Community assembly and metaphylogeography of soil biodiversity: insights from haplotype-level community DNA metabarcoding within an oceanic island

Carmelo Andujar[1], Paula Arribas[2], Heriberto López[3], Yurena Arjona [4], Antonio Pérez-Delgado[5], Pedro Oromí[3], Alfried Vogler[6], and Brent Emerson[5]

[1]The Natural History Museum of London
[2]IPNA
[3]Affiliation not available
[4]Real Jardin Botanico
[5]IPNA-CSIC
[6]Imperial College/Natural History Museum of London

January 14, 2022

Abstract

Most of our understanding of island diversity comes from the study of aboveground systems, while the patterns and processes of diversification and community assembly for belowground biotas remain poorly understood. Here we take advantage of a relatively young and dynamic oceanic island to advance our understanding of eco-evolutionary processes driving community assembly within soil mesofauna. Using whole organism community DNA (wocDNA) metabarcoding and the recently developed metaMATE pipeline, we have generated spatially explicit and reliable haplotype-level DNA sequence data for soil mesofaunal assemblages sampled across the four main habitats within the island of Tenerife. Community ecological and metaphylogeographic analyses have been performed at multiple levels of genetic similarity, from haplotypes to species and supraspecific groupings. Broadly consistent patterns of local-scale species richness across different insular habitats have been found, whereas local insular richness is lower than in continental settings. Our results reveal an important role for niche conservatism as a driver of insular community assembly of soil mesofauna, with only limited evidence for habitat shifts promoting diversification. Furthermore, support is found for a fundamental role of habitat in the assembly of soil mesofauna, where habitat specialism is mainly due to colonisation and the establishment of preadapted species. Hierarchical patterns of distance decay at the community level and metaphylogeographical analyses support a pattern of geographic structuring over limited spatial scales, from the level of haplotypes through to species and lineages, as expected for taxa with strong dispersal limitations. Our results demonstrate the potential for wocDNA metabarcoding to advance our understanding of biodiversity.

Introduction

Colonisation and speciation, together with extinction, are key processes contributing to island diversity and core processes within models of island biogeography (e.g. MacArthur & Wilson, 1963, 1967; Hubbell, 2001; Rosindell *et al.* , 2011). Most of our understanding of island diversity, and the mechanisms of diversification and community assembly on islands, comes from the study of aboveground systems (e.g., Gillespie & Roderick, 2002; Warren *et al.* , 2014; Patiño *et al.* , 2017), while the patterns and processes of importance for underground biotas remain poorly understood (FAO report, 2020). This lack of knowledge presents a major limitation to understanding island biodiversity and dynamics, as patterns and processes are not necessarily coupled between aboveground and belowground components of ecosystems (Bardgett & van der Putten, 2014; Shade et al., 2018)

Soil biodiversity, in particular soil mesofauna (i.e. small-bodied invertebrates measuring between 0.1 and 2 mm), is globally poorly understood (Cameron et al., 2018; Decaëns, 2010; White et al., 2020). Knowledge regarding fundamental biological and ecological traits of soil mesofauna is absent for most species. For example, dispersal dynamics within soil fauna remains an open and central question in soil biodiversity research (Ettema & Wardle, 2002; Thakur et al., 2019). Within insular settings, soil faunal diversity is expected to be strongly influenced by variation among species for dispersal capacity and niche breadth, as these traits underpin both island colonization and within island processes of population structure and speciation (Emerson & Gillespie, 2008; Gillespie et al., 2012; Kisel & Barraclough, 2010; Warren et al., 2014). Thus, insular systems provide an important focus for the development of a broader understanding of how dispersal and niche traits shape soil mesofaunal biodiversity.

Arthropod mesofaunal lineages typically exhibit various adaptations to soil environments, including the reduction of wings, eyes, and legs, and are thus likely to be limited in their propensity for active dispersal (Decaëns, 2010; Wardle, 2002). When extrapolated over extended periods of evolutionary time, such dispersal limitation is consistent with the high turnover across limited spatial scales and high local endemicity that has been reported for soil mesofaunal lineages (e.g., Andújar et al., 2017; Arribas, Andújar, Salces-Castellano, Emerson, & Vogler, 2021; Cicconardi, Nardi, Emerson, Frati, & Fanciulli, 2010; Collins, Hogg, Convey, Barnes, & McDonald, 2019; Morek, Surmacz, López-López, & Michalczyk, 2021). However, it has also been argued that their small body size and often high local abundances may increase the probability of passive dispersal and long-distance movement (Ettema & Wardle, 2002; Thakur et al., 2019), supporting the "Everything is everywhere but environment selects" hypothesis for soil mesofauna (Fenchel & Finlay, 2004; Finlay, 2002). In the context of oceanic islands, if passive dispersal is sufficiently high, island colonisation by soil fauna lineages should be a recurrent process maintaining species cohesion between islands and source regions, and panmictic populations at intra-island scales (Fig. 1A). In contrast, if passive dispersal is strongly constrained for soil fauna, it is reasonable to assume that colonization will occur primarily through sporadic events of long-distance dispersal (i.e. LDD events, Nathan, 2005), and that geographic speciation, even within islands, will play a more important role in community assembly (Fig. 1A).

While island colonisation will depend on dispersal capacity, successful establishment is also reliant upon species-specific traits related to climatic niche breadth. In general, islands have been proposed to favour generalist species, either by colonization filters that select for species with wide niche breadth (ecological tolerance) (Gaston, 2003; Reaka, 1980) or through lower levels of competition favouring ecological release following colonisation (Olesen, Eskildsen, & Venkatasamy, 2002). It has also been demonstrated that climatic gradients within islands can be characterised by very differentiated invertebrate communities, comprising species with strong habitat specificity (Lim et al., 2021). Ecological speciation involving climatic-niche shifts has been described as an essential process generating diversity within oceanic island biotas (Gillespie, Roderick, & Howarth, 2001). However, recent studies focused on arthropod assemblages have highlighted an important role for climatic niche conservatism as a driver of community assembly and diversification within islands (Lim et al., 2021; Salces-Castellano et al., 2020).

Habitat specialisation and climatic niche conservatism across soil fauna lineages has been poorly explored. However, previous studies on the community assembly of soil mesofauna have shown strong evidence for specialisation to open versus forested vegetation types (Arribas, Andújar, Salces-Castellano, et al., 2021; Caruso, Taormina, & Migliorini, 2012), with further evidence for specialisation among different forest types (Noguerales et al., 2021). Oceanic islands that have remained geographically isolated over evolutionary timescales and present variation in habitat types provide near-ideal conditions to explore further the relative contribution of generalist and specialist species composing soil island biotas and the role of habitat-shifts in the process of diversification within insular settings.

Here we take advantage of a relatively young and dynamic oceanic island to advance our understanding of eco-evolutionary processes driving community assembly within soil mesofauna. We achieve this by appling whole organism community DNA (wocDNA) metabarcoding to soil mesofaunal communities sampled across the four dominant habitats within the island of Tenerife. Tenerife is one of the seven principal Canary

Islands, an archipelago within the subtropical region of the North Atlantic Ocean. The oldest massif of Tenerife emerged approximately 9 Ma, but most of its 2,034 km^2 landscape dates back to less than 3 Ma, with extensive volcanic activity in the last 2 Ma (Ancochea, Maria, Ibarrola, Cendrero, & Coello, 1990; Carracedo et al., 2004). Maximum altitude exceeds 3,000 m, giving rise to an altitudinal-zonal distribution of main habitat types, strongly mediated by trade winds.

We use spatially explicit and reliable (Andújar et al., 2021) haplotype-level DNA sequence data for the mtDNA COI gene to conduct community ecological and metaphylogeographic (Turon, Antich, Palacín, Præbel, & Wangensteen, 2019) analyses at multiple levels of genetic similarity, from the level of haplotypes, through to species and supraspecific groupings. We estimate local, habitat-level, and island-level richness, together with measures of local endemicity and the structuring of community variation across habitats and geographic distance. We use these data for a joint evaluation of the patterns and processes driving the diversity and structure of soil mesofauna from the level of the community down to individual lineages, and address the following four questions. Is dispersal limitation of soil mesofauna sufficient to drive geographic structuring of communities and lineage diversification? How do habitat specificity and habitat shift contribute to community assembly? What is the relative importance of spatial*vs* environmental processes as drivers of community structure and lineage diversification? How do wocDNA diversity estimates compare with more traditional assessments, and how do they compare to similar estimates from comparable continental soils?

Material and Methods

Soil sampling and mesofauna extraction

Fifty-two sites were sampled across the main habitats of the island of Tenerife (Canary Islands), including 16 sites on laurel forest, 12 on thermophilous woodland, 12 on pine forest, and 12 on dry scrubland (Fig. 1B). Distances between sites ranged from a few meters to a maximum of 75 km (Fig. 1B, Table S1). Each site was sampled for: (i) the superficial soil layer (SUP) by removing one square metre of leaf litter and humus, and; (ii) the corresponding deep soil layer (DEEP), collected by extracting 20 litres of soil to a depth of approximately 25-30 cm below where the superficial layer was collected. SUP and DEEP soil samples were processed following the flotation–Berlese–flotation protocol (FBF) of Arribas *et al* . (2016). Briefly, the FBF protocol is based on the flotation of soil in water, which allows the extraction of the organic (floating) matter containing the soil mesofauna from soil samples. Subsequently, the organic portion is placed in a modified Berlese apparatus to capture specimens alive and preserve them in absolute ethanol. The last part of the FBF protocol includes additional flotation and filtering steps of the ethanol-preserved arthropods using 1-mm and 0.45-µm wire mesh sieves to yield macrofaunal (retained in the 1-mm mesh) and mesofaunal fractions (retained in the 0.45-µm mesh). Additional manual sorting was performed to pool together Coleoptera specimens from both fractions. The remaining macrofauna was stored and not used for this study. This procedure generates two 'clean' bulk specimen subsamples for each soil layer, one including all adult and larval Coleoptera (beeltles) and a second with the smallest mesofauna typically dominated by Acari (mites) and Collembola (springtails).

DNA extraction, PCR amplification, and Illumina sequencing

Bulk specimen subsamples were DNA extracted using the DNeasy Blood and Tissue Kit (Qiagen) in a Kingfisher Flex robotic system (Thermo Scientific). The bulk of Coleoptera was extracted non-destructively by splitting specimens into parts or puncturing the body. The mesofauna sample was homogenized in 1.5 ml vials with glass pestles. DNA extracts were quantified using Nanodrop 1000 UV–Vis spectrophotometer (Thermo Scientific), and the corresponding Coleoptera and mesofaunal subsample pairs were combined at a ratio of 1:10 in the amount of DNA (in accordance with the expected species diversity for these two fractions (Arribas et al., 2021)). The bc3' fragment, corresponding to the 3' 418 bp of the COI barcode region was amplified, using tailed primers corresponding to the Illumina P5 and P7 sequencing adapters for subsequent library preparation. Three independent PCR reactions were performed for each sample, and amplicons were pooled. All information regarding primers, PCR reagents and conditions is provided in Table S2. Amplicons were then cleaned using Ampure XP magnetic beads and used as the template for limited-cycle secondary

PCR amplification to add dual-index barcodes and the Illumina sequencing adapters (Nextera XT Index Kit; Illumina, San Diego, CA, USA). Metabarcoding libraries were then sequenced on an Illumina MiSeq sequencer (2 x 300 bp paired-end reads), dedicating approximately 1% of a flow cell to each library, producing paired reads (R1 and R2) with a dual tag combination for each sample.

Bioinformatics read processing

Raw reads were quality checked in Fastqc (Babraham Institute, 2013). Primers were trimmed using fastx_trimmer and reads processed in Trimmomatic (Bolger, Lohse, & Usadel, 2014) using *TRAILING:20* . Individual libraries were further processed, implementing several steps of the Usearch (Edgar, 2013) pipeline: reads were merged (option*mergepairs -fastq_minovlen 50, -fastq_maxdiffs 15*), quality-filtered (*Maxee = 1*), trimmed to full length amplicons of 416-420 bp (*-sortbylength*), dereplicated (*-fastx_uniques*), and denoised and chimera checked (*-unoise3, -minsize 2*). Denoised reads from all 104 libraries, representing putative haplotypes, were then combined and dereplicated to yield a set of unique sequences across all samples, referred to as amplicon sequence variants (ASVs from here on; Callahan *et al.* , 2016). MEGAN V5 (Huson, Auch, Qi, & Schuster, 2007) with the lowest common ancestor (LCA) algorithm was used to compute the taxonomic affinity of each ASV. This classification was based on the result of a BLAST search (*blastn -outfmt 5 -evalue 0.001*) against a reference library including the NCBI *nt* database (Accessed at June 2018) together with an additional 559 unpublished taxonomically assigned Iberian sequences of Acari, Collembola, and Coleoptera.

ASVs classified by MEGAN as Acari, Collembola, and Coleoptera were processed with metaMATE (Andújar et al., 2021). MetaMATE evaluates the survival of ASVs under alternative filtering procedures based on the relative abundance of co-distributed ASVs. Briefly, the application of metaMATE involves a six-step procedure: (i) identification of verified authentic ASVs (va-ASVs) by 100% matching against a reference COI sequence; (ii) identification of ASVs including indels or STOP codons as verified non-authentic ASVs (vna-ASVs); (iii) generation of a community table with read-counts (ASV abundance) by sample against the complete collection of reads (i.e., before the dereplicating and denoising steps) using Usearch (-search_exact option); (iv) filtering with a range of criteria and threshold values; (v) evaluation of the survival of va-ASVs and vna-ASVs, and (vi) estimation of the predicted number of a-ASVs and na-ASVs, for every filtering iteration. Filtering parameters can thus be chosen according to desired stringency for the survival of a-ASVs and na-ASVs. (see Andújar *et al.* , 2021 for further details)

The following input files were used to run MetaMATE: (i) the set of unique ASVs (-A option); (ii) a reference dataset (-R) for the identification of va-ASVs, including all BOLD sequences for Acari, Collembola, and Coleoptera (downloaded at May 2020) plus 1,011 sequences from specimens collected at the Iberian Peninsula and the Canary Islands; (iii) all reads prior to the dereplicating and denoising steps (-L), and; (4) the specification file including filtering criteria and parameters to be evaluated (-S) (parameters used: –refmatchlength 350 –refmatchpercent 100 –expectedlength 418). Filtering was explored using both (i) minimum absolute and minimum percentage abundance by library and (ii) minimum percentage abundance by library and lineages at 20% divergence, and all pair combinations of these (See MetaMATE tutorial for details). Analyses were conducted independently for Acari, Coleoptera, and Collembola. Filtering parameters were selected for each taxon to maximize the number of surviving va-ASVs while maintaining the predicted contribution of na-ASVs to the final dataset to be [?] 5%. Finally, the filtered set of ASVs was further filtered to reduce any potential cross-contamination problems across samples by removing ASVs with four or fewer reads from each library. Community tables of fully filtered haplotypes were then transformed into incidence (presence/absence) data for further analyses.

Community richness and structure at multiple thresholds of genetic similarity

Filtered ASVs were used to generate a UPGMA tree using F84 corrected genetic distances, within which haplotypes were grouped into clusters of genetic similarity at 0.5%, 1%, 2%, 3%, 5%, 8% and 15% thresholds for the analysis of α and β diversity from intraspecific haplotype level (h) variation through to supraspecific lineages. Subsequent community-level analyses were performed for either a selection of hierarchical levels (h

, 3%, and 15% clustering) or the complete set of thresholds.

To test for significant differences in community richness (α diversity) among different habitats and soil layers for h , 3%, and 15%-level clusters, repeated-measures ANOVAs were conducted using habitat and soil layer as grouping factors and sampling site as a within-subject factor. DEEP and SUP samples were then combined within each sampling site (n=52), and Kruskal-Wallis rank sum tests were conducted using habitat as a grouping factor to assess whether α diversity differed between the communities of each of the four habitats. Endemicity at the scale of individual sampling sites was also calculated for h , 3%, and 15%-level clusters measured as the proportion of total lineages within a given sampling site that occur exclusively within that sampling site. Kruskal-Wallis rank sum tests were conducted to test for differences in community endemicity among the four habitats. Total observed richness (g diversity) and accumulation curves (random method, 1000 permutations, *specaccum* function) were estimated for each habitat for h , 3% and 15%-level clusters, and total extrapolated richness (Chao equation, *specpool* function) by habitat was estimated. Total community dissimilarity across the communities of each habitat was estimated at all clustering levels, and pairwise community matrices were generated using total β diversity (Sorensen index, βsor) and its additive turnover (Simpson index, βsim) and nestedness (βsne) components (Baselga & Orme, 2012). Community composition matrices were used for non-parametric multidimensional scaling (NMDS) for h , 3% and 15%-level clusters, and plots were created with the *ordispider* option to visualise the compositional ordination of communities according to their respective habitat. Permutational ANOVAs were conducted over the community dissimilarity matrices using 999 permutations and the habitat as the grouping factor.

Variation in community composition with spatial distance was assessed following the 'multi-hierarchical macroecology' approach of Baselga*et al.* (2013), where distance decay of similarity is contrasted across hierarchical levels. For each habitat, the relationship between community similarity and spatial distance between sampled sites (1 – pairwise β diversity, see above) was assessed for each clustering level. The spatial distance was calculated using the R package *gdistance*(van Etten, 2017), which uses Tobler's hiking function to provide the shortest route between two points given the slope of the terrain (m) (Tobler, 1993). Pairwise calculations were made among sites within the same habitat. The lowest and highest elevations of each habitat within our sampling sites were used to constrain altitudinal movement, to avoid shortest paths transgressing a different habitat. A negative exponential function was used to adjust a generalised linear model (GLM) with Sorensen similarity as the response variable, spatial distance as the predictor, log link and Gaussian error, and maintaining untransformed spatial distances (Gómez-Rodríguez & Baselga, 2018). Fractal patterning (power-law function) among distance-decay curves was assessed by a log-log Pearson correlation across clustering levels for (a) the number of lineages, (b) the initial similarity, and (c) the mean similarity of the distance-decay curves. High correlation values are indicative of self-similarity in lineage branching (i.e., number of lineages) and spatial geometry of lineage distributional ranges (i.e., initial and mean similarity; Baselga *et al.* , (2015)), which are predicted under a predominant neutral process of community evolution. Analyses were also conducted to assess the relationship between community similarity and environmental distance, computed using Gower's distance over the elevation and 19 bioclimatic variables (from WORLDCLIM at 30 arc-seconds resolution), characterising each sampling site (Table S3). When significant relationships were found, variance partitioning was conducted to assess the fractions of variance in community dissimilarity that are uniquely and jointly explained by spatial and environmental distance.

Finally, we compared biodiversity measures for haplotypes and 3% OTUs for the four habitat types in Tenerife with those obtained by Arribas*et al.* (2020) in three forest and three grassland sampling regions in a continental setting (n=12 for each habitat on each sampling region). Kruskal-Wallis rank sum tests were used to compare α diversity by sample with insularity (Tenerife island n=52; continent n = 72) and sampling region as grouping factors. Comparisons of β diversity by sampling region were restricted to a comparable spatial scale of 15 km, conducting a Kruskal-Wallis rank sum test with insularity as a grouping factor. Comparisons of β diversity were repeated for intervals of spatial distance between 0-5 km, 5-10 km, and 10-15 km. Finally, g diversity (total species richness) was estimated for each habitat and region using accumulated haplotypes and 3% OTUs across 12 community samples (using *specaccum* function when the available number of samples was higher). All analyses were performed using the R-packages*vegan* (Oksanen

et al., 2016), *cluster* (Maechler, Rousseeuw, Struyf, Hubert, & Hornik, 2021), *PMCMR* (Pohlert, 2014), *hier.part* (Mac Nally & Walsh, 2004), *ecodist*(Goslee & Urban, 2007), and *betapart* (Baselga & Orme, 2012).

Lineages characterisation and meta-phylogeographical patterns

Three per cent and a 15% similarity clusters were used, whereby 3% clusters are considered a proxy to species, and from here on referred to as "OTUs"; while 15% clusters are lineages of one or more species and are hereon referred to as "15% lineages". We evaluated the genetic diversity, distribution, and degree of habitat specificity for each OTU and 15% lineage. We then tested the relative roles of the habitat and the geographical distance in the diversification of soil fauna within the island. The number of haplotypes was recorded as a measure of the genetic richness of each OTU, and OTUs were classified as "single haplotype" or "multiple haplotypes". At the level of 15% lineages and under the assumption that each arises from a single colonisation of Tenerife, the number of OTUs within each 15% lineage was used to classify each lineage as "non-diversified" or "diversified" according to whether they included one or multiple OTUs within the island. BLAST search (*blastn -outfmt 5 -evalue 0.001*) against a reference library including all sequences on BOLD (database downloaded at 3-07-2020), together with COI sequences from southern Iberia (Arribas*et al.,* 2020), and COI Collembola sequences from Cicconardi et al. (2017) from outside the Canary Islands, were used to classify OTUs as 'non-endemic' if similarity with non-Canarian sequences was [?]97%; and 'likely introduced' if the similarity was [?]99%.

To explore OTU and 15% lineage distributions, the number of sampling sites with a presence (number of occurrences), the maximum geographical distance of occurrences, and the different habitats with occurrences were recorded for each OTU and 15% lineage, the latter summarised using Venn diagrams. Habitat specificity was estimated for each entity using the proportion of occurrences in a particular habitat, considering those with 80-100% of occurrences in one habitat as entities with high habitat specificity. Estimations of habitat specificity were performed for those entities sampled in n or more sites, with n = 3, 4, 5, and 6. Finally, we explore the structure of genetic diversity for each OTU and 15% lineage with a product of its number of sites by its number of haplotypes [?] 15. Firstly, we tested the relationship between the genetic distance (F84 model) and geographic distance (Euclidean distance between sampling sites). The relationship between both distances was estimated by randomising spatial distances 1000 times and computing the proportion of times in which the model deviance was smaller than the randomised model deviance, adjusting a linear model using the *glm* function (link = "identity") as in Gómez-Rodríguez & Baselga (2018). Geographic distances were calculated using the R package *gdistance* as before, with calculations performed for each pair of sites with the lowest and highest limit of permitted movements restricted to the highest (plus 100 meters) and lowest (minus 400 meters) values of the two sites. We applied these restrictions to avoid shortest paths transgressing unfavourable habitats over the top of the island, while also allowing paths to cross the valley separating the central region of Tenerife from the Anaga peninsula, and facilitating connectivity over cliffs separating coastal sites. In addition we also tested the correlation between genetic distance (F84 model) among haplotypes and their distribution in the four habitats, using permutational ANOVAs with 999 permutations and the habitat as a grouping factor. To graphically summarise patterns of haplotype relatedness and habitat association, we estimated and plotted haplotype networks for all 15% lineages including four or more haplotypes using the function*mjn* of the R package *pegas* (Paradis, 2010). For 15% lineages with more than 40 haplotypes (four cases), the *mjn*function could not be applied, and networks were alternatively estimated with the *haploNet* function, which uses an infinite site model and uncorrected distances.

Results

Metabarcode data

Overall, 12,621,754 raw reads were obtained, distributed across 104 libraries, of which 1,405,224 passed initial cleaning and denoising steps and were classified as Acari, Collembola, and Coleoptera, resulting in 19,304 ASVs. Of these, 1,813 ASVs (1,278,294 reads) passed metaMATE filtering, applying parameters to maximize the number of surviving va-ASVs while limiting na-ASVs to comprise [?]5% of the final dataset. Parameters used and estimated contributions of a-ASVs and na-ASVs to the filtered dataset are provided in Table 1.

Final filtering to remove records with less than five reads in a library resulted in the retention of 1791 ASVs (i.e. 98.7%) in the final community dataset. Summary data per library is provided in supplementary table S4, and the final set of ASVs and the community table is provided as a supplementary file.

Community richness and structure at multiple thresholds of genetic similarity

Superficial layers tend to have higher richness than their corresponding deep soil layers across all four habitats, with significant richness differences between soil layers found for thermophilous woodland and pine forest (Fig. S1). After combining superficial and deep soil layers for all 52 sites, mean site richness (α diversity) within habitats ranged 55 - 73.5 for haplotypes, 38.5 - 49 for 3% clusters and 34.5 - 43 for 15% clusters (Fig. 2A). Differences in richness by sample among habitats were small and maintained across different clustering thresholds, and pointed to dry scrubland community samples as poorer (lower richness by site) compared to the other habitats (Fig. 2A).

Mean endemicity by site (proportion of lineages that occur exclusively in that site) ranged from 24.0% to 48.8% at the haplotype level, from 13.5% to 22.7% for 3% clusters, and from 6.8% to 15.4% for 15% clusters (Fig. 2B). Comparisons among habitats revealed that endemicity was significantly higher for dry scrubland communities than for laurel forest communities (Fig. 2B). Compositional dissimilarity among communities (β diversity, βsor) was high and was dominated by lineage turnover (βsim), rather than nestedness (βsne), with βsor values ranging 0.87-0.96 across all clustering levels and habitats. Dry scrubland communities showed the highest levels of compositional dissimilarity across the different clustering thresholds (Fig. 2C).

Total observed richness at the island scale (g diversity) by habitat ranged from 534 - 588 haplotypes, 278 - 316 lineages at 3% and 194 - 255 lineages at 15% (Fig. 2C), while extrapolated values (Chao index) nearly doubled observed values (Fig. 2D). Differences in g diversity among habitats were not consistent across different clustering thresholds, with thermophilous woodland showing the lowest number of haplotypes but the highest number of lineages at the 15% clustering threshold (Fig. 2C). Accumulation curves reveal no plateau in the accumulation of entities across samples for any habitat or genetic threshold, with the laurel forest showing the lowest rates of accumulation (Fig. 2D).

Comparisons with biodiversity measures obtained by Arribas *et al.*(2020) in forest and grassland sites in a continental setting revealed that richness by sample (α diversity) was lower in the samples of Tenerife compared with continental soils (Kruskal $p < 0.001$; Fig. S2). Comparisons of β diversity values restricted to a comparable spatial scale of 15 km resulted in significantly lower β diversity values in Tenerife for haplotypes ($p < 0.001$) but not for 3% OTUs (Fig. S2). Finally, g diversity by sampling region, as estimated by the total number of haplotypes and OTUs recorded, was similar for the different habitats of Tenerife (534 – 588 haplotypes and from 278 - 316 3% OTUs) and the six continental settings in Arribas et al. (2020) (558 - 623 haplotypes, and 276 – 319 OTUs) (Fig. S2).

NMDS for the compositional dissimilarity of the communities of Tenerife soils showed habitat as a major driver of the ordination of samples, and accordingly, for all clustering levels, a significant proportion of variance ($0.18 < r2 < 0.28$; $p < 0.001$) was explained by the habitat factor (Fig. 3A). In addition, dry scrubland communities showed the highest dispersal, while the laurel forest communities were the least scattered (Fig. 3A).

Analyses of community similarity (1-pairwise β diversity) with spatial distance within each habitat revealed significant distance decay for all clustering levels in all habitats, except for dry scrubland (Fig. 3B). For laurel forest, pine forest, and thermophilous woodland, slopes of the exponential decay curves were very similar at all threshold levels, and assemblage similarity increased with each level (Fig. 3B). Genetic similarity showed a high and significant log-log correlation with the number of lineages ($0.97 < r2 < 0.99$, $p < 0.001$), initial similarity ($0.92 < r2 < 0.99$; $p < 0.001$), and mean similarity of communities ($0.97 < r2 < 0.99$; $p < 0.001$) (Table S5), as expected if community variation across hierarchical levels of similarity is described by a fractal geometry (Baselga et al., 2013, 2015).

A decrease in community similarity with environmental distance (Fig. S3) was only significant for the laurel

forest and some clustering levels in the pine forest (Table S6). However, variance partitioning showed that variance uniquely explained by environmental distance (i.e. independently of the spatial distance) was lower (3.2% - 9.0% of explained variation at all levels) than the uniquely explained variance by the spatial distance (6.9% - 45.0% of explained variation).

Lineages characterisation and meta-phylogeographical patterns

Across all 52 samples across Tenerife island, a total of 813 OTUs (3% clustering) and 533 15% lineages (15% clustering) were found, with a mean of 2.2 haplotypes by OTU and a mean of 1.5 OTUs by 15% lineage. Table 2 shows the number of OTUs and 15% lineages obtained and extrapolated values (Chao index) for Acari, Collembola, and Coleoptera across the 52 sites. Among OTUs, 488 (60%) included a single haplotype (single-haplotype OTUs), and 325 (40%) were classified as*multi-haplotype OTUs* (Fig. 4). The most diverse OTU included 40 haplotypes and corresponded to a species of Acari from the order Sarcoptiformes, not represented in public sequence repositories. Among the 533 15% lineages, 413 (77%) included a single OTU (*non-diversified lineages*), and 122 (23%) included 2 or more OTUs and were classified as *diversified lineages* . (Fig. 4). The most diverse 15% lineage included 21 OTUs (77 haplotypes), corresponding to the weevil genus *Laparocerus* Schoenherr, 1834, the most diverse beetle genus in Tenerife (Machado, Rodríguez-Expósito, López, & Hernández, 2017). Among the 813 OTUs, 135 (16.6%) were classified as *non-endemic OTUs* because they have a similarity [?]97% with sequences of external (non-Canarian) databases. Of these, 115 OTUs (14.1%) showed a similarity [?]99% and so were additionally categorised as *likely introduced OTUs* (Table S7).

Each OTUs was found on average on 2.9 sampling sites and each 15% lineage on 3.9 sites. Four hundred and five OTUs (49.8%) were detected in a single site, and the remaining 408 (50.2%) in two or more sites (Fig. 4). Two-hundred 15% lineages (37.5%) were detected in a single site, and the remaining 333 (63.5%) were found in two or more sites. The most widespread 15% lineage, including a single OTU, was found in 37 sites and corresponded to a likely introduced species identified as*Ceratophysella gibbosa* (Bagnall, 1940), having similarity >99% with specimens from France and Australia (Table S7). Regarding the distributions of OTUs and 15% lineages across the habitats, habitat specificity was estimated for those entities sampled in n or more sites, with $n = 3$, 4, 5, and 6. The percentage of OTUs considered *OTUs with high habitat specificity* , with at least 80% of occurrences within the same habitat, ranged from 31% to 29% (Fig. 5A). Similarly, *lineages with high habitat specificity* ranged from 30% to 26% (Fig. 5B). Patterns of shared OTUs and 15% lineages among habitats revealed that spatially (and climatically) adjacent habitats presented higher numbers of shared OTUs and lineages (e.g., laurel forest and thermophilous woodland; 102 shared 15% lineages), compared to spatially disconnected habitats (e.g. laurel forest and dry scrubland; 50 shared 15% lineages, of which 45 are also shared with the thermophilous woodland typically located in between) (Fig. 5). Individual Venn diagrams for Acari, Collembola, and Coleoptera were highly consistent with this general pattern (Fig. S4).

Regarding the structure of the genetic diversity within OTUs and 15% lineages, the analyses were restricted to those entities showing a product of the number of sites by the number of haplotypes [?] 15; n = 107 OTUs and 128 15% lineages. The proportion of these entities with a significant geographical structure of genetic diversity constituted 29.0% of the OTUs and 30.5% of the 15% lineages (Fig. 6). The proportion of entities with a significant structure of genetic diversity associated with the factor habitat was lower and represented 8.4% of OTUs and 16.4% of 15% lineages (Fig. 6). The overlap between the entities structured by spatial distance and habitat revealed that 13 of the 21 entities structured by habitat were also structured by spatial distance (Fig. 6).

Discussion

Using the recently developed metaMATE pipeline (Andujar et al., 2021), we have generated a stringently filtered dataset of amplicon sequence variants (ASVs) for mesofaunal soil communities sampled across an oceanic island. By achieving a level of spurious sequences estimated to be no more than 5% of ASVs in the final dataset, we have been able to undertake both phylogeographic and community ecological analyses

at different hierarchical levels of relatedness. These data reveal both ecological patterns and evolutionary processes, providing novel insights into community assembly within soil mesofauna at an unprecedented taxonomic scale. In doing so, we demonstrate wocDNA metabarcoding to be a powerful tool for understanding ecological and evolutionary processes within dark taxa – highly diversified lineages for which described species are estimated to be only a limited proportion of true species richness (Hartop, Srivathsan, Ronquist, & Meier, 2021).

The (unknown) diversity of soil mesofauna within insular soils

The diversity of soil mesofauna within oceanic islands has been poorly explored. Literature on the topic is limited (i.e. Koh *et al.* , 2002; Maraun *et al.* , 2007; Fattorini, 2009; Cicconardi *et al.* , 2017), and even basic species inventory data are in general scarce for this ecologically important biodiversity fraction. Within the Canary Islands, the Biodiversity Databank of the Canary Islands (*https://www.biodiversidadcanarias.es/biota/*; from hereon referred to as BIOTA) is a constantly updated public database containing all species records for the archipelago published in the scientific literature. BIOTA currently reports 287 species of Acari, 88 of Collembola, and 1360 species of Coleoptera from the island of Tenerife. Our results demonstrate that current knowledge of species diversity within the island is greatly underestimated. By sampling only 52 soil communities (approximately a 2.6 x 10^{-8} of total island surface area) across the four dominant habitats of the island of Tenerife, we have recovered nearly 1800 mtDNA haplotypes from Acari, Collembola, and Coleoptera that clustered into 813 putative species (OTUs at 3%), 434 Acari, 129 Collembola and 250 Coleoptera (Table 2). Even with a substantially more conservative dissimilarity threshold of 15%, total lineage number remains above 500. For Acari and Collembola, sampled OTU numbers exceed the number of species recorded until the date for the island (Table 2). In the case of Coleoptera, it should be noted that many of the 1360 recorded species in BIOTA for Tenerife are not associated with soil, while our sampling is strictly focussed on soil lineages, so a direct comparison is difficult. Overall, our results reveal that the soils of Tenerife are much richer in mesofauna than previously reported, and highlight the generally appreciated problems of the Linnaean and Wallacean shortfalls (Cardoso, Erwin, Borges, & New, 2011; Hortal et al., 2015) for soil arthropod biodiversity.

By comparing obtained ASVs against public molecular repositories, we found that 135 (16.7%) of the 813 OTUs matched (97% similarity) non-Canarian records, and can thus be considered as non-endemic species, being either native or introduced (Table S7). However, attributing all remaining OTUs to endemic species is not possible, because of the incomplete nature of public molecular repositories. Comparison to public molecular repositories identifies 34 Acari (8%), 39 Coleoptera (16%), and 49 Collembola (38%) OTUs with high sequence similarity ([?]99%) to individuals from other regions. It is plausible that most, if not all, are recent human-mediated introductions, rather than worldwide distributed species requiring unrealistic passive dispersal kernels to maintain species cohesion. These findings are in agreement with Cicconardi et al. (2017), who concluded from genome skimming data that 88% of the 25 Collembola species they sampled from laurel forests in Tenerife result from human-mediated introductions. Distinguishing between native and introduced origins for soil-adapted species is challenging, and focused studies are needed to elucidate the extent of species introductions within oceanic islands (Andersen et al., 2019).

Alpha, beta, and gamma diversity estimations at the OTU and haplotype levels point to lower diversity values in island soils compared to continental soils (Fig. S2). In contrast, high endemicity by sample and significant community differences among habitats are similar to patterns found in continental soils (see the section below). Using the same field, laboratory, and bioinformatic protocols, Arribas *et al.*(2020) sampled 12 sites within each of three forested and three grassland areas in Iberia. While sample sizes are comparable between both studies, spatial scale differs, with maximum distances between samples of 70 km within Tenerife habitats compared to only 15 km in Arribas *et al.* (2020). Within habitats, both α and β diversity (restricted to a comparable spatial scale of 15 km) were significantly lower in Tenerife, suggesting that insular soil mesofaunal diversity may be lower, compared to continental areas of a similar size, consistent with previous suggestions for aboveground plant and animal communities (Kier et al., 2009; Whittaker & Fernández-Palacios, 2007). While interesting, the generality of this pattern awaits further investigation. With

appropriate measures to harmonise methodology and optimise data comparability, the generality of patterns observed here can feasibly be assessed across independent studies (Arribas, Andújar, Bidartondo, et al., 2021).

Dispersal limitation as a driver of the assemblage and diversification of insular soil mesofauna

Dispersal is a key process shaping island biotas, being fundamental for colonization and consequential within islands for the geographic structuring of genetic variation within species, speciation, and intra-island diversification (Gillespie et al., 2012; Salces-Castellano et al., 2020; Warren et al., 2014). Integrating across the distances and frequencies over which active and passive dispersal processes contribute to species cohesion and speciation (Fig. 1) provides a predictive framework for evolutionary trajectories at the level of individual lineages (Gillespie et al., 2012). Competing models can be proposed for the likely shape of the dispersal kernel for the typically tiny and flightless component of mesofaunal soil species, with differing implications for their spatial patterns of diversity (Fig. 1) (Andújar et al., 2017). The first is a model of limited active but high passive dispersal potential, mediated by the small size of soil mesoarthropods, according to the "everything is everywhere hypothesis" (Fenchel & Finlay, 2004; Finlay, 2002) which predicts large spatial distances for species cohesion. The second model is one of limited active and passive dispersal potential, and thus predicts a limited spatial scale for speciation (Andújar et al., 2017; Arribas, Andújar, Salces-Castellano, et al., 2021).

Analyses of mesofauna from continental soils have led to contrasting inferences for how dispersal shapes their community assembly and diversification. Strong dispersal constraints have rarely been recognised for soil mesofauna, and long-distance dispersal has been considered to characterise soil mesofauna, largely mediated by passive dispersal by air, water or in marine plankton (Decaëns, 2010; Thakur et al., 2019; Wardle, 2002). In contrast, molecular studies of soil mesofaunal lineages and communities frequently reveal dispersal limitation, associated with both diversification and community turnover across limited spatial scales (Andújar et al., 2017; Arribas et al., 2021; Francesco Cicconardi, Fanciulli, & Emerson, 2013; Collins et al., 2019). The BIOTA inventory for the island of Tenerife reveals that 236 of 297 recorded species of Acari (79%) , 62 of 88 Collembola (70%), and 699 of 1360 Coleoptera (51%), are considered to be non-endemic, having populations outside of the Canary Islands. These data are more consistent with a model of high dispersal potential for soil mesofauna. However, our metabarcode data provide greater support for a model where dispersal is limited, where island populations are evolutionarily independent entities, within which futher diversification can occur.

Following island colonisation and establishment, dispersal limitation may favour subsequent intra-island genetic differentiation, the extent of which will be mediated by species traits (e.g. niche, species-specific dispersal ability), and the selective landscape (e.g. spatial variation in biotic and abiotic conditions). Under this model, spatially structured lineages and communities are expected to emerge, and there are clear signatures for this within our data. Within each of the studied habitats, for haplotype, species, and supraespecific levels of variation, community similarity is a function of geographic distance (Fig. 2C and 3). This self-similarity of distance decay at haplotype and species level (Fig. 3B) is consistent with a role for dispersal limitation driving community assembly (Baselga et al., 2015; Gómez-Rodríguez & Baselga, 2018). The influence of dispersal constraints within the soil matrix appears to act at short spatial distances, and the evident high turnover with physical distance suggests that our sampled communities within each habitat are not from a single panmictic metacommunity. At the lineage level, our results reveal multiple signals of dispersal limitation constraining diversification. Many of the soil mesofaunal OTUs recovered from our wocDNA metabarcode data are not recorded (at least molecularly) outside the island (Fig. 4; Table S7), have restricted distributions within the island, and present spatially structured genetic variation (Fig. 6). Additionally, among the 533 15% lineages recovered, 122 comprises two or more OTUs. If we assume each 15% lineage represents a single colonization event into Tenerife, 49.2% of all OTUs may be derived from intra-island divergence events. Thirty-nine OTUs show a significant correlation between genetic and spatial distances, 34 of these comprising two or more OTUs, further supporting *in situ* spatial structuring and diversification within lineages (Fig. 4 and 6).

Habitat and the diversity and structure of insular soil mesofauna

Across all communities, the greatest differentiation was among communities from the different habitats (Fig. 3A), and we find up to 30% of OTUs and lineages that are consistent with high habitat specialisation (Fig. 5). Habitat specificity in soil mesofauna has been previously reported, with strong evidence for specialisation between open versus forested vegetation types (Arribas *et al.* , 2020, Caruso et al. 2012) and different forest types (Noguerales et al., 2021). Our results extend the generality of these patterns to the soils of an oceanic island.

Islands have been suggested to favour generalist species, either by colonization and persistence filters that select for species with wide niche breadth (ecological tolerance) (Gaston, 2003; Reaka, 1980) or through lower levels of competition favouring ecological release following colonisation (Olesen et al., 2002). However, our results are not consistent with these proposals, revealing that for much of the soil mesofauna, habitat features could be driving a scenario of species sorting (Leibold et al., 2004), with the existence of largely separate (still overlapping) metacommunities inhabiting the different habitats within the island. Two contrasting but not mutually exclusive models can be evoked to explain these patterns of habitat specificity. The first involves niche conservatism, with colonising species establishing into habitats to which they are preadapted and with intraisland diversification primarily constrained within the same habitats (Lim et al., 2021; Salces-Castellano et al., 2020). The second involves niche lability, and it has been described as an essential process generating diversity within oceanic island biotas through selection gradients across different habitats (Gillespie et al., 2001). Our results reveal that among the 533 lineages that are assumed to be independent colonisations to Tenerife, 312 are restricted to a single habitat. Furthermore, among the 128 15% lineages where genetic differentiation associated with habitat type was tested for, 21 presented a significant association (Fig. 6). Thus, our data provides only limited evidence for habitat shifts promoting diversification, suggesting an important role for climatic niche conservatism driving ecological assembly of soil mesofauna within the island.

Despite contrasting biotic and abiotic features among the sampled habitats (del Arco Aguilar, González-González, Garzón-Machado, & Pizarro-Hernández, 2010), g and mean α diversities were similar within each, albeit with some differences between dry scrublands and the remaining three habitats (Fig. 2). Dry scrubland soils have significantly lower species richness by sample, whereas lineage accumulation across multiple sites resulted in similar values of g diversity. This pattern is mediated by significantly higher local endemicity within dry scrubland soils, and thus higher turnover not spatially structured (Fig. 3). Habitat specific differences related current and past habitat patchiness and connectivity could be driving such differences. Under the habitat stability hypothesis (Ribera & Vogler, 2000; Southwood, 1977), lineages with high dispersal potential are expected to be primarily selected within more ephemeral habitats. Within this framework, it can be hypothesised that a lower habitat stability for scrublands, due to higher exposure to sea-level changes in geological times, could be contributing to their observed lower local richness and more limited spatial structuring of their soil mesofaunal communities. However, the number of spatially structured OTUs and 15% lineages was very similar among habitats, and habitat specificity at different hierarchical levels of analysis was comparable among them (data not shown). Other factors, such as fine-scale habitat heterogeneity, may have eroded the signature of geography into the assembly of mesofaunal communities within the dry scrublands. Further studies are needed to explore the contrasting ecological and evolutionary processes that drive the community assembly within different habitat types, such as those described here.

Conclusions

Our results reveal an important role for niche conservatism as a driver of insular soil mesofaunal community assembly, with limited evidence for habitat shifts promoting diversification. These results also support a fundamental role of habitat features in the assembly of soil mesofauna, in agreement with previous studies (Arribas et al., 2021; Noguerales et al., 2021), with much habitat specialism being explained as the result of independent colonisation and establishment of preadapted species. Hierarchical patterns of distance decay at the community level and metaphylogeographical analyses reveal geographic structuring over limited spatial scales from the level of haplotypes through to species and lineages, as expected for taxa under strong dispersal

limitations. We also reveal broadly consistent patterns of local-scale species richness across different insular habitats and find that local insular richness is lower than in broadly comparable continental settings. These results demonstrate the potential for wocDNA metabarcoding to advance our understanding of biodiversity, particularly for the so called dark taxa – important fractions of biodiversity that have traditionally been difficult to work with.

Acknowledgements

This work was supported by project CGL2015-74178-JIN (AEI, Spain/FEDER, EU) awarded to CA, and CGL2017-85718-P and PID2020-116788GB-I00 (AEI, Spain/FEDER, EU) awarded to BCE. CA was additionally supported by Fundación Caja Canarias/Obra Social "La Caixa" (2017RCE03). BCE, together with PA was additionally supported by the H2020 iBioGen project, funded by the European Research Council, Award Number: 810729. We extend our gratitude to the regional governments of Andalucía, Canarias, and and the local council (Cabildo) of Tenerife for facilitating collecting of samples, and to Jesús Arribas for assistance with field sampling and drawings of habitats.

Author contributions

CA, PA and BCE conceived the study, that was led by CA; CA and PA designed the methodology; CA, PA, HL, AP-D, PO, APV, and BCE provided the data; CA, PA, and YA analysed the data. CA, PA, and BCE interpreted results and wrote the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Orcid

Carmelo Andújar*https://orcid.org/0000-0001-9759-7402*

Paula Arribas*https://orcid.org/0000-0002-0358-8271*

Alfried P. Vogler*https://orcid.org/0000-0002-2462-3718*

Brent C. Emerson*https://orcid.org/0000-0003-4067-9858*

Antonio J. Pérez-Delgado*https://orcid.org/0000-0002-3797-4058*

Heriberto López*https://orcid.org/0000-0001-6988-5204*

Yurena Arjona*https://orcid.org/0000-0002-1851-1664*

References

Ancochea, E., Maria, J., Ibarrola, E., Cendrero, A., & Coello, J. (1990). Volcanic evolution of the island of Tenerife (Canary Islands) in the light of new K-Ar data CANARY, *44* (1771).

Andersen, J. C., Oboyski, P., Davies, N., Charlat, S., Ewing, C., Meyer, C., … Roderick, G. K. (2019). Categorization of species as native or nonnative using DNA sequence signatures without a complete reference library. *Ecological Applications* , *29* (5), 1–11. doi:10.1002/eap.1914

Andújar, C., Creedy, T. J., Arribas, P., López, H., Salces-Castellano, A., Pérez-Delgado, A. J., … Emerson, B. C. (2021). Validated removal of nuclear pseudogenes and sequencing artefacts from mitochondrial metabarcode data. *Molecular Ecology Resources* ,*21* (6), 1772–1787. doi:10.1111/1755-0998.13337

Andújar, C., Pérez-González, S., Arribas, P., Zaballos, J. P., Vogler, A. P., & Ribera, I. (2017). Speciation below ground: Tempo and mode of diversification in a radiation of endogean ground beetles.*Molecular Ecology* , *26* (21), 6053–6070. doi:10.1111/mec.14358

Arribas, P., Andújar, C., Bidartondo, M. I., Bohmann, K., Coissac, É., Creer, S., ... Emerson, B. C. (2021). Connecting high-throughput biodiversity inventories: Opportunities for a site-based genomic framework for global integration and synthesis. *Molecular Ecology* . doi:10.1111/mec.15797

Arribas, P., Andújar, C., Hopkins, K., Shepherd, M., & Vogler, A. P. (2016). Metabarcoding and mitochondrial metagenomics of endogean arthropods to unveil the mesofauna of the soil. *Methods in Ecology and Evolution* , *7* (9), 1071–1081. doi:10.1111/2041-210X.12557

Arribas, P., Andújar, C., Salces-Castellano, A., Emerson, B. C., & Vogler, A. P. (2021). The limited spatial scale of dispersal in soil arthropods revealed with whole-community haplotype-level metabarcoding.*Molecular Ecology* , *30* (1), 48–61. doi:10.1111/mec.15591

Babraham Institute. (2013). FastQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc. Retrieved from www.bioinformatics.babraham.ac.uk/projects/fastqc

Bardgett, R. D., & van der Putten, W. H. (2014). Belowground biodiversity and ecosystem functioning. *Nature* , *515* (7528), 505–511. doi:10.1038/nature13855

Baselga, A., Fujisawa, T., Crampton-Platt, A., Bergsten, J., Foster, P. G., Monaghan, M. T., & Vogler, A. P. (2013). Whole-community DNA barcoding reveals a spatio-temporal continuum of biodiversity at species and genetic levels. *Nature Communications* , *4* (May), 1892, DOI: 10.1038/ncomms2881. doi:10.1038/ncomms2881

Baselga, A., Gómez-Rodríguez, C., & Vogler, A. P. (2015). Multi-hierarchical macroecology at species and genetic levels to discern neutral and non-neutral processes. *Global Ecology and Biogeography* , *24* (8), 873–882. doi:10.1111/geb.12322

Baselga, A., & Orme, C. D. L. (2012). betapart: an R package for the study of beta diversity. *Methods in Ecology and Evolution* ,*3* (5), 808–812. doi:10.1111/j.2041-210X.2012.00224.x

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* ,*30* (15), 2114–2120. doi:10.1093/bioinformatics/btu170

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* , *13* (7), 581–583. doi:10.1038/nmeth.3869

Cameron, E. K., Martins, I. S., Lavelle, P., Mathieu, J., Tedersoo, L., Gottschall, F., ... Eisenhauer, N. (2018). Global gaps in soil biodiversity data. *Nature Ecology and Evolution* , *2* (7), 1042–1043. doi:10.1038/s41559-018-0573-8

Cardoso, P., Erwin, T. L., Borges, P. A. V., & New, T. R. (2011). The seven impediments in invertebrate conservation and how to overcome them.*Biological Conservation* , *144* (11), 2647–2655. doi:10.1016/j.biocon.2011.07.024

Carracedo, J. C., Guillou, H., Paterne, M., Scaillet, S., Rodríguez Badiola, E., Paris, R., ... Hansen Machín, A. (2004). Analisis del riesgo volcanico asociado al flujo de lavas en Tenerife (Islas Canarias): Escenarios previsibles para una futura erupcion en la isla.*Estudios Geologicos* .

Caruso, T., Taormina, M., & Migliorini, M. (2012). Relative role of deterministic and stochastic determinants of soil animal community: a spatially explicit analysis of oribatid mites. *The Journal of Animal Ecology* , *81* (1), 214–21. doi:10.1111/j.1365-2656.2011.01886.x

Cicconardi, F, Nardi, F., Emerson, B. C., Frati, F., & Fanciulli, P. P. (2010). Deep phylogeographic divisions and long-term persistence of forest invertebrates (Hexapoda: Collembola) in the North-Western Mediterranean basin. *Molecular Ecology* , *19* (2), 386–400. doi:10.1111/j.1365-294X.2009.04457.x

Cicconardi, Francesco, Borges, P. A. V., Strasberg, D., Oromí, P., López, H., Pérez-Delgado, A. J., ... Emerson, B. C. (2017). MtDNA metagenomics reveals large-scale invasion of belowground arthropod communities by introduced species. *Molecular Ecology* ,*26* (12), 3104–3115. doi:10.1111/mec.14037

Cicconardi, Francesco, Fanciulli, P. P., & Emerson, B. C. (2013). Collembola, the biological species concept and the underestimation of global species richness. *Molecular Ecology* , *22* (21), 5382–5396. doi:10.1111/mec.12472

Collins, G. E., Hogg, I. D., Convey, P., Barnes, A. D., & McDonald, I. R. (2019). Spatial and temporal scales matter when assessing the species and genetic diversity of springtails (Collembola) in Antarctica. *Frontiers in Ecology and Evolution* , *7* (MAR), 1–18. doi:10.3389/fevo.2019.00076

Decaëns, T. (2010). Macroecological patterns in soil communities. *Global Ecology and Biogeography* , *19* (3), 287–302. doi:10.1111/j.1466-8238.2009.00517.x

del Arco Aguilar, M. J., González-González, R., Garzón-Machado, V., & Pizarro-Hernández, B. (2010). Actual and potential natural vegetation on the Canary Islands and its conservation status. *Biodiversity and Conservation* , *19* (11), 3089–3140. doi:10.1007/s10531-010-9881-2

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods* , *10* (10), 996–8. doi:10.1038/nmeth.2604

Emerson, B. C., & Gillespie, R. G. (2008). Phylogenetic analysis of community assembly and structure over space and time. *Trends in Ecology & Evolution* , *23* (11), 619–30. doi:10.1016/j.tree.2008.07.005

Ettema, C. H., & Wardle, D. a. (2002). Spatial soil ecology. *Trends in Ecology and Evolution* , *17* (4), 177–183. doi:10.1016/S0169-5347(02)02496-5

FAO, ITPS, GSBI, SCBD, & EC. (2020). *State of knowledge of soil biodiversity – Status, challenges and potentialities. Report 2020* . Roma: FAO. doi:https://doi.org/10.4060/cb1928en

Fattorini, S. (2009). On the general dynamic model of oceanic island biogeography. *Journal of Biogeography* , *36* (6), 1100–1110. doi:10.1111/j.1365-2699.2009.02083.x

Fenchel, T., & Finlay, B. J. (2004). The ubiquity of small species: Patterns of local and global diversity. *BioScience* , *54* (8), 777–784. doi:10.1641/0006-3568(2004)054[0777:TUOSSP]2.0.CO;2

Finlay, B. J. (2002). Global dispersal of free-living microbial eukaryote species. *Science* , *296* (5570), 1061–1063. doi:10.1126/science.1070710

Gaston, K. J. (2003). *The structure and dynamics of geographic ranges* . Oxford, UK: Oxford University Press.

Gillespie, R. G., Baldwin, B. G., Waters, J. M., Fraser, C. I., Nikula, R., & Roderick, G. K. (2012). Long-distance dispersal: A framework for hypothesis testing. *Trends in Ecology and Evolution* ,*27* (1), 47–55. doi:10.1016/j.tree.2011.08.009

Gillespie, R. G., & Roderick, G. K. (2002). Arthropods on islands: Colonization, speciation, and conservation. *Annual Review of Entomology* , *47* (February), 595–632. doi:10.1146/annurev.ento.47.091201.145244

Gillespie, R. G., Roderick, G. K., & Howarth, F. G. (2001). Adaptive Radiation. *Encyclopedia of Biodiversity: Second Edition* ,*1* (ii), 21–36. doi:10.1016/B978-0-12-384719-5.00002-2

Gómez-Rodríguez, C., & Baselga, A. (2018). Variation among European beetle taxa in patterns of distance decay of similarity suggests a major role of dispersal processes. *Ecography* , *41* (11), 1825–1834. doi:10.1111/ecog.03693

Goslee, S. C., & Urban, D. L. (2007). The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software* , *22* (7), 1–19. doi:10.18637/jss.v022.i07

Hartop, E., Srivathsan, A., Ronquist, F., & Meier, R. (2021). Large-scale Integrative Taxonomy (LIT): resolving the data conundrum for dark taxa.*Https://Www.Biorxiv.Org/Content/10.1101/2021.04.13.439467V2* , 1–34.

Hortal, J., De Bello, F., Diniz-Filho, J. A. F., Lewinsohn, T. M., Lobo, J. M., & Ladle, R. J. (2015). Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. *Annual Review of Ecology, Evolution, and Systematics* , *46* (December), 523–549. doi:10.1146/annurev-ecolsys-112414-054400

Hubbell, S. P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography* . *The Unified Neutral Theory of Biodiversity and Biogeography* . Princeton: Princeton University Press. doi:10.1515/9781400837526

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research* , *17* (3), 377–386. doi:10.1101/gr.5969107

Kier, G., Kreft, H., Tien, M. L., Jetz, W., Ibisch, P. L., Nowicki, C., ... Barthlott, W. (2009). A global assessment of endemism and species richness across island and mainland regions. *Proceedings of the National Academy of Sciences of the United States of America* ,*106* (23), 9322–9327. doi:10.1073/pnas.0810306106

Kisel, Y., & Barraclough, T. G. (2010). Speciation has a spatial scale that depends on levels of gene flow. *The American Naturalist* ,*175* (3), 316–34. doi:10.1086/650369

Koh, L. P., Sodhi, N. S., Tan, H. T. W., & Peh, K. S. H. (2002). Factors affecting the distribution of vascular plants, springtails, butterflies and birds on small tropical islands. *Journal of Biogeography* , *29* (1), 93–108. doi:10.1046/j.1365-2699.2002.00657.x

Leibold, M. A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J. M., Hoopes, M. F., ... Gonzalez, A. (2004). The metacommunity concept: A framework for multi-scale community ecology. *Ecology Letters* ,*7* (7), 601–613. doi:10.1111/j.1461-0248.2004.00608.x

Lim, J. Y., Patino, J., Noriyuki, S., Simmari, L. C., Gillespie, R. G., & Krehenwinkel, H. (2021). Climatic niche conservatism shapes the ecological assembly of Hawaiian arthropod communities. *BioRxiv* , 26.

Mac Nally, R., & Walsh, C. J. (2004). Hierarchical partitioning public-domain software. *Biodiversity and Conservation* ,*13* (3), 659–660. doi:10.1023/B:BIOC.0000009515.11717.0b

MacArthur, R. H., & Wilson, E. O. (1963). An Equilibrium Theory of Insular Zoogeography. *Evolution* , *17* (4), 373–387.

MacArthur, R. H., & Wilson, E. O. (1967). *The Theory of Island Biogeography.* (P. U. Press, Ed.). Princeton.

Machado, A., Rodríguez-Expósito, E., López, M., & Hernández, M. (2017). Phylogenetic analysis of the genus laparocerus, with comments on colonisation and diversification in macaronesia (Coleoptera, Curculionidae, Entiminae). *ZooKeys* , *2017* (651), 1–77. doi:10.3897/zookeys.651.10097

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2021). cluster: Cluster Analysis Basics and Extensions. R package version 2.1.2. Retrieved from https://cran.r-project.org/package=cluster

Maraun, M., Schatz, H., & Scheu, S. (2007). Awesome or ordinary? Global diversity patterns of oribatid mites. *Ecography* , *30* (2), 209–216. doi:10.1111/j.2007.0906-7590.04994.x

Morek, W., Surmacz, B., López-López, A., & Michalczyk, Ł. (2021). "Everything is not everywhere": Time-calibrated phylogeography of the genus Milnesium (Tardigrada). *Molecular Ecology* , *30* (14), 3590–3609. doi:10.1111/mec.15951

Nathan, R. (2005). Long-distance dispersal research: Building a network of yellow brick roads. *Diversity and Distributions* , *11* (2), 125–130. doi:10.1111/j.1366-9516.2005.00159.x

Noguerales, V., Meramveliotakis, E., Castro-Insua, A., Andújar, C., Arribas, P., Creedy, T. J., ... Papadopoulou, A. (2021). Community metabarcoding reveals the relative role of environmental filtering and spatial

processes in metacommunity dynamics of soil microarthropods across a mosaic of montane forests. *Molecular Ecology* , (2), 1–19. doi:10.1111/mec.16275

Oksanen, J., Blanchet, F. Guillaume Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Gavin, L., ... Wagner. (2016). vegan: Community Ecology Package. R package, (December 2016), 0–291. Retrieved from https://cran.r-project.org/package=vegan

Olesen, J. M., Eskildsen, L. I., & Venkatasamy, S. (2002). Invasion of pollination networks on oceanic islands: Importance of invader complexes and endemic super generalists. *Diversity and Distributions* ,*8* (3), 181–192. doi:10.1046/j.1472-4642.2002.00148.x

Paradis, E. (2010). Pegas: An R package for population genetics with an integrated-modular approach. *Bioinformatics* , *26* (3), 419–420. doi:10.1093/bioinformatics/btp696

Patiño, J., Whittaker, R. J., Borges, P. A. V., Fernández-Palacios, J. M., Ah-Peng, C., Araújo, M. B., ... Emerson, B. C. (2017). A roadmap for island biology: 50 fundamental questions after 50 years of The Theory of Island Biogeography. *Journal of Biogeography* ,*44* (5), 963–983. doi:10.1111/jbi.12986

Pohlert, T. (2014). The Pairwise Multiple Comparison of Mean Ranks Package (PMCMR). R package. Retrieved from https://cran.r-project.org/package=PMCMR

Reaka, M. L. (1980). Geographic range, life history patterns, and body size in a guild of coral-dwelling Mantis shrimps. *Evolution* ,*34* (5), 1019. doi:10.2307/2408010

Ribera, I., & Vogler, A. P. (2000). Habitat type as a determinant of species range sizes: The example of lotic-lentic differences in aquatic Coleoptera. *Biological Journal of the Linnean Society* ,*71* (1), 33–52. doi:10.1006/bijl.1999.0412

Rosindell, J., Hubbell, S. P., & Etienne, R. S. (2011). The Unified Neutral Theory of Biodiversity and Biogeography at Age Ten. *Trends in Ecology and Evolution* , *26* (7), 340–348. doi:10.1016/j.tree.2011.03.024

Salces-Castellano, A., Patiño, J., Alvarez, N., Andújar, C., Arribas, P., Braojos-Ruiz, J. J., ... Emerson, B. C. (2020). Climate drives community-wide divergence within species over a limited spatial scale: evidence from an oceanic island. *Ecology Letters* . doi:10.1111/ele.13433

Shade, A., Dunn, R. R., Blowes, S. A., Keil, P., Bohannan, B. J. M., Herrmann, M., ... Chase, J. (2018). Macroecology to Unite All Life, Large and Small. *Trends in Ecology and Evolution* , *33* (10), 731–744. doi:10.1016/j.tree.2018.08.005

Southwood, T. (1977). Habitat, the templet for ecological strategies?*Journal of Animal Ecology* , *46* (2), 336–365.

Thakur, M. P., Phillips, H. R. P., Brose, U., De Vries, F. T., Lavelle, P., Loreau, M., ... Cameron, E. K. (2019). Towards an integrative understanding of soil biodiversity. *Biological Reviews* ,*31* , brv.12567. doi:10.1111/brv.12567

Tobler, W. (1993). Three Presentations on Geographical Analysis and Modelling. *National Center for Geographic Information and Analysis Technical Report* , *93* (1).

Turon, X., Antich, A., Palacín, C., Præbel, K., & Wangensteen, O. S. (2019). From metabarcoding to metaphylogeography: separating the wheat from the chaff. *Ecological Applications* , *30* (2), e02036. doi:10.1101/629535

van Etten, J. (2017). R package gdistance: Distances and routes on geographical grids. *Journal of Statistical Software* ,*76* (1). doi:10.18637/jss.v076.i13

Wardle, D. (2002). *Communities and Ecosystems: Linking the aboveground and belowground components* . Princeton: Princeton University Press.

Warren, B. H., Simberloff, D., Ricklefs, R. E., Aguilée, R., Condamine, F. L., Gravel, D., ... Thébaud, C. (2014). Islands as model systems in ecology and evolution: prospects fifty years after MacArthur-Wilson.*Ecology Letters* , n/a-n/a. doi:10.1111/ele.12398

White, H., Leon-Sanchez, L., Burton, V., Cameron, E., Caruso, T., Cunha, L., ... Caplat, P. (2020). Methods and approaches to advance soil macroecology. *Global Ecology and Biogeography (in Press)* , (June), 1–17. doi:10.1111/geb.13156

Whittaker, R. J., & Fernández-Palacios, J. M. (2007). *Island biogeography* . Oxford, UK: Oxford University Press.

Data Accessibility and Benefit-Sharing

The data that support the findings of this study are openly available in Dryad at*https://doi.org/10.5061/dryad.mw6m905z3*

Tables

Table 1. Filtering parameters selected with MetaMATE and summary results on the number of va-ASVs, vna-ASVs, a-ASVs, and na-ASVs in the dataset before and after filtering. Filtering parameters were selected for a final contribution of na-ASVs below 5%. Values for a-ASVs and na-ASVs are estimated within MetaMATE. Removed va-ASVs were added to the final dataset; surviving vna-ASVs were excluded from the final dataset.

		Before MetaMATE	Before MetaMATE	Before MetaMATE	Before MetaMATE	After MetaMATE	After MetaMATE	After MetaMATE	After MetaMATE
	Filtering criteria (always by library)	va-ASV	vna-ASV	a-ASV	na-ASV	va-ASV	vna-ASV	a-ASV	na-ASV
Acari	Minimum N reads = 40 and Minimum percentage abundace = 0.7%	32	514	1594.0	9508.0	16	2	797.0	37.0
Collembola	Minimum N reads = 50 andMinimum percentage abundace by 20% lineage = 4%	67	235	512.5	2887.5	35	1	267.7	12.3

		Before MetaMATE	Before MetaMATE	Before MetaMATE	Before MetaMATE	After MetaMATE	After MetaMATE	After MetaMATE	After MetaMATE
Coleoptera	Minimum N reads = 8 and Minimum percentage abundace by 20% lineage = 3%	77	763	734.9	4067.1	65	5	620.3	26.7

Table 2. Number of species of Acari, Collembola, and Coleoptera recorded on the Biodiversity Databank of the Canary Islands (*https://www.biodiversidadcanarias.es/biota/*; BIOTA) and number of OTUs (clusters 3%) and lineages (cluster 15%) observed and extrapolated (Chao index) across the 52 sampling sites.

	BIOTA Canarias	BIOTA Tenerife	OTUs (3%) Observed/Chao	Lineages (15%) Observed/Chao
Acari	469	287	434/733	276/382
Collembola	138	88	129/172	105/129
Coleoptera	2234	1360	250/503	152/225

Figure legends

Figure 1. Hypothetical dispersal kernels for soil mesofaunal lineages with different passive dispersal potential (**A**) and map of Tenerife with the distribution of sampled sites (left) and zonal distribution of habitats on the island (right) (**B**). Within (**A**), the high passive dispersal of species in (left) allows species cohesion over larger geographic distances than in (right), with a lower passive dispersal. Modified from Andújar *et al.* (2017).

Figure 2. Richness of soil mesofaunal lineages by sample (alpha diversity, **A**), mean endemicity by sample (**B**), total accumulated richness (local-scale richness or gamma diversity,**C**), mean β diversity among samples (**C**), and richness accumulation curves (**D**) for haplotypes (left), 3% OTUs (middle), and 15% lineages (right) by habitat (laurel forest, pine forest, dry scrublands, and thermophilous woodlands). The significance of Kruskal-Wallis rank sum test (post-hoc comparisons using Bonferroni correction) is indicated for panels (**A**) and (**B**).

Figure 3. Non-parametric multidimensional scaling (NMDS) ordinations (**A**) and distance decay of genetic similarity (**B**) for soil mesofaunal samples. NMDSs represent the variation in community composition (Simpson index, βsim) for haplotype, 3% similarity OTUs, and 15% similarity lineages. Explained variation (r2) and significance (p) of habitat as a grouping factor from the permutational ANOVAs over the community dissimilarity matrixes are shown. Distance decay is plotted at multiple levels of genetic similarity (from haplotypes, black, to 15% genetic similarity, pale grey) within the four habitats (laurel forest, La; pine forest, Pi; dry scrubland, Ds; and thermophilous woodland, Tw).

Figure 4. Histograms showing the distribution of OTUs (**A**) and 15% lineages (**B**) according to the number of sites (left) and the number of haplotypes (middle). Graphs on the right indicate the number of OTUs and 15% lineages found in either one or multiple sites, and the number of OTUs and 15% lineages with one or multiple haplotypes. Also indicated for (**A**) is the number of OTUs with a similarity match

[?] 97% and [?] 99 with reference sequences from outside the Canary Islands, and for (B) the number of lineages with one or multiple OTUs.

Figure 5. Venn diagrams showing distribution of OTUs (**A**) and 15% lineages (**B**) among habitats (laurel forest, La; pine forest, Pi; dry scrubland, Ds; and thermophilous woodland, Tw). Venn diagrams illustrate patterns of exclusive and shared OTUs and 15% lineages. In parenthesis the following are respectively indicated: the total number of OTUs and 15% lineages ("L"), and the number of those collected in a single site ("S"). Barplots on the right represent the proportion and number of OTUs and 15% lineages sampled in more than two, three, four, and five sites that are considered to have high habitat specificity ([?] 80% of sampled sites from a single habitat).

Figure 6. Graph showing the number of 15% lineages with genetic diversity significantly structured with increasing spatial distance (orange) and with genetic diversity significantly structured by habitat (purple) (**A**), and examples of lineages with and without significant habitat and spatial structure (**B** , **C** , and**D**). Estimations in A considered only lineages for which the product of the number of sites by the number of haplotypes is [?] 15. Graphs in B, C, and D represent the correlation between corrected geographical distance (x axis) and genetic similarity (y axis) (left) and haplotype networks (right). Circle size represents the number of sites where haplotypes are found, and colour represents the habitat (laurel forest, yellow; pine forest, red; dry scrubland, blue; and thermophilous woodland, green).