186/1472-6785-11-18

Manuscript received: 16 February 2016
Manuscript accepted: 22 July 2016
Published on: 7 February 2017
Guest editors: Line Le Gall, Frédéric Delsuc, Stéphane Hourdez, Guillaume Lecointre and Jean-Yves Rasplus
Desk editor: Danny Eibye-Jacobsen

Information on Supplementary file

Sheet S1. Species list. n_otu: number of OTUs, relevant and only provided for named species. nap: not applicable. md: missing data. n_specimens: number of specimens stored in individual tubes. Additional_specimens: indicates whether additional specimens, mass stored in ethanol, are available in addition to those indicated in the "n_specimens" column. DNA_tissue: body part used for DNA extraction. n_DNA_extract: number of specimens from which DNA was extracted and *CO1* sequencing was attempted. biogeo_status: biogeographical status (P: previously reported taxa, either indigenous or of unknown status; E: endemic taxa, endemic to a single or several islands, or to a single or several archipelagos of French Polynesia; I: introduced taxa). BIN: list of Barcode Index Numbers included in this species. OTUs: list of Operational Taxonomic Units (using our internal OTU index) included in this species. Clustering_method: method used to create the OTU(s) making a species, when relevant (RESL for most species, 3%_threshold for a few species that did not pass the BOLD quality filter). Morpho-species: a morpho-species ID, given only for specimens that did not receive a species ID (that is, specimens that were not barcoded, not assigned to a barcoded species based on morphological characters and not named at the species level).

Sheet S2. Collecting events. md: missing data.

Sheet S3. Specimens. md: missing data. symbiocode_ID: unique identifiers for specimens, internal to this study (matches with BOLD and Biocode IDs are also given, as well as links to specimen pages, including photographs and all metadata). col_event: collecting event (see sheet S2 for details). specimen_available: indicates whether the specimen is available upon request for further study. species_id: unique identifier for species, as listed in sheet S1. included_in_biodiversity_survey: indicates whether the specimen was included in the species count, which is the case for specimens belonging to barcoded species or named at the species level on the basis of morphological characters. BIN: Barcode Index Number as defined in the BOLD database. OTU: our internal Operational Taxonomic Unit number, corresponding to the BIN when a sequence was assigned to a BIN. Clustering_method: "RESL" if the sequence was directly assigned a BIN number in BOLD, "RESL_indirect" for lower quality sequences that were assigned to a BIN or "3%_threshold" for lower quality sequences that were not assigned to a BIN, but clustered in groups at a maximum of 3% divergence.

Sheet S4. Mitochondrial diversity in 11 species containing more than one OTU based on the RESL algorithm. Pi: mean of the raw genetic distances calculated among all individuals. Max_dist: maximum raw genetic distance within this species. BINs: list of Barcode Index Numbers included in each species. Additional relevant information for each species can be found in sheet S1.

Sheet S5. New species records for French Polynesia.