DR. BRENT EMERSON (Orcid ID : 0000-0003-4067-9858)

Received Date : 09-Sep-2016
Revised Date : 12-Dec-2016
Accepted Date : 24-Dec-2016
Article type : From the Cover

MtDNA metagenomics reveals large-scale invasion of belowground arthropod communities by introduced species

Francesco Cicconardi[1], Paulo A. V. Borges[2] , Dominique Strasberg[3], Pedro Oromí[4], Heriberto López[5], Antonio J. Pérez-Delgado[5], Juliane Casquet[6], Juli Caujapé-Castells[7], José María Fernández-Palacios[8], Christophe Thébaud[6] & Brent C. Emerson[5,9]

[1]Institute of Ecology, University of Innsbruck, Technikerstrasse 25, a-6020 Innsbruck, Austria.
[2]CE3C – Centre for Ecology, Evolution and Environmental Changes / Azorean Biodiversity Group and Universidade dos Açores –Departamento de Ciências Agrárias, Rua Capitão João d'Ávila s/n, 9700-042, Angra do Heroísmo, Açores, Portugal.
[3]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.
[4]Departamento de Biología Animal y Edafología y Geología, Universidad de La Laguna, C/ Astrofísico Francisco Sánchez, 38206 La Laguna, Tenerife, Canary Islands, Spain.
[5]Island Ecology and Evolution Research Group, IPNA-CSIC, 38206 La Laguna, Tenerife, Canary Islands, Spain.
[6]Laboratoire Evolution & Diversité Biologique, UMR 5174 CNRS-Université Paul Sabatier-ENFA, 31062 Toulouse Cedex 9, France.
[7]Departamento de Biodiversidad Molecular y Banco de ADN, Jardín Botánico Canario 'Viera y Clavijo' - Unidad Asociada CSIC, Cabildo de Gran Canaria, Camino del Palmeral 15 de Tafira Alta, 35017 Las Palmas de Gran Canaria, Spain.
[8]Island Ecology and Biogeography Research Group. Instituto de Enfermedades Tropicales y Salud Pública de Canarias (IUETSPC), Universidad de La Laguna, Tenerife, Canary Islands 38206, Spain.
[9]School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK.

Key words: mesofauna, soil, introduced species, island biogeography, invertebrate

Corresponding author: Brent Emerson, Island Ecology and Evolution Research Group, Instituto de Productos Naturales y Agrobiología (IPNA-CSIC), C/ Astrofísico Francisco Sánchez 3, 38206 - La Laguna, Santa Cruz de Tenerife - Canary Islands, Spain. Ph: +34 922 256 847 (ext. 283). Fax: +34 922 260 135. Email: bemerson@ipna.csic.es

Short title: Soil mesofauna assembly by species introductions

Abstract

Using a series of standardised sampling plots within forest ecosystems in remote oceanic islands, we reveal fundamental differences between the structuring of aboveground and belowground arthropod biodiversity that are likely due to large-scale species introductions by humans. Species of beetle and spider were sampled almost exclusively from single islands, while soil dwelling Collembola exhibited more than tenfold higher species sharing among islands. Comparison of Collembola mitochondrial metagenomic data to a database of more than 80,000 Collembola barcode sequences revealed almost 30% of sampled island species are genetically identical, or near identical, to individuals sampled from often very distant geographic regions of the world. Patterns of mtDNA relatedness among Collembola implicate human-mediated species introductions, with minimum estimates for the proportion of introduced species on the sampled islands ranging from 45-88%. Our results call for more attention to soil mesofauna to understand the global extent and ecological consequences of species introductions.

Introduction

To understand soil ecosystem functioning, with reference to phenomena that extend beyond soil itself, such as potential cascading effects across trophic levels or the impact of introduced and potentially invasive non-native species on ecosystem processes (e.g. Ehrenfeld 2010; Wardle *et al.* 2004; Yang *et al.* 2009), advances are needed to bridge the gap between belowground and aboveground terrestrial systems. However, such advances are limited by the paucity of biodiversity data for soil, which has been referred to as the "third biotic frontier", along with tropical forest canopies and ocean abysses (André *et al.* 1994). Forest soils are especially challenging, as a single square metre of

temperate forest soil may contain more than 1000 species of invertebrates, most of which are less than 2mm in length (Schaefer & Schauermann 1990). Much of the invertebrate species diversity of soil remains uncatalogued, meaning that there is probably no soil where we are able to identify, or even quantify all resident invertebrates (Decaëns 2010; Wall *et al.* 2005). This lack of primary data on species identity complicates the study and measurement of soil invertebrate biodiversity, which is critically needed as the taxonomic composition of an ecosystem determines the diversity of forms and functions. This functional component of biodiversity, which acts as a key driver of ecosystem functioning (Violle *et al.* 2015), and this is likely to be of great importance in soil ecosystems (Bardgett & Van der Putten 2014; Dominati *et al.* 2010; Heemsbergen *et al.* 2004; Lavelle *et al.* 2006).

For more than a decade, it has been recognised that DNA sequence analysis can provide some relief to the taxonomic impediment – the limitation to science imposed by the difficulty in identifying living species, most of which remain undescribed (Gaston 1991; Lomolino *et al.* 2010). The original methods for DNA barcoding (Hebert *et al.* 2003) have been developed into powerful and effective metabarcoding protocols that are particularly well-suited for analysing species-rich assemblages of taxa like invertebrates (e.g. Ramírez-González *et al.* 2013; Yu *et al.* 2012). More recently, shotgun metagenomic sequencing of mixed insect species templates, with a particular focus on beetle phylogenetics, has yielded numerous reads corresponding to the mitochondrial DNA (mtDNA) genome that can be assembled into full or partial mitogenomes (Andújar *et al.* 2015; Crampton-Platt *et al.* 2015; Gillett *et al.* 2014; Gómez-Rodríguez *et al.* 2015; Tang *et al.* 2014). It has been pointed out that, in addition to phylogenetic reconstruction, such extensive mtDNA genome data also offer great potential for understanding how community assembly and structure influence functioning in ecosystems by providing a powerful way to reveal biodiversity that was previously "invisible" (Andújar *et al.* 2015).

Here we employ mitochondrial metagenomics to compare soil-dwelling and aboveground arthropod communities sampled across three remote oceanic archipelagos, two of which are located in the northern Atlantic (Canary Islands and Azores) and the third (Mascarene Islands) being located in the southwestern Indian Ocean. For soil-dwelling arthropods we focus on the ubiquitous and dominant soil mesofaunal taxon Collembola. Species identification of Collembola is complicated by (i)

small adult size that can typically range between 0.2 - 2 mm (Decaëns 2010), (ii) pervasive cryptic species (Emerson *et al.* 2011) that can result in underestimates of morphologically-derived species richness by more than an order of magnitude (Cicconardi *et al.* 2013; Cicconardi *et al.* 2010), and (iii) changes in adult morphology attributable to ecomorphosis, epitoky and cyclomorphosis (Hopkin 1997).

Metabarcoding has previously been used to suggest that a substantial proportion of the Collembola fauna of the Canary Island of Tenerife is of recent origin (Ramírez-González *et al.* 2013). However, the limited mtDNA sequence length of 220bp obtained by Ramírez-González *et al.* (2013) resulted in taxonomic uncertainty for many sequences, rendering quantitative comparisons of taxonomic relatedness unreliable. We address this limitation by using a mitochondrial metagenomic approach to first robustly assign DNA sequences to the class Collembola by means of phylogenetic analysis, and then evaluate their distribution limits beyond the island where they were sampled. We sample soil Collembola communities from forest ecosystems on the islands of Tenerife (Canary Islands), Terceira (Azores) and Réunion (Mascarene Islands). We first compare sharing of Collembola species among islands, placing our results into context by also sampling and comparing aboveground arthropod communities of beetle and spider species from the same sampling sites. We then use a publicly available database of more than 80,000 geographically referenced Collembola barcode sequences to quantify the proportion of Collembola taxa exhibiting high mtDNA genome identity to individuals sampled from regions beyond the three sampled islands. Our study provides significant and novel insight into how biogeographical patterns of species diversity can be influenced by human activities at a very large-scale in the soil biome, while subaerial biomes remain largely unaffected.

Materials and methods

Sampling

Ten sampling plots measuring 50m x 50m were established in forest habitat on each of the islands of Tenerife in the Canary Islands (laurel forest), Terceira in the Azores (elfin cloud forest) and Réunion in the Mascarene Islands (lowland rainforest) (Fig. 1). The 30 plots were subjected to a standardised sampling protocol for beetles and spiders, using a modification of the protocols of Cardoso *et al.* (2008a; 2009; 2008b). In each of the 30

sites 10 cylindrical soil cores of 8 cm width and approximately 10 cm depth (depending upon soil depth) were randomly sampled, with a minimum distance of 5 m between cores, to obtain between 4-5 litres of soil. Cores were mixed and 3.6 litres of soil was distributed across 3 Tullgren funnels and extracted into ethanol over a period of 7 days under a 40W incandescent light. Collembola were then separated from soil and other organic matter and stored at 4°C prior to DNA extraction.

DNA extraction, library preparation and sequencing

All sampled Collembola from a given sampling plot were combined for a single DNA extraction, performed using the DNeasy blood and tissue extraction kit (Qiagen). Illumina TruSeq libraries were constructed for each community extract for sequencing on the MiSeq platform (600-cycle 2x300 bp) using the DNA sequencing service of the Department of Biochemistry at the University of Cambridge. The 30 TruSeq libraries were divided into two groups of 15, and each group sequenced on a single Illumina MiSeq run.

Mitogenomic assembly pipeline

Raw paired-end reads were quality filtered using TRIMMOMATIC V.0.32 (Bolger *et al.* 2014) without trimming (Settings: WINDOWS=30, QUAL=10, MINLEN=300), and used directly to generate the assembly. The iterative de Bruijn Graph metagenome assembler IDBA-UD V.1.1.1 (Peng *et al.* 2012) was used to assemble reads, without precorrection, using a range of k-mers (23-123), a step of 10, a similarity threshold of 0.99, and a max mismatch of error correction of 0 (default options were used for all other assembler parameters). The resulting metagenome was filtered to retain only scaffolds with a length between 1 and 19 kbp. A mitochondrial nucleotide (nt) BLAST database (DB) was constructed using all complete arthropod mtDNA, extracted from the METAMIGA database (Feijão *et al.* 2006), and all Collembola mitochondrial gene sequences from GenBank (Benson *et al.* 2015), for a total alignment of 20,984 nts. This DB was used as a reference for a BLASTN (Camacho *et al.* 2013) search in order to filter for mtDNA scaffolds from the dataset (*e*-value ≤ $1 \bullet 10^{-10}$). As there are only 12 complete mtDNA genome sequences for Collembola, misidentification of scaffolds may arise from missing references. To address this and increase the sensitivity of our analysis, we undertook an

iterative search, adding scaffolds hits to the reference DB until no new putative mtDNA scaffolds were recovered.

Annotation and sequence alignment

All putative mtDNA scaffolds from the previous step were submitted to a modified version of MITOS WEBSERVER (Bernt *et al.* 2013), that accepts multifasta input files. The MITOS pipeline is designed to compute a *de novo* annotation of mitogenomic sequences, searching for amino acid (aa) homology to REFSEQ sequences, discriminating for gene duplication, and also annotating non-coding RNA sequences. Since mtDNA genes do not undergo splicing, have very short or no intergenic regions, and have a very conserved orientation and order, a pipeline was created to increase sensitivity and specificity by discriminating between false positive and true positive mtDNA scaffolds. The pipeline filters out scaffolds when the ratio between transcribed and non-transcribed regions, gene orientation or gene order are different from parameters found in known Collembola mtDNA. To align protein-coding sequences we used MACSE V.1.01B (Ranwez *et al.* 2011), a multiple alignment tool for coding sequences that uses translated aa sequences, while for non-coding sequences we used CLUSTAL W V.2 (Larkin *et al.* 2007). In order to reduce long gaps in the alignment due to incomplete gene sequences, an iterative alignment was carried out to remove sequences shorter than the 33% and 75% of the overall alignment, for coding and non-coding nt sequences respectively, realigning sequences after the short sequences were discarded. Gene alignments (nt and aa) were then concatenated. Scripts are available in the Github repository (https://github.com/francicco/IslandBiogeography).

Phylogenetic analysis

The concatenated alignment was partitioned according to codon positions when applicable using PARTITIONFINDER V.1.1.1_MAC (Lanfear *et al.* 2014), adopting the greedy search algorithm to select the partitioning scheme and the best model of evolution under the corrected Akaike Information Criterion (AICC). The proportion of invariant sites (I) was excluded from all models to avoid overfitting of the data (Cicconardi *et al.* 2013). To identify sequences of non-Collembola origin, we performed a phylogenetic analysis, extending the scaffold alignment with 17 known mtDNA genomes, including all 12 available for Collembola, three orders of Insecta (KF163965 [Lepidoptera: *Agrotis*

ipsilon], NC_000857 [Diptera: *Ceratitus capitata*], NC_003081 [Coleoptera: *Tribolium castaneum*]) and two classes of Myriapoda (HQ457012 [Pauropoda: *Pauropus longiramus*], NC_008453 [Sympyla: *Scutigerella causeyae*]). Phylogenetic analysis was carried out using Maximum Likelihood (ML), as implemented in RAxML v.8.1.20 (Stamatakis 2014), running 20 independent searches, using the rapid hill-climbing algorithm (-f d), and a rapid bootstrap analysis (settings: -f a, -# autoMRE, MRE-based bootstopping criterion), with an accuracy of 0.1 log likelihood units. By default, the bootstrap searches used the CAT approximation, while the search for the ML tree used the GTRGAMMA model for nucleotides.

Scaffold assignment to sampling site and genome geographic range analysis

To assign scaffolds to their sampling sites, raw reads from each sampling site were aligned back to scaffolds using Bowtie 2 v.2.2.4 (Misale 2014), using the local alignment 'very sensitive' setting. To assess the genomic similarity of scaffolds to Collembola sampled from geographic regions other than the three islands included in the present study, an alignment was produced for the 658bp barcode region from all scaffolds containing this region, retaining sequences with the complete, or almost complete (>80%) barcoding region. The alignment was then submitted to the Barcode of Life (BOLD) specimen identification database.

Results

Sampling

A total of 10,273 Collembola individuals were sampled across all 30 sampling sites (Fig. 1). Mean sample sizes per sampling site within each of the three islands were 105, 206 and 716 individuals for Terceira, Réunion and Tenerife respectively. Sampling of beetle and spider species from the same 30 sites yielded a total of 303 beetle species, of which two (0.6%) were found on more than one island (*Anaspis proteus* and *Ocypus aethiops*, sampled on both Tenerife and Terceira), and 175 spider species, of which three (1.7%) were found on more than one island (*Lathys dentichelis*, *Microlinyphia johnsoni* and *Steatoda grossa*, all three sampled on both Tenerife and Terceira).

Mitogenome assembly

Illumina MiSeq sequencing (300 bp, paired-ends) yielded 26 Gb of raw data from all 30 sampling sites. Low quality data was filtered out, and the remaining reads (70±2%, 19Gb) (Fig. 2a) used to build a *de novo* meta-assembly. A 1.1Gb metagenome was built (scaffolds length ≥ 1kb; N50: 3,011; N90: 1,344), comprised of 395,153 scaffolds. The iterative BLAST search gave a metagenome of 303 scaffolds (1.7Mb), with a mean of 5,555 bp and a maximum length of 29,645 bp (N50: 11,312; N90: 2,193). All 303 scaffolds were then annotated for coding and non-coding mitochondrial genes prior to filtering out scaffolds not consistent with Collembola mtDNA genomes based on the orientation and order of Collembola mitochondrial genes, reducing the scaffold number to 185.

Phylogenetic analysis

Following gene concatenation and alignment, a total of 36 partitions were found. Thirty-one partitions included only one codon, while five were described by two codons, for a total of 36 models. The GTR+Γ model was the most commonly represented, selected for 21 of the 36 partitions (See Table S1 in Supporting Information). Within the full ML phylogeny of all 185 scaffolds and complete mtDNA reference genomes, the two Myriapoda mtDNA genomes clustered together as a monophyletic (bs:93) sister clade to the Hexapoda (bs:93). The three Insecta mtDNA genomes clustered with two non-Collembola scaffolds forming a monophyletic (bs:100) sister clade to the other monophyletic clade (bs:96) comprised of all 12 Collembola mtDNA genomes and the remaining 183 putative Collembola scaffolds (Fig. 3). The two non-Collembola scaffolds clustered together with the Diptera and Lepidoptera reference genomes (bs:100). The more divergent of the two scaffolds contained the barcode region and all top hits to the BOLD specimen identification database were assigned to the family Chironomidae. Both scaffolds were sampled from Tenerife. The length distribution of the final assembly was skewed for short (< 5,000nt) and long scaffolds (> 10,000nt) (Fig. 2b), with 17 scaffolds containing only a single non-tRNA gene, 37 scaffolds contained all 13 protein coding genes (PCGs), and 13 scaffolds contained all 37 mitochondrial genes (Fig. 2c).

Scaffold assignment to sampling site

Through read mapping we were able to assign scaffolds to each of the 30 localities from where they were sampled. Tenerife and Terceira island yielded similar scaffold counts with 73 and 75 scaffolds respectively, while Réunion presented 101, with an average of 34 scaffolds per sampling site. Mapping scaffolds to sampling sites revealed a number of instances where multiple incomplete scaffolds had identical or near identical distributions and similar abundance patterns across sampling sites, often with very high abundances. We interpret these as scaffolds belonging to the same mtDNA lineage where intraspecific variation, or genomes from closely related species, are expected to complicate genome assembly (Nagarajan & Pop 2013). Limiting subsequent analyses to scaffolds no less than half the maximum scaffold size obtained (16,411 nts) guarantees that no two scaffolds belong to the same mtDNA lineage. We were able to extend this minimum threshold of 8,205 nts to 5,285 nts, as no scaffolds between 5,285 and 8,205 were found to share distribution and abundance patterns. This yielded a total of 68 scaffolds ranging from 5,285 nts to 16,411 nts inferred to belong to different presumed biological species.

Analyses of genome geographic range

Of the 68 scaffolds inferred to belong to different presumed biological species, 19 (28%) were sampled from more than one island. Seven were sampled on both Terceira and Tenerife, two on Terceira and Réunion, four on Tenerife and Réunion, and six were sampled on all three islands. From the 94 *cox1* (partial and complete) sequences, 70 have full (658bp) or near full length (>80%) barcode region sequences. These were submitted to the BOLD specimen identification database, with 21 (30%) having 99% or higher similarity to Collembola sampled from other geographic areas, of which 18 (26%) were a 100% match (Tables 1 & 2, Fig. 4). The mean matching of the remaining 49 sequences was 86.6%, with a standard deviation (SD) of 4.9%. The mean genetic *p*-distance among the 70 barcode sequences was 24%, with a SD of 4.5% and a smallest *p*-distance between barcode sequences of 3.2%. As it is possible that closely related barcode sequences could be representative of the same biological species, we repeated the analysis collapsing sequences with a *p*-distance divergence from each other of less than 10%. This resulted in the collapsing of five pairwise comparisons reducing data to

65 barcode sequences, yielding 20 matches of 99% similarity or higher to Collembola sampled from other geographic areas.

Discussion

To preserve and maintain fundamental soil ecosystem processes, it is essential to understand the vulnerability of soil communities under ongoing global change (Bardgett 2005; Wall & Nielsen 2012), something that first requires the characterisation of community composition and structure. This poses an immense challenge because of the inherent difficulties associated with mesofaunal species identification. In general, studies of the diversity of soil mesofauna based on morphology have frequently documented species with broad geographic distributions (Decaëns 2010), a pattern similar to that described for microbial communities (e.g. Chu *et al.* 2010). However, recent molecular studies have revealed high levels of phylogenetic and spatial structuring within traditionally recognised morphological species, indicating a high proportion of cryptic diversity with limited dispersal over deep evolutionary timescales. This is particularly true within the Collembola (Cicconardi *et al.* 2013; Cicconardi *et al.* 2010; Emerson *et al.* 2011; Garrick *et al.* 2008; Garrick *et al.* 2007; Porco *et al.* 2012; Stevens *et al.* 2006; Timmermans *et al.* 2005; Torricelli *et al.* 2009), where widely distributed morphospecies may thus represent one of two possibilities: (i) single species with broad geographic ranges, (ii) a complex of cryptic species, each with a more restricted geographic range.

Here we have taken a mitochondrial metagenomic approach to compare communities of Collembola from the mesofaunal component of the soil biome with aboveground arthropod communities, between which the nature of linkages remain poorly understood. Our bioinformatic pipeline recovered 185 mitochondrial scaffolds, of which 183 were confirmed to be of Collembola origin through phylogenetic analysis. The two non-Collembola scaffolds were inferred to be of dipteran origin through phylogenetic clustering, and taxonomic assignment through the BOLD specimen identification database. Because both dipteran scaffolds occurred at high frequency within single sites on Tenerife, they are most plausibly explained by sorting error when Collembola were separated from soil and other organic matter after Tullgren extraction.

Contrasting patterns of community similarity between aboveground and belowground arthropods

Of the 183 Collembola scaffolds, 68 could be unambiguously inferred to belong to different biological species. While many of the remaining 115 scaffolds may also belong to different biological species, assembly limitations associated with intraspecific variation, or genomes from closely related species, limit robust inferences concerning these. Of the 68 scaffolds, 21 were sampled on more than one island, six of which were sampled on all three archipelagos, and eight of which were shared between one island of Macaronesia and Réunion (see Fig. S1). This high community similarity at a large spatial scale within the soil mesofauna, with 31% of Collembola species shared among islands, contrasts dramatically with beetle and spider species, for which only 1% of species are found on more than one island, with no species sharing between the islands of Macaronesia and Mascarenes. The high similarity among soil mesofaunal communities implicates the dispersal of Collembola over large geographic distances, which is in contrast to recent molecular results that demonstrate dispersal limitation for Collembola species (Cicconardi *et al.* 2013; Cicconardi *et al.* 2010; Emerson *et al.* 2011; Garrick *et al.* 2008; Garrick *et al.* 2007; Stevens *et al.* 2006; Timmermans *et al.* 2005; Torricelli *et al.* 2009). These studies have shown that even over very limited geographic distances, measured in tens of kilometres, dispersal limitation may have maintained distinct community assemblages over timescales extending into millions of years (Cicconardi *et al.* 2013; Cicconardi *et al.* 2010; Garrick *et al.* 2008; Garrick *et al.* 2007). Our geographic analysis of barcode sequences also reveals that many species sampled from the three islands have large range sizes. Of the 70 complete or near complete barcode sequences, 30% were identical or near identical to individuals sampled from other geographic areas, often involving substantial geographical distances (Tables 1 & 2, Fig. 4).

Biological invasions as drivers of the loss of uniqueness among soil communities

Evidence for biological invasions by Collembola has been documented for several sub-Antarctic Islands (e.g. Gabriel *et al.* 2001; Greenslade 2002a; Greenslade & Wise 1984), New Zealand (Salmon 1941), Australia (e.g. Greenslade 2002b; King *et al.* 1985), Tenerife (Ramírez-González *et al.* 2013) and North America (Porco *et al.* 2013). With

regard to North America, Porco *et al.* (2013) used mtDNA sequencing for 5 species of Collembola introduced from Europe to North America to reveal multiple and apparently recurrent introductions of species, involving large numbers of founding individuals. Pyrosequencing has revealed a recent origin for some Collembola species on the island of Tenerife (Ramírez-González *et al.* 2013), but limited power for taxonomic assignment precluded inferences about the relative contribution of non-native introduced species to this pattern, and their contribution to community assembly. Thus, while previous studies have demonstrated biological invasions by Collembola species, and in the case of Porco *et al.* (2013), the potentially high frequency of introductions well-beyond the native distribution range, they say little about the scale of introduction events and subsequent invasions themselves.

By directly comparing belowground and aboveground insular forest arthropod communities we have been able to show that community similarity among Collembola, as measured by species sharing, is approximately 30x higher than that observed within aboveground arthropod communities. We also reveal that 30% of sampled Collembola species have mtDNA sequences that are identical, or near identical, to other regions of the world. The typically high genetic dissimilarity of the remaining species (Table 2: mean matching = 86.6%, SD = 4.9%) supports human-mediated introductions, as opposed to natural long distance dispersal, for species with identical, or near identical mtDNA sequences. This bimodal distribution of sequence similarities is consistent with high species establishment over an ecological time-scale (identical or near-identical mtDNA genomes) within a background of natural colonisation over an evolutionary time-scale (divergent mtDNA genomes).

Of the 19 Collembola scaffolds that both (i) contain the barcode region, and (ii) were sampled from more than one island, 9 (47%) are identical or near identical to Collembola sampled from other geographic regions, consistent with the non-indigenous origin of species occurring on more than one island. Our data reveal that introduced non-native species can dominate soil mesofaunal communities in three remote oceanic archipelagos. Of the 70 complete or near complete barcode sequences, 44 were sampled on the island of Réunion, of which 20 (45%) are inferred to be non-native species (Table 2). On the island of Terceira 17 of the 27 barcode sequences (63%) are consistent with a non-indigenous origin, while for Tenerife 22 of 25 (88%) are inferred to belong to non-native species (Table 2). These can be considered minimum estimates, as it is plausible

that other barcode sequences may belong to exotic species, but that these are not represented on the BOLD database.

Conclusion

Our findings reveal an unprecedented contribution of introduced non-native species to the Collembola component of soil mesofauna, greatly in excess of that observed for arthropods in adjacent subaerial biomes. These results suggest that it may be difficult to infer the integrity of soil mesofaunal communities by simple extrapolation from subaerial arthropod community patterns. It remains unknown to what extent ecosystem function may be compromised by introduced species and species invasion in the soil biome, and to address this we call for more attention on soil mesofaunal diversity and its spatial structure. Our comparison of mtDNA genome similarity to a global database of barcode sequences revealed many insular species are also found in multiple disparate geographic locations, consistent with intra and intercontinental introductions as well as those to remote oceanic islands. To understand the broader extent of introductions to soil mesofaunal community composition requires extending taxonomic sampling to other mesofaunal groups, such as Acari and Nematoda, and the analysis of other geographic regions and ecosystems.

Acknowledgements. We thank for following for assistance with field sampling and specimen sorting: Joëlle Sadeyen, Jacques Fournel, Samuel Danflous, Dominique Hoareau and Noémie Mollaret (Réunion); Isabel E. Amorim, Pedro Cardoso, Rui Carvalho, Simone Fattorini, Maria T. Ferreira, Orlando Guerreiro, Rui Nunes, Fernando Pereira, Ana Picanço, Carla Rego, François Rigal (Terceira); Rienk Apperloo, Manuel Arechavaleta, Salvador de La Cruz, Carla Díaz, Sara Ravagni, Benito Vispo, Guillermo Sánchez, Isabel Sancibrián, Nuria Macías, Nieves Zurita (Tenerife). This research was supported by the ERA-Net Net-Biome research framework, financed through: Canary Island Government ACIISI grants SE-12/02 (PO), SE-12/03 (JCC), SE-12/04 (BE), cofinanced by FEDER; Portuguese FCT-NETBIOME grant 0003/2011 (PB); French ANR-NETBIOME grant n°11-EBIM-001-01 (CT); Région Réunion council for research activities (DS), Université de La Réunion contract DGADD/PE/20120585 (DS). BCE was

also supported by Spanish MINECO grant CGL2013-42589-P, co-financed by FEDER, and Spanish MAGRAMA grant S20141203_002597 from the Organismo Autonomo Parques Nacionales. The field research station of Mare Longue (P.O.E. Reunion 2.02) and OSU Réunion provided logistical support for sampling on Réunion. The Natural Park of Terceira (Azores), The National Park of Réunion and the Cabildo of Tenerife (Canary Islands) provided the necessary authorizations for sampling. We are also grateful to the two anonymous reviewers who assessed an earlier version of this manuscript.

References

André HM, Noti M, -I., Lebrun P (1994) The soil fauna: the other last biotic frontier. *Biodiversity and Conservation*, **3**, 45-56.

Andújar C, Arribas P, Ruzicka F*, et al.* (2015) Phylogenetic community ecology of soil biodiversity using mitochondrial metagenomics. *Molecular Ecology*.

Bardgett RD (2005) *The biology of soil: a community and ecosystem approach* Oxford University press, USA.

Bardgett RD, Van der Putten WH (2014) Belowground biodiversity and ecosystem functioning. *Nature*, **515**, 505-511.

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

Bernt M, Donath A, Jühling F*, et al.* (2013) MITOS: Improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetic and Evolution*, **69**, 313-319.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114-2120.

Camacho C, Madden T, Ma N*, et al.* (2013) BLAST Command Line Applications User Manual. BLAST® Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US).

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, 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.

Chu HY, Fierer N, Lauber CL*, et al.* (2010) Soil bacterial diversity in the Arctic is not fundamentally different from that found in other biomes. *Environmental Microbiology*, **12**, 2998-3006.

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.

Crampton-Platt A, Timmermans JTN, Gimmel ML*, et al.* (2015) Soup to tree: the phylogeny of beetles inferred by mitochondrial metagenomics of a Bornean rainforest sample. *Molecular Biology and Evolution*.

Decaëns T (2010) Macroecological patterns in soil communities. *Global Ecology and Biogeography*, **19**, 287-302.

Dominati E, Patterson M, Mackay A (2010) A framework for classifying and quantufying the natural capital and ecosystem services of soils. *Ecological Economics*, **69**, 1858-1868.

Ehrenfeld JG (2010) Ecosystem Consequences of Biological Invasions. *Annual Review of Ecology Evolution and Systematics*, **41**, 59-80.

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.

Feijão PC, Neiva LS, de Azeredo-Espin AML, Lessinger AC (2006) AMiGA: the arthropodan mitochondrial genomes accessible database. *Bioinformatics*, **22**, 902-903.

Gabriel AGA, Chown SL, Barendse J*, et al.* (2001) Biological invasions of Southern Ocean islands: the Collembola of Marion Island as a test of generalities. *Ecography*, **24**, 421-430.

Garrick RC, Rowell DM, Simmons CS, Hillis DM, Sunnucks P (2008) Fine-scale phylogeographic congruence despite demographic incongruence in two low-mobility saproxylic springtails. *Evolution*, **62**, 1103-1118.

Garrick RC, Sands CJ, Rowell DM, Hillis DM, Sunnucks P (2007) Catchments catch all: long-term population history of a giant sprigtail from the southeast Australian highlands - a multigene approach. *Molecular Ecology*, **16**, 1865-1882.

Gaston KJ (1991) The magnitude of global insect species richness. *Conservation Biology*, **5**, 283-296.

Gillett CPDT, Crampton-Platt A, Timmermans JTN*, et al.* (2014) Bulk De Novo Mitogenome Assembly from Pooled Total DNA Elucidates the Phylogeny of Weevils (Coleoptera: Curculionoidea). *Molecular Biology and Evolution*, **31**, 2223-2237.

Gómez-Rodríguez C, Crampton-Platt A, Timmermans JTN, Baselga A, Vogler AP (2015) Validating the power ofmitochondrial metagenomics for community ecology and phylogenetics of complex assemblages.

Greenslade P (2002a) Assessing the risk of exotic Collembola invading subantarctic islands: prioritising quarantine management. *Pedobiologia*, **46**, 338-344.

Greenslade P (2002b) Systematic composition and distribution of Australian cave collembolan faunas with notes on exotic taxa. *Helictite*, **38**, 11-16.

Greenslade P, Wise KAJ (1984) Additions to the Collembolan fauna of the Antarctic. *Transactions of the Royal Society of Australia*, **108**, 203-206.

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

Heemsbergen DA, Berg MP, Loreau M*, et al.* (2004) Biodiversity affects on soil processes explained by interspecific functional dissimilarity. *Science*, **306**, 1019-1020.

Hopkin SP (1997) *Biology of the Springtails (Insecta: Collembola).* Oxford University Press, Oxford.

King KL, Greenslade P, Hutchinson KJ (1985) Collembolan associations in natural versus improved pastures of the New-England tableland, NSW: distribution of native and introduced species. *Australian Journal of Ecology*, **10**, 421-427.

Lanfear R, Calcott B, Kainer D, Mayer C, Stamatakis A (2014) Selecting optimal partitioning schemes for phylogenomic datasets. *BMC Evolutionary Biology*, **14**, 82.

Larkin MA, Blackshields G, Brown NP*, et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947-2948.

Lavelle P, Decaëns T, Aubert M*, et al.* (2006) Soil invertebrates and ecosystem services. *European Journal of Soil Biology*, **42**, S3-S15.

Lomolino MV, Riddle BR, Whittaker RJ, Brown JH (2010) *Biogeography*, 4th edn. Sinauer Associates, Inc., Sunderland.

Misale C (2014) Accelerating Bowtie2 with a lock-less concurrency approach and memory affinity. In: *22nd Euromicro International Conference on Parallel, Distributed, and Network-Based Processing*, pp. 578-585, Torino.

Nagarajan N, Pop M (2013) Sequence assembly demystified. *Nature Reviews Genetics*, **14**, 157-167.

Peng Y, Leung HCM, Yiu SM, Chin FYL (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, **28**, 1420-1428.

Porco D, Bedos A, Greenslade P*, et al.* (2012) Challending species delimitation in Collembola: cryptic diversity among common springtails unveiled by DNA barcoding. *Invertebrate Systematics*, **26**, 470-477.

Porco D, Decaëns T, Deharveng L*, et al.* (2013) Biological invasions in soil: DNA barcoding as a monitoring tool in a multiple taxa survey targeting European earthworms and springtails in North America. *Biological Invasions*, **15**, 899-910.

Ramírez-González R, Yu DW, Bruce C*, et al.* (2013) PyroClean: Denoising pyrosequences from protein-coding amplicons for the recovery of interspecific and intraspecific genetic variation. *PLoS ONE*, **8(3):e57615**.

Ranwez V, Harispe S, Delsuc F, Douzery EJP (2011) MACSE: Multiple alignment of coding sequences accounting for frameshifts and stop codons. *PLoS ONE*, **6**, e22594.

Salmon JT (1941) The Collembolan Fauna of New Zealand, including a discussion of its distribution and affinities. *Transactions of the Royal Society of New Zealand*, **70**, 282-431.

Schaefer M, Schauermann J (1990) The soil fauna of beech forests: comparison between a mull and a moder soil. *Pedobiologia*, **34**, 299-314.

Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312-1313.

Stevens MI, Greenslade P, Hogg ID, Sunnucks P (2006) Southern hemisphere springtails: could any have survived glaciation of Antarctica? *Molecular Biology and Evolution*, **23**, 874-882.

Tang M, Tan M, Meng G*, et al.* (2014) Multiplex sequencing of pooled mitochondrial genomes - a crucial step toward biodiversity analysis using mito-metagenomics. *Nucleic Acids Research*, **42**, e166.

Timmermans MJTN, Ellers J, Marien J*, et al.* (2005) Genetic structure in *Orchesella cincta* (Collembola): strong subdivision of European populations inferred from mtDNA and AFLP markers. *Molecular Ecology*, **14**, 2017-2024.
Torricelli G, Carapelli A, Convey P*, et al.* (2009) High divergence across the whole mitochondrial genome in the "pan-Antarctic" springtail *Friesea grisea*: evidence for cryptic species? *Gene*, **449**, 30-40.
Violle C, Reich PB, Pacala SW, Enquist BJ, Kattge J (2015) The emergence and promise of functional biogeography. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 13690-13696.
Wall DH, Fitter AH, Paul EA (2005) Developing new perspectives from advances in soil biodiversity research. In: *Biological diversity and function in soils.* (eds. Bardgett RD, Usher MB, Hopkins DW), pp. 3-27. Cambridge University Press, Cambridge.
Wall DH, Nielsen UN (2012) Biodiversity and ecosystem services: is it the same below ground? *Nature Education Knowledge*, **3**, 8.
Wardle DA, Bardgett RD, Klironomos JN*, et al.* (2004) Ecological linkages between aboveground and belowground biota. *Science*, **304**, 1629-1633.
Yang J, Kloepper JW, Ryu CM (2009) Rhizosphere bacteria help plants tolerate abiotic stress. *Trends in Plant Science*, **14**, 1-4.
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.

Data accessibility statement: All DNA sequence data from this study has been submitted to the NCBI Sequence Read Archive (SRA) and is available under BIOproject accession number PRJNA357014.

Author contributions: BCE, CT, PAV, PO, DS, JMFP and JJC conceived the study design and sampling program. BCE, PAV and DS coordinated Collembola sampling. PO, HL, AJPD, DS, JC and PAV coordinated beetle and spider sampling and species identification. BCE and FC generated and analysed the metagenomic data and wrote the manuscript. All authors commented on the final version of the manuscript.

Figure 1. Map of sampling sites within the oceanic islands of Terceira (Azores archipelago), Tenerife (Canary Islands) and Réunion (Mascarene Islands). Sampling sites are numbered 1-10 within each island.

Figure 2. Raw data, scaffolds and annotation statistics. (a) Histogram showing the number of raw and quality filtered reads for each sampling site. (b) Distribution of the frequency of lengths for the assembled mtDNA scaffolds. (c) The frequency of protein coding genes (PCGs) per scaffold. The highest frequencies are represented by 43 scaffolds having two PCGs and 37 scaffolds with all 13 PCGs.

Figure 3. Maximum likelihood phylogeny of the 185 mtDNA scaffolds with the 17 reference mtDNA genomes (12 Collembola, 3 Insecta and 2 Myriapoda). Scaffolds matching barcode sequences (99% or higher similarity) are also shown with the affiliated species name or family, when known. Bootstrap support values are shown on branches, with values below 75 not shown. Bootstrap values are also represented by branch colour and width (see legend). For scaffolds longer than 5,200nt a heat map shows the occurrence of the scaffold on each of the thirty sampling sites (blue shading), while its length with a histogram (red shading).

Figure 4. Diagram showing species sharing among sampled islands and ecoregions for 21 Collembola mtDNA scaffolds with 99% or higher matching to the BOLD database. Each coloured ribbon shows the presence of a species on the three focal islands and ecoregions of the world. Ribbon width is proportional to numbers of localities within ecoregions.

Table 1. Scaffolds matching to barcodes in Barcode Of Life Database (BOLD). For each match the island occurrence of the scaffold and the ecoregion location of the reference barcoded sample are listed. Ne: Nearctic; Pa: Palearctic; At: Afrotropical; Im: Indo-Malayan; Oc: Oceania; Au: Australasia; Nt: Neotropical.

	Sample sites						Number of localities per ecoregion						
Scaffold	**Terceira**	**Tenerife**	**Réunion**	**Sp. Id. (%)**	**Order**	**Species**	**Ne**	**Pa**	**At**	**Im**	**Oc**	**Au**	**Nt**
6241	0	1	0	100	Poduromorpha	*Gen. sp.*			1				
6287	2	6	0	100	Poduromorpha	*Ceratophysella gibbosa*		1				1	
12586	3	7	1	100	Poduromorpha	*Gen. sp.*		2					
7124	3	0	0	99	Poduromorpha	*Gen. sp.*		1	1				
62376	2	6	0	100	Poduromorpha	*Mesaphorura sp.*	1	2				1	
6379	0	0	9	100	Poduromorpha	*Gen. sp.*		1	1	1			
6532	0	0	9	100	Poduromorpha	*Gen. sp.*		2				1	
6543	0	0	2	100	Poduromorpha	*Gen. sp.*	2						
6480	0	4	0	100	Poduromorpha	*Deuteraphorura sp.*							
36935	0	4	0	100	Poduromorpha	*Protaphorura sp.*		2					
6653	7	0	1	100	Entomobryomorpha	*Gen. sp.*		1					
7289	0	0	4	100	Entomobryomorpha	*Desoria sp.*	1		1				
8783	4	10	1	100	Entomobryomorpha	*Parisotoma notabilis*	1	10					
5537	7	2	0	100	Entomobryomorpha	*Parisotoma notabilis L2*	2	8					
112296	3	1	0	100	Entomobryomorpha	*Tomocerus minor*	1	1					
6464	1	1	0	100	Entomobryomorpha	*Lepidocyrtus curvicollis*		2					
6802	0	0	6	100	Entomobryomorpha	*Gen. sp.*					1		
35470	0	0	6	100	Entomobryomorpha	*Gen. sp.*			1			1	
19920	0	0	4	99	Entomobryomorpha	*Paronellinae sp.*				2			
7305	0	8	0	100	Neelipleona	*Gen. sp.*			1				
6565	1	8	2	99	Neelipleona	*Megalothorax minimus*		2	1				1

Table 2. Summary of the island occurrence and matching to barcodes in Barcode Of Life Database (BOLD) for the 70 scaffolds with full (658bp) or near full length (>80%) barcode region sequences. Light shading indicates scaffolds inferred to be from non-native species either because (i) they are sampled on more than one island, (ii) match to geographically distant samples from BOLD, or both. Dark shading indicates scaffolds that are not inferred to be from non-native species.

scaffold	Terceira	Tenerife	Réunion	Highest BOLD match
5537	1	1	0	100
6053	0	0	1	85.11
6208	1	0	0	86.09
6241	0	1	0	100
6245	1	1	1	84.25
6287	1	1	0	100
6379	0	0	1	100
6463	1	0	1	86.85
6464	1	1	0	100
6480	0	1	0	100
6483	1	0	0	85.63
6513	1	0	0	94.8
6532	0	0	1	100
6543	0	0	1	100
6565	1	1	1	99.69
6593	0	0	1	91.07
6641	0	1	0	98.93
6653	1	0	1	100
6702	0	0	1	85.79
6800	1	0	0	84.05
6802	0	0	1	100
6817	0	0	1	82.8
6864	0	1	1	86.53
7003	1	1	1	82.97
7032	0	0	1	85.83
7088	0	0	1	81.64
7124	1	0	0	99.36
7194	0	0	1	81.36
7289	0	0	1	100
7305	0	1	0	100
7540	1	0	0	86.85
7544	0	0	1	84.52
7582	0	0	1	83.79
7942	0	1	1	86.43
8054	0	0	1	85.39
8612	0	1	0	87.85
8783	1	1	1	100

9320	0	0	1	82.39
9391	0	1	1	81.11
9749	0	0	1	82.8
12586	1	1	1	100
12624	1	1	0	85.3
13090	0	0	1	88.74
15103	0	1	1	86.04
16605	1	1	1	96.86
17413	0	0	1	88.16
19920	0	0	1	99.08
22943	1	1	0	92.66
23520	0	0	1	88.15
30106	0	0	1	85.19
35470	0	0	1	100
35676	0	0	1	85.66
36935	0	1	0	100
41303	0	0	1	80.21
42411	1	0	0	80.95
48363	0	0	1	81.66
52670	1	0	0	93.43
62376	1	1	0	100
69515	0	0	1	85.57
73182	0	0	1	83.94
76953	1	0	0	86.55
83813	0	0	1	85.13
104491	0	0	1	86.11
104533	0	0	1	85.71
112296	1	1	0	100
113570	1	0	0	83.93
121342	1	0	0	95.7
153098	0	0	1	85.71
167705	0	1	0	91.06
212789	1	1	0	97.86

Total	**27**	**25**	**44**	**70**
Matches to BOLD	11 (41%)	12 (48%)	11 (25%)	21 (30%)
Multi-island	14 (52%)	18 (72%)	19 (43%)	20 (29%)
Presumed introductions	17 (63%)	22 (88%)	20 (45%)	32 (46%)