MOLECULAR ECOLOGY

Molecular Ecology (2015) **24**, 3603–3617 doi: 10.1111/mec.13195

Phylogenetic community ecology of soil biodiversity using mitochondrial metagenomics

CARMELO ANDÚJAR,*† PAULA ARRIBAS,*† FILIP RUZICKA,*‡ ALEX CRAMPTON-PLATT,*‡ MARTIJN J.T.N. TIMMERMANS*†[1] and ALFRIED P. VOGLER*†

Department of Life Sciences, Natural History Museum, Cromwell Road, London SW7 5BD, UK, †Department of Life Sciences, Imperial College London, Silwood Park Campus, Ascot SL5 7PY, UK, ‡Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT, UK

Abstract

High-throughput DNA methods hold great promise for the study of taxonomically intractable mesofauna of the soil. Here, we assess species diversity and community structure in a phylogenetic framework, by sequencing total DNA from bulk specimen samples and assembly of mitochondrial genomes. The combination of mitochondrial metagenomics and DNA barcode sequencing of 1494 specimens in 69 soil samples from three geographic regions in southern Iberia revealed >300 species of soil Coleoptera (beetles) from a broad spectrum of phylogenetic lineages. A set of 214 mitochondrial sequences longer than 3000 bp was generated and used to estimate a well-supported phylogenetic tree of the order Coleoptera. Shorter sequences, including *cox1* barcodes, were placed on this mitogenomic tree. Raw Illumina reads were mapped against all available sequences to test for species present in local samples. This approach simultaneously established the species richness, phylogenetic composition and community turnover at species and phylogenetic levels. We find a strong signature of vertical structuring in soil fauna that shows high local community differentiation between deep soil and superficial horizons at phylogenetic levels. Within the two vertical layers, turnover among regions was primarily at the tip (species) level and was stronger in the deep soil than leaf litter communities, pointing to layer-mediated drivers determining species diversification, spatial structure and evolutionary assembly of soil communities. This integrated phylogenetic framework opens the application of phylogenetic community ecology to the mesofauna of the soil, among the most diverse and least well-understood ecosystems, and will propel both theoretical and applied soil science.

Keywords: beta diversity, Coleoptera, Genome skimming, NGS, phylogenetic community structure, soil biodiversity

Received 16 January 2015; revision received 3 April 2015; accepted 8 April 2015

Introduction

Knowledge about the magnitude, distribution and assembly of biodiversity is essential to the understanding of ecosystem processes and environmental change (Gaston 2000). However, to date only a fraction of the existing species diversity on Earth has been catalogued, and most of the described species remain poorly studied (Wilson 2002). The greatest knowledge gaps concern so-called biotic frontiers (André *et al.* 1994), that is highly species-rich habitats that are remote, inaccessible or simply too diverse to be studied with conventional tools of taxonomy (André *et al.* 1994). Possibly 25% of all multicellular species on Earth reside in the soil where they form diverse biological communities (Curtis *et al.* 2002; Nielsen *et al.* 2011). A large proportion of taxonomic and functional diversity in the soil is made

Correspondence: Carmelo Andújar Fernández, Fax: 0207 942 5229; E-mail: c.andujar@nhm.ac.uk

[1] Present address: Department of Natural Sciences, Middlesex University, Hendon Campus, London NW4 4BT, UK

up by species-rich groups of invertebrates. Their trophic networks affect properties of the soil and leaf litter through complex feedback relationships between the activity of the detritivore (decomposer) community, their predators and the physicochemical environment (Ponge 2013).

The important role of biodiversity in soil ecosystems is broadly recognized but their true diversity remains poorly known (Decaëns 2010). While DNA sequencing methods are increasingly applied to study the microbial diversity of the soil (e.g. Bates *et al.* 2013; Ranjard *et al.* 2013), this is not the case for small invertebrates that make up a large proportion of the soil mesofauna. The study of species richness of arthropods in soil communities has been difficult due to a combination of minute body size, poor taxonomic background knowledge, high abundance, the difficulties of linking life stages, and high levels of cryptic diversity (Bardgett 2002; Cicconardi *et al.* 2010; Decaëns 2010). In addition, little is known about the evolutionary origins and community structure of soil-inhabiting organisms. This general lack of taxonomic and evolutionary knowledge has hampered the study of soil biodiversity and its effects on ecosystem function (Wardle 2002; Heemsbergen *et al.* 2004; Decaëns 2010; Nielsen *et al.* 2011).

This study evaluates the role of geographic turnover and soil layer on the taxonomic and phylogenetic composition of soil arthropod communities. Spatial trends in soil biodiversity depend on the degree of dispersal among local sites, as well as the vertical distribution and the response to ecological heterogeneity. Many soil arthropods are secondarily flightless and live concealed deep in the soil layer, limiting their movement, while their small size may facilitate passive dispersion. It is unclear to what degree dispersal capacity constrains the distribution of soil mesofaunal species at regional and global scales (Decaëns 2010), which leaves great uncertainties about the magnitude of biodiversity in soils and its geographic turnover (Fierer *et al.* 2009; Decaëns 2010; Wu *et al.* 2011). As another potential factor driving soil community assembly, vertical stratification has been proposed as a determinant of species diversification in mesofaunal assemblages, but its broader relevance for structuring soil biota is still debated, and the existence of true endogeic components in some mesofaunal groups remains unclear (Ducarme *et al.* 2004). Finally, the lack of a phylogenetic framework for many soil arthropods ignores the evolutionary context of such diversity patterns and precludes phylogenetic community analyses that integrate processes of local species assembly and the evolutionary history of cooccurring lineages (Graham & Fine 2008).

High-throughput sequencing arguably will overcome the 'taxonomic impediment' to the study of inaccessible arthropod biodiversity (Emerson *et al.* 2011; Yu *et al.* 2012). Current methods of DNA barcoding (Hebert *et al.* 2003) and 'metabarcoding' (e.g. Yu *et al.* 2012) for the sequencing of communities mainly target short PCR amplicons for OTU recognition, but have little power for phylogenetic analysis. Instead, phylogenetic informative markers may be obtained by shotgun sequencing of total DNA to extract the high-copy fraction of genomes through 'genome skimming' (Straub *et al.* 2012; Malé *et al.* 2014). Shotgun metagenomic sequencing of bulk community samples yields numerous reads corresponding to mitochondrial DNA, which can be assembled into full or partial mitogenomes (Dettai *et al.* 2012; Zhou *et al.* 2013; Gillett *et al.* 2014; Tang *et al.* 2014; Crampton-Platt *et al.* 2015). 'Mitochondrial metagenomics' to date has been applied mainly for phylogenetics (Gillett *et al.* 2014) but also offers great potential for the study of community assembly, particularly for inaccessible hyperdiverse groups such as soil mesofauna, by allowing the mapping of reads against reference mitochondrial sequences in a similar approach to that used for the study of microbial communities (e.g. Martin *et al.* 2012; Riesenfeld & Pollard 2013). As we show here, mitochondrial metagenomics contributes well-supported community-level evolutionary trees and allows for the study of phylogenetic composition and structure of soil mesofaunal communities.

We focus our efforts on the communities of Coleoptera (beetles), the presumed largest radiation of living organisms, which constitute a major component of the soil biota. Soil-inhabiting beetles comprise most major evolutionary lineages and functional guilds of the Coleoptera, including predators, scavengers, fungivores and herbivores feeding on roots (Burges & Raw 1967). There are different degrees of associations with the soil biome. While some ground-dwelling groups use the superficial leaf litter-humus (epigeic species), frequently for shelter only, others occur in the deeper horizons (humiculous-endogeic species), either in the larval stages, as several groups of root feeders, or for their entire life cycle. Some species exhibit specific adaptations to life underground, such as atrophy of the eyes and a reduction in body size, among others (Jeannel 1963). The variation in lifestyle and dependency on the soil would suggest differences in population structure. For example, the movement of agile, flighted predators in Staphylinidae (rove beetles) and Carabidae (ground beetles) foraging in the leaf litter is not constrained by the soil habitat. Likewise, larval root feeders including leaf beetles in the subfamily Galerucinae and predators such as Cantharidae (soldier beetles), although less mobile, may still disperse widely in the adult stages present in aboveground habitats. In contrast, permanently endogeic, flightless lineages may not disperse easily and form locally differentiated variants.

The great diversity of Coleoptera makes them a useful group to clarify the community assembly of soil mesofauna across regions and soil layers. This study focuses on soil communities from southern Spain, as a model for the analysis of species richness and turnover in largely undisturbed soil ecosystems. Soil arthropod communities grow in complexity with the geological age of soils, as evident from reduced species diversity in recent postglacial compared to Pleistocene soils (Zaitsev *et al.* 2012). The general stability of Iberian ecosystems that persisted during the Pleistocene glaciations can be expected to have produced a complex soil profile structuring the resident communities over evolutionary timescales. Therefore, Iberian soil organisms are highly suitable to investigate how contemporary ecological factors and evolutionary lineage history determine community composition. Using the phylogenetic trees from mitochondrial genomes obtained by sequencing pools of all locally encountered species, we were able to characterize these local assemblages in the framework of phylogenetic community ecology comparing compositional and phylogenetic diversity (Webb *et al.* 2002; Graham & Fine 2008). Full knowledge of species diversity and phylogenetic history at the whole-community level helps to disentangle the role of the geographic turnover vs. vertical stratification and to identify the community-level processes driving species diversification, spatial structure and evolutionary assembly in deep and superficial soil layers.

Material and methods

Soil sampling

Soil samples were collected from the southern Iberian Peninsula at Sierra de Grazalema, Cádiz (*CAD*), and Sierra de Cabra, Córdoba (*COR*), situated south of the Guadalquivir river in the Betic geologic domain, and Sierra Madrona, Ciudad Real (*CR*), to the north at the border of the ancient Iberian Massif (Fig. S1, Supporting information). Samples were collected from 28, 20 and 21 soil pits at *CAD*, *COR* and *CR*, respectively, representing a defined set of environments from open grassland to ancient *Quercus* forest (Table S1, Supporting information). Each soil pit was divided into two samples, corresponding to (i) the superficial layer of the soil (*SUP*; 1 m^2 of leaf litter up to 5 cm deep) and (ii) the deeper fraction (*DEEP*; sampling volume of 2500 cm^3 up to 40 cm deep). *SUP* samples were sifted with a Winkler apparatus (0.5 × 0.5 cm mesh) and subsequently extracted using a modified Berlese apparatus. *DEEP* samples were initially floated in water; sediments were discarded and the water was filtered with a 100 × 100 μm mesh to obtain a bulk of organic matter and soil fauna, which was processed using a Berlese apparatus. All larval and adult Coleoptera were preserved in absolute ethanol.

DNA extraction, sequencing and NGS data processing

Specimens from each sample were classified to morphospecies. DNA extractions were conducted on individual specimens. The 5′ portion of the *cox1* gene (barcode fragment) was PCR-amplified and sequenced with the Sanger method and ABI technology. According to our focus on the effect of the geographic location and soil layer on soil diversity of beetles, six Illumina TruSeq DNA libraries, one per region and soil layer, were prepared by pooling DNA extracts to generate roughly equimolar DNA concentration per specimen. Aliquots of 2, 4, 10, 20 and 40 μL per DNA extract were pooled according to their DNA concentration (respectively, (i) >250 ng/μL; (ii) 100–250 ng/μL; (iii) 40–100 ng/μL; (iv) 20–40 ng/μL; and (v) 0.1–20 ng/μL) as measured in Nanodrop 8000 UV–Vis Spectrophotometer (Thermo Scientific). The six Illumina TruSeq DNA libraries were sequenced on the MiSeq platform (2 × 250 bp) at about 30–50% of a flow cell each.

Each DNA library was assembled using Celera Assembler v7.0 (Myers 2000) (Data S1, Supporting information). Mitochondrial contigs were filtered against a reference database including 245 nearly complete coleopteran mitochondrial genomes (M. Timmermans, C. Barton, J. Haran, D. Ahrens, L. Culverwell, S. Dodsworth, P.G. Foster, L. Bocak & A. Vogler, unpublished data) and subsequently annotated using COVE v2.4.4 (Eddy & Durbin 1994) and tRNA covariance models. FeatureExtract (Wernersson 2005) was used to extract inter-tRNA regions corresponding to the protein-coding genes, which were individually aligned, edited and reconcatenated to get the final contigs (details in Data S1, Supporting information).

Two data sets were generated from the contigs with the aim of generating a backbone phylogenetic tree to place shorter metagenomic contigs and Sanger *cox1* sequences. (i) The *3KB* data set includes contigs >3000 bp plus the 245 reference sequences that were combined for a 'minimum contig-length' supermatrix. Orthology of contigs is not certain because multiple nonoverlapping contigs may correspond to a single mitochondrial genome. Putatively nonoverlapping contigs were combined into a unique sequence, if they occupy a similar position in the phylogenetic tree and show low divergence from each other in preliminary phylogenetic trees obtained using BEAST (Drummond *et al.* 2012; details in Data S1, Supporting information). Phylogenetic trees were also explored to identify and remove noncoleopteran sequences, after confirmation of top hits in the NCBI database. (ii) The *BC* data set was generated from contigs of any length containing a fragment of minimally 100 bp of the *cox1* barcode (positions 1808–1907 of the *T. castaneum* mitogenome), for a '*cox1* barcode-centred' supermatrix. This contig set ensures orthology of all terminals in the matrix.

Phylogenetic inference

The *3KB* data set was used for phylogenetic inference on amino acid sequences in PhyloBayes (Lartillot & Philippe 2004), running two independent chains under a CAT-Poisson model for 168 h. Trees retrieved from both chains were combined after discarding 50% initial trees as burn-in, and the maximum clade probability tree was estimated using TreeAnnotator (Drummond *et al.* 2012). The amino acid sequences were used in this step as they provided an improved phylogeny of Coleoptera at the higher taxonomic level (M. Timmermans, C. Barton, J. Haran, D. Ahrens, L. Culverwell, S. Dodsworth, P.G. Foster, L. Bocak & A. Vogler, unpublished data). The tree obtained was used as backbone constraint in RAxML (Stamatakis *et al.* 2008) using the -r function and a combined DNA alignment (*3KB+BC*) to place contigs from the *BC* data set into the backbone phylogeny, using a GTR+Γ model and conducting 10 searches for the best ML tree and 100 bootstrap pseudoreplicates. Similarly, the resulting tree (including both *3KB* and *BC* contigs and reference sequences) was used as a backbone to place the Sanger sequenced *cox1* barcodes. Branch lengths of final trees were re-estimated to be ultrametric using nucleotide data and fixing the topology in BEAST (operators *arrowExchange*, *wideExchange*, *wilsonBalding*, *subtreeSlide* inactivated). Analyses were run for 1–2.5 × 10^7 generations sampling every 5000th generation under a GTR+G model, an uncorrelated log-normal (ULN) clock and taking median values for branch lengths after discarding 50% initial trees as burn-in. Tree searches were conducted on the CIPRES portal (Miller *et al.* 2010).

Species delimitation and community composition

Ultrametric trees were used for species delimitation applying the single threshold algorithm of the generalized mixed Yule coalescent (GMYC) model (Pons *et al.* 2006). The GMYC was applied to trees generated from (i) the *3KB* contigs, (ii) the *BC* contigs and (iii) the *BC* contigs + Sanger *cox1* sequences. Respectively, the tree used for each data set was obtained by pruning all constrained terminal branches from the tree generated with the backbone approach as described above.

Contigs and Sanger barcodes were used to screen for species presence in each Illumina library by plotting the number and distribution of reads that match each sequence (*matched reads*). Positive identification of a species in the sample required a minimum of two *matched reads* of 150 bp and 100% similarity with an existing contig. Sister GMYC species that shared the match of two or more reads were collapsed into a single species (for consistency with the identification based on *matched reads*). Using this information, we built a matrix for presence of GMYC species in each library and a corresponding species tree by retaining a single terminal per GMYC species.

Total phylogenetic diversity (PD) for each community was quantified as the total branch length spanned by the tree including all its member species (Faith *et al.* 2009). We used the function *phylocurve.perm* (Nipperess & Matsen 2013) with 999 randomizations to estimate the expected PD for each community, after rarefaction for the minimum species number in any community to normalize for species richness (Gotelli & Colwell 2001).

Compositional dissimilarity of communities was estimated using the Sørensen index and its additive turnover and nestedness components (Baselga & Orme 2012). For estimates of phylobetadiversity, we used the analogous *1-Phylosor* index (Bryant *et al.* 2008) that considers the fraction of branch length shared among communities and ranges from 0 (no dissimilarity) to 1 (complete dissimilarity). We compared the observed phylobetadiversity between pairs of communities with a null model where species richness and turnover were fixed and only the identity of the species was randomized (999 iterations) (Graham *et al.* 2009; Leprieur *et al.* 2012). Additionally, to visualize the ordination of the communities based on compositional and phylogenetic information, we performed a principal coordinates analysis (PCoA) and used the function *envfit* (Oksanen *et al.* 2015) with 999 permutations to check for the correlation between its main axes and the regional and soil layer vectors. These analyses were performed using the R-packages *vegan* (Oksanen *et al.* 2015), *ade4* (Thioulouse *et al.* 1997) and *betapart* (Baselga & Orme 2012).

Based on the presence/absence matrices, species were classified by geographic region (CAD, CAR and CR) and by the soil layer they inhabit (exclusively found in the deep layer, exclusively in the superficial layer, or both). The phylogenetic clustering of these groups was assessed comparing their observed phylogenetic diversity (PD) and mean nearest taxon distance (MNTD) to the pattern expected under a null model of 999 community randomizations (independent swap algorithm; Graham *et al.* 2009). The resultant indexes (PD_{SES} and $MNTD_{SES,}$ respectively) and their associated *P*-values indicate whether species in a community are phylogenetically more closely related (clustered; <0) or less closely related (overdispersed; >0) than expected by chance (Webb *et al.*

2002). The species pool used in null model analyses included all the species found in the soil samples of the different regions, thus allowing to identify the potential role of the dispersal limitation at such regional scale (Cornell & Harrison 2013). The phylogenetic clustering indexes and randomization tests were performed using *Picante* (Kembel *et al*. 2010). Figure S2 (Supporting information) summarizes the proposed workflow.

Results

DNA extractions were performed on up to three specimens per morphospecies and local sample, for a total 535 adult and 959 larvae of presumed Coleoptera. DNA pools were generated from the vouchered extractions separately for the three sites and the two soil layers. The resulting six Illumina libraries included a total of >46 million paired reads, and assemblies in Celera yielded 273488 contigs, of which 0.96% were identified as mitochondrial sequences using BLAST (Table 1). All contigs with a minimum length of 3000 bp were combined into the *3KB* data set. Preliminary phylogenetic analyses recognized several non-Coleoptera contigs apparently resulting from misidentified larval specimens, which were removed. The final *3KB* set included 214 new mitochondrial contigs and 245 Coleoptera reference sequences of M. Timmermans, C. Barton, J. Haran, D. Ahrens, L. Culverwell, S. Dodsworth, P.G. Foster, L. Bocak and A. Vogler (unpublished data). In addition, the *cox1*-centred *BC* data set was created containing 264 new contigs (Table 1; Tables S2–S4, Supporting information). Sanger sequencing for the vouchered specimens resulted in 1128 sequences (75% success), of which 518 and 295 sequences, respectively, for the *DEEP* and *SUP* samples clustered with Coleoptera, while 315 apparently noncoleopteran barcodes were discarded (Table S5, Supporting information).

Phylogenetic analyses on amino acid sequences for the *3KB* data set resulted in an overall well-supported tree (Fig. 1), closely matching the topology of basal relationships in Coleoptera by the reference set alone (M. Timmermans, C. Barton, J. Haran, D. Ahrens, L. Culverwell, S. Dodsworth, P.G. Foster, L. Bocak & A. Vogler, unpublished data). The phylogenetic tree generated from the *BC* data set and *cox1* barcodes revealed numerous clusters of closely similar terminals, frequently composed of (near-)identical sequences from barcoding and metagenomic contigs, in particular if obtained from the same site (Fig. 2; Fig. S2, Supporting information). The tree indicated the congruence of results from either approach, and linked identifications made on *cox1*-barcoded vouchers to metagenomic contigs (Fig. 2; Fig. S2, Supporting information). The tree showed good recovery of families and superfamilies within Coleoptera, with both monophyletic Adephaga and Polyphaga. Within the latter, Scarabaeiformia, Bostrichiformia and Cucujiformia were retrieved as monophyletic, with paraphyletic Elateriformia and Staphyliniformia. At the superfamily level, Elateroidea, Buprestoidea, Staphylinoidea, Histeroidea, Cleroidea, Tenebrionoidea, Chrysomeloidea and Curculionoidea were monophyletic. Our contigs were distributed widely across the tree, with notable clusters in Carabidae (16 and 30 terminals for *3KB* and *BC*, respectively), Elateroidea (27 and 36), Staphylinoidea (68 and 79), Tenebrionoidea (24 and 29) and Curculionoidea (33 and 35).

Species delimitation with GMYC resulted in 166 and 196 species for the *3KB* and *BC* data sets (excluding ref-

Table 1 Number of studied specimens, obtained reads from Illumina sequencing and assembled contigs in the *3KB* and *BC* data sets for the studied soil communities (one per region and soil layer)

	CAD DEEP	CAD SUP	COR DEEP	COR SUP	CR DEEP	CR SUP	Total
Number of specimens	327	471	203	157	166	170	1494
Reads (millions) ×2	11.9	9.4	7.1	4.8	5.9	7.3	46.3
Reads without adapters (millions) ×2	9.3	8.5	5.8	4.5	4.8	6.8	39.6
Number of contigs	65 577	39 640	40 069	39 763	63 785	24 754	27 3588
Mitochondrial contigs (>50 bp after tRNA-based gene extraction)	529	655	352	164	342	319	2361
Contigs >10 000 bp	30	18	14	13	13	9	97
Contigs >3000 bp	90	72	54	29	39	43	327
Contigs (including cox1_100 bp)	88	76	44	27	42	40	317
Contigs in *3KB**	62	44	37	19	28	24	214
Contigs in *3KB* >10 000 bp*	36	11	14	11	13	10	95
Contigs in *BC**	78	55	42	21	35	33	264
Contigs in *BC* >10 000 bp*	25	9	10	10	8	5	67

Regions: CAD, Cádiz; COR, Córdoba; CR, Ciudad Real; SUP, superficial; DEEP, endogeic.

*Data refer to the 13 mitochondrial protein-coding genes.

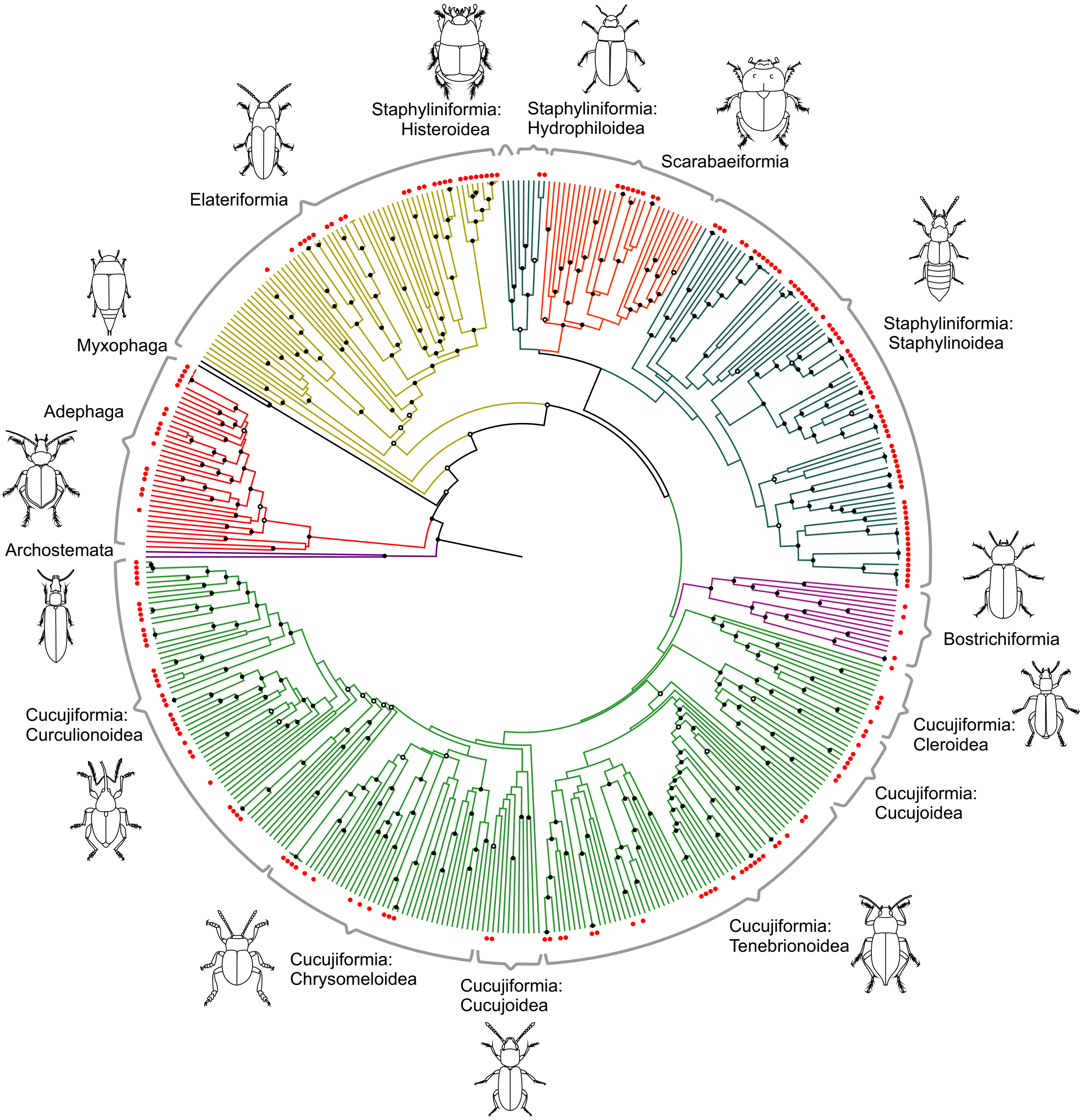
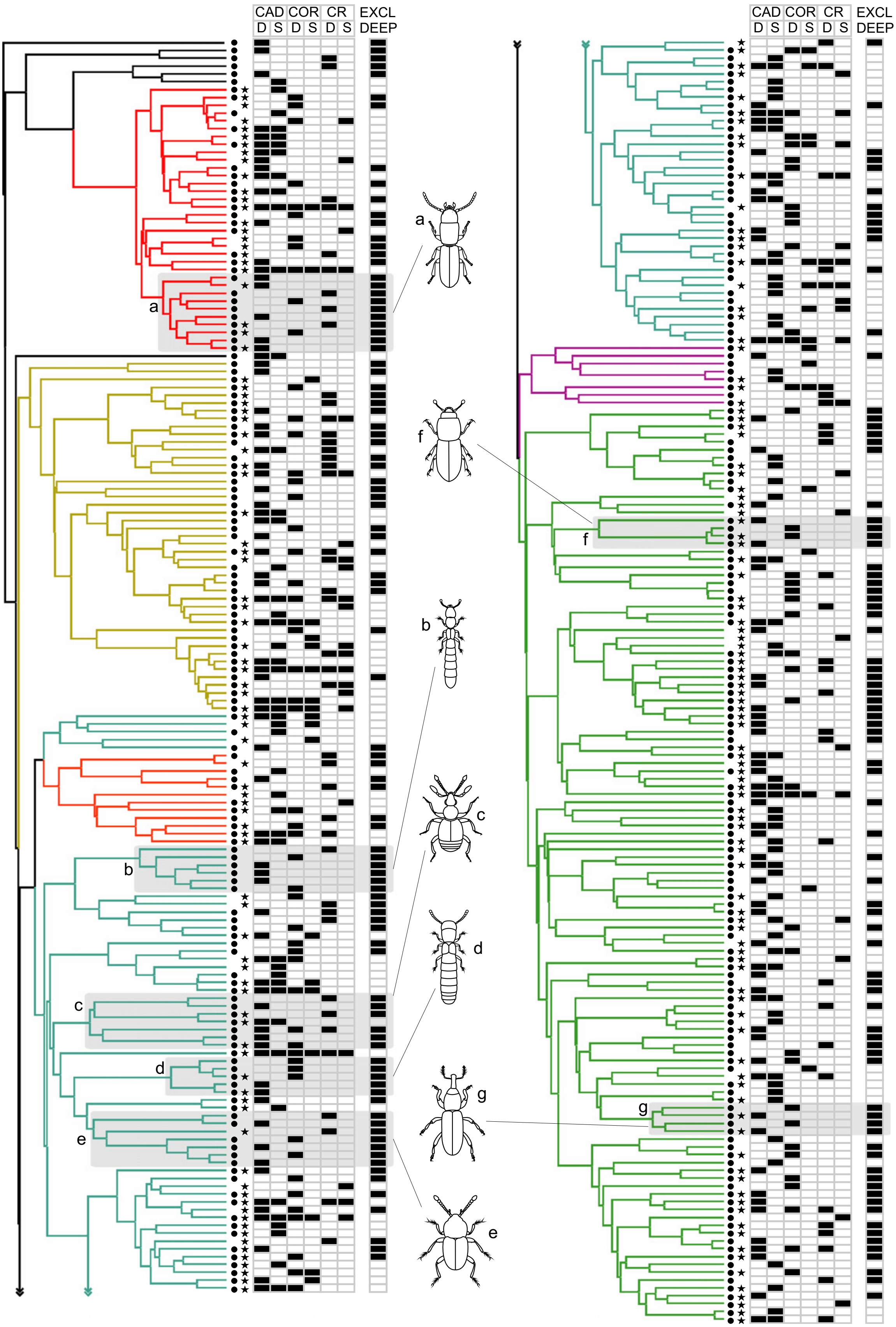


Fig. 1 Ultrametric Bayesian tree obtained in PhyloBayes and BEAST for the data set including 214 contigs longer than 3000 bp and 245 reference mitogenomes for Coleoptera (*3KB* data set). Posterior probability support ≥0.9 is indicated by black circles and support of 0.8–0.89 by white filled circles. Red circles on tips mark the mitogenomic contigs.

Fig. 2 Ultrametric Bayesian tree collapsed to GMYC species for the *cox1*-centred mitogenomic contigs and the Sanger-sequenced *barcodes* (*BC+Sanger*). Colour of clades: main lineages within Coleoptera as in Fig. 1. Circles on tips mark species with barcodes; stars mark species with mitogenomic contigs. Panels: presence (black)/absence (white) of each species in each community (in columns, from left to right: Cádiz deep layer, Cádiz superficial layer, Córdoba deep layer, Córdoba superficial layer, Ciudad Real deep layer, Ciudad Real superficial layer and finally, taxa only present in deep soil layers. Highlighted in grey are the clades restricted to the deep soil with unique species for each region (and for which adult specimens were found): a) Anillini, b) Leptotyphlini, c) Pselaphidae, d) Osoriini, e) Scydmaenidae, f) Anommatini, g) Torneumatini.


CAD COR CR EXCL
D S D S D S DEEP
a
b
c
d
e
f
g
CAD COR CR EXCL
D S D S D S DEEP

erence sequences), whereas 324 species were obtained for the combined *BC* and Sanger barcodes, of which 152 (47%) were shared, 36 (11%) were obtained exclusively with metagenomics and 136 (42%) exclusively with barcoding (Fig. 2). All major phylogenetic lineages were captured by either methodology, although some sublineages, for example the tribe Leptotyphlini (Staphylinidae) and several Scydmaenidae, were picked up mainly with the PCR-based approach, possibly because of their minute body size and consequently low DNA yield in the metagenomic mixtures.

All reads from the six Illumina libraries, including those not incorporated into the contigs, were matched against the contigs and barcodes (see Material and methods), to establish the distribution of each GMYC group across all libraries. By considering sequence reads, rather than assembled contigs, the discovery rates using metagenomic sequencing increased greatly. For example, when mapped against the 813 *cox1* Sanger sequences (288 GMYC species), 733 sequences (243 GMYC species) were identified with at least one *matched read* of 100% and ≥100 bp, representing 90% and 84% of sequences and species, respectively. Two sister taxa shared ≥2 *matched reads* and were collapsed into a single species for estimates of species numbers in seven pairs of sister GMYC species for the *3KB* and one pair for the *BC* data sets.

Compositional and phylogenetic diversity of communities

Species richness decreased from south to north, that is from CAD to COR to CR (Table S6, Supporting information), but at each region the species numbers were higher in *DEEP* than in *SUP* communities. Phylogenetic diversity closely matched the patterns of species richness. After PD rarefaction based on the overall lowest number of 24 species in any community, differences between communities were clearly reduced (Table S6, Supporting information).

The multisite beta diversity (compositional dissimilarity) and phylobetadiversity (phylogenetic dissimilarity) showed strong differences among communities, with values of >0.8 and >0.7, respectively (Table S7, Supporting information). By separating turnover and nestedness components of beta diversity and phylobetadiversity, approximately 85% of the total could be assigned to turnover for both metrics. Compositional and phylogenetic dissimilarities were greater among the three regions than between the two layers, and the turnover components among regions were higher (11 of 12 comparisons) for deep layers than superficial layers (Fig. 3; Table S7, Supporting information). For the three data sets, the PCoA ordinations on species presence/absence across axis 1 pointed to regional differences as the main factor driving compositional dissimilarities between communities (Fig. 3), and concordantly, the *envfit* function detected strong and significant correlations between axis 1 and the region vector (Table S8, Supporting information). Ordinations clearly showed that differences between soil layers split communities along axis 2 (Fig. 3), but such relationship was not found significant across the different data sets (Table S8, Supporting information).

Phylobetadiversity in most comparisons of deep and superficial communities was higher than expected from simple species beta diversity, a finding that was mainly evident in the *BC+Sanger* data set with its most complete species coverage. No such increase in phylogenetic differences compared to compositional differences between communities was seen for comparisons within either layer (Table 2). Applying the PD and MNTD metrics widely used in phylogenetic community ecology, the composition of communities in the three regional species pools showed no phylogenetic clustering or overdispersion (evenness) (Table 3). In contrast, species present only in the deep layer showed significant clustering across the tree, while no such pattern of phylogenetic clustering was evident in the exclusively superficial-layer species and, only partially significant for taxa present in both superficial and deep layers (Table 3).

Discussion

Compositional and phylogenetic diversity of soil beetle communities

Phylogenetic community ecology provides an evolutionary framework of biodiversity (Webb *et al.* 2002; Graham & Fine 2008), but its application to species-rich and poorly known taxa remains limited. The integrated approach of mitochondrial metagenomics for *de novo* sequencing of mitogenomes and the direct mapping of reads against reference sequences can overcome several challenges of community-level phylogenetic studies and so revealed the magnitude, pattern and potential drivers of compositional and phylogenetic diversity of soil beetle communities.

First, the study showed the magnitude of beetle species diversity and the broad representation of major lineages of Coleoptera (Fig. 1). Sampling 69 soil samples from three geographic regions in southern Iberia produced 324 putative beetle species from combined mitogenomes and *cox1* barcodes (*BC+Sanger*), of which 179 species were exclusive to the deep-soil layer. Species-level entities at these sites were delimited and placed into known lineages of Coleoptera based on the

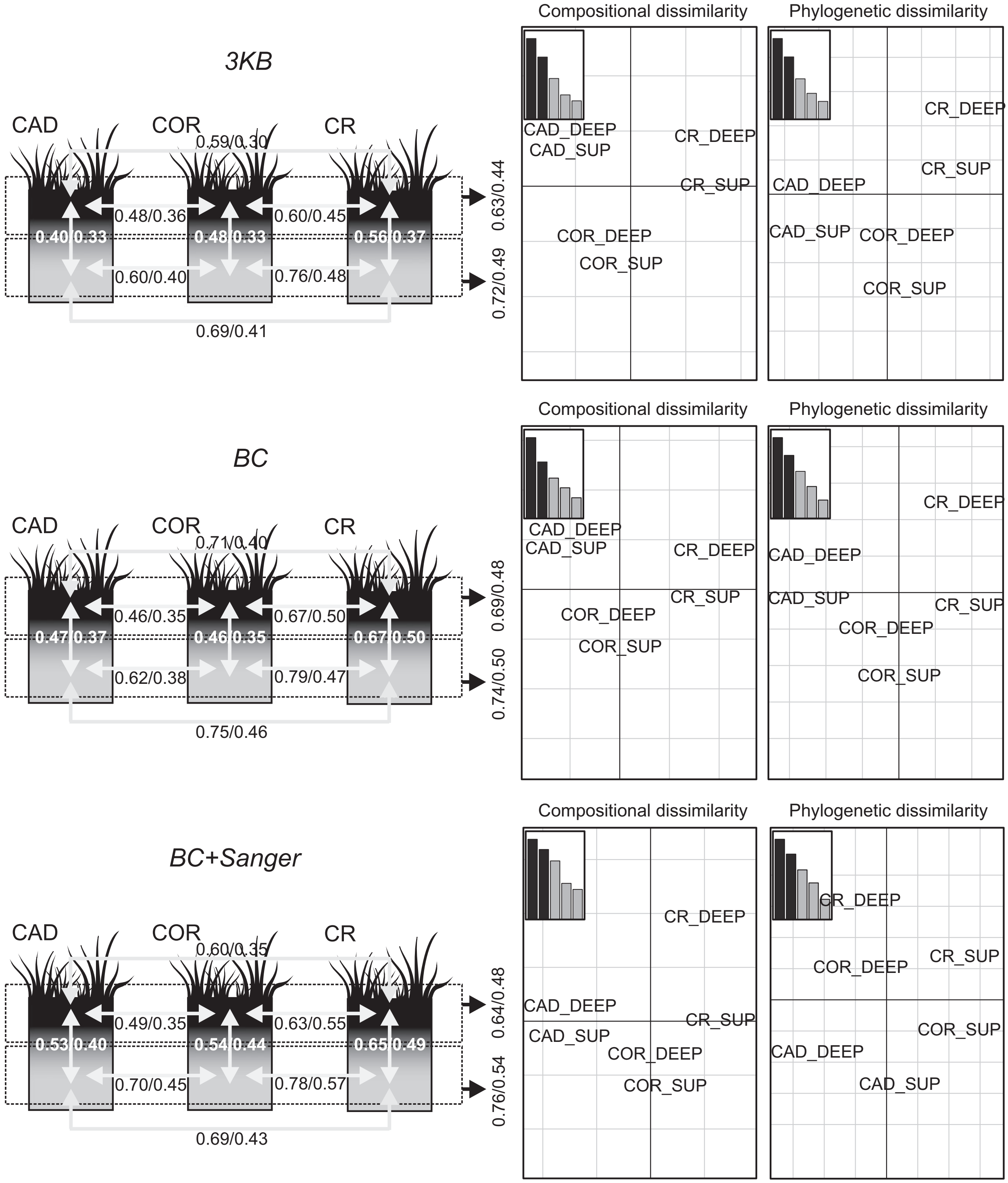


Fig. 3 Compositional and phylogenetic dissimilarities between communities in the *3KB, BC* and *BC+Sanger* data sets. Right panel: PCoA ordinations of communities using taxonomical (*Sørensen* index, sor) and phylogenetic (*1-Phylosor* index, psor) dissimilarity matrixes. Left panel: Compositional (*Simpson* index, sim)/phylogenetic (*1-Phylosor$_{turn}$* index, psim) turnover between the different pairs of communities. Regions: CAD, Cádiz; COR, Córdoba; CR, Ciudad Real. Soil layer: SUP, superficial; DEEP, endogeic.

phylogenetic power of long mitochondrial sequences which allowed for well-supported community-level evolutionary trees, a resource until now unavailable but essential for the performance of the phylogenetic community ecology analyses. Nearly half of these species were encountered in the larval stages (Fig. S2, Support-

Table 2 Comparisons of compositional and phylogenetic pairwise dissimilarities between communities in the *3KB*, *BC* and *BC+Sanger* data sets

	3KB		*BC*		*BC+Sanger*	
Communities	psor$_{SES}$	*P*-value	psor$_{SES}$	*P*-value	psor$_{SES}$	*P*-value
Between vertical layers						
CAD_DEEP-CAD_SUP	**2.883**	**0.009**	1.374	0.183	**2.962**	**0.005**
CAD_DEEP-COR_SUP	**3.567**	**0.003**	**2.299**	**0.023**	**2.867**	**0.007**
CAD_DEEP-CR_SUP	−0.968	0.325	0.229	0.845	**2.751**	**0.011**
COR_DEEP-CAD_SUP	0.908	0.357	1.841	0.059	**4.168**	**0.001**
CR_DEEP-CAD_SUP	**1.988**	**0.035**	1.237	0.221	**3.551**	**0.001**
COR_DEEP-COR_SUP	1.476	0.159	0.998	0.329	1.908	0.053
COR_DEEP-CR_SUP	−0.499	0.599	−0.019	0.963	**2.145**	**0.031**
CR_DEEP-COR_SUP	0.535	0.597	−0.431	0.645	1.662	0.091
CR_DEEP-CR_SUP	−0.555	0.583	1.021	0.281	**2.758**	**0.005**
Within vertical layers						
CAD_DEEP-COR_DEEP	0.917	0.365	−1.008	0.311	1.104	0.277
CAD_DEEP-CR_DEEP	0.355	0.741	0.032	0.973	0.735	0.467
COR_DEEP-CR_DEEP	−0.585	0.551	0.745	0.445	0.785	0.437
CAD_SUP-COR_SUP	1.264	0.203	1.446	0.155	1.264	0.203
CAD_SUP-CR_SUP	−1.343	0.177	−0.005	0.983	0.745	0.445
COR_SUP-CR_SUP	1.379	0.165	−0.098	0.917	1.806	0.063

Regions: CAD, Cádiz; COR, Córdoba; CR, Ciudad Real. Soil layer: SUP, superficial; DEEP, endogeic. Standardized effect sizes of the phylogenetic dissimilarity (psor$_{SES}$: *1-Phylosor* index) and *P*-values as obtained for null model comparisons where species richness and turnover among communities were fixed and only the identity of the species was randomized (999 iterations). Positive values in this index indicate a higher phylogenetic dissimilarity than expected from the compositional dissimilarity and negative values lower than expected. Significant values are in bold.

Table 3 Phylogenetic clustering of lineages of each region and soil layer in the *3KB*, *BC and BC+Sanger* data sets

Metric	CAD	COR	CR	DEEP	Both	SUP
3KB						
PD$_{SES}$	0.4428	0.0407	−1.124	**−2.561**	**−1.951**	0.574
PD$_{SES}$ *P*-value	0.662	0.498	0.124	**0.005**	**0.028**	0.703
MNTD$_{SES}$	1.264	−0.326	−1.646	**−2.485**	−1.634	0.700
MNTD$_{SES}$ *P*-value	0.894	0.377	0.055	**0.007**	0.051	0.747
BC						
PD$_{SES}$	0.452	1.707	−1.437	**−2.706**	−1.152	1.718
PD$_{SES}$ *P*-value	0.687	0.960	0.080	**0.005**	0.136	0.947
MNTD$_{SES}$	0.9899	0.9507	−0.7593	**−2.310**	−0.807	1.148
MNTD$_{SES}$ *P*-value	0.840	0.810	0.213	**0.017**	0.209	0.877
BC+Sanger						
PD$_{SES}$	0.436	−1.498	−1.267	**−2.414**	**−2.198**	0.773
PD$_{SES}$ *P*-value	0.665	0.076	0.107	**0.011**	**0.017**	0.214
MNTD$_{SES}$	−0.458	**−1.862**	−0.566	**−2.782**	−1.396	−0.947
MNTD$_{SES}$ *P*-value	0.329	**0.035**	0.296	**0.005**	0.091	0.173

Regions: CAD, Cádiz; COR, Córdoba; CR, Ciudad Real; DEEP, species found exclusively in deep layer; SUP, species found exclusively in superficial layer; Both, species present in deep and superficial layers. PD$_{SES}$, MNTD$_{SES}$: Standardized effect sizes of the phylogenetic diversity (PD) and the mean nearest taxon distance (MNTD) and *P*-values as obtained for null model comparisons (independent swap, 999 randomizations). Negative values in these indexes indicate phylogenetic clustering and positive values phylogenetic evenness (overdispersion). Significant values are in bold.

ing information), whose species circumscription and phylogenetic placement would have been very difficult with conventional methods. Phylogenetic lineages in the soil correspond to diverse functional groups, including predators, such as Staphylinidae, Scydmaenidae, Cantharidae and Carabidae; detritivores in Tenebrionidae and Scarabaeidae; or root feeders in Curculionidae and Chrysomelidae (Fig. 1). The approach here performed integrates species discovery and biodiversity analysis from local specimen collections for a much needed global taxonomic database of soil organisms. The phylogenetic position relative to identified lineages immediately provides information on functional ecology and probable guild membership and thus permits further developments including estimations of functional diversity and the potential ecosystem services provided by soil communities.

Second, the three major study sites provide a spatial perspective to the composition of soil communities. The site at Cadiz produced approximately twice as many species as the two other sites. Phylogenetic diversity (PD) was also higher, but when controlling for species number PD was remarkably similar for the three regions, pointing to a mostly uniform representation of the main Coleopteran lineages in all sites (Table S6, Supporting information). Possible higher habitat heterogeneity of the Cádiz region, due to the high floristic diversity and uniqueness of the Sierra de Grazalema, or its milder climatic conditions compared with the more continental Córdoba and Ciudad Real regions could be responsible for the higher diversity. Yet, compositional and phylogenetic dissimilarity of communities could be attributed almost entirely to 'turnover', in particularly for the deep-soil communities (Fig. 3, Table S7, Supporting information), which points to the importance of factors causing vicariant ranges to constrain soil community composition (Graham *et al.* 2009; Leprieur *et al.* 2012).

The patterns suggest strong dispersal limitations acting in soil beetle communities even at regional scales, in contrast to soil microbial patterns, but similar to initial findings in other mesofaunal groups (e.g. Erdmann *et al.* 2012). Recent studies on other soil taxa already revealed that at the global scale very few species are shared among sites (Wu *et al.* 2011; Nielsen *et al.* 2014). Local differentiation over fine geographic scales has already been established for Mediterranean soil communities of Collembola (Cicconardi *et al.* 2010), and high community turnover may result in potentially large undiscovered diversity and underestimation of species numbers due to incomplete geographic sampling (Cicconardi *et al.* 2013). These results highlight the restricted scale of the 'local community' for soil Coleoptera and support the idea of a primacy of neutral and/or dispersive processes driving the assembly of mesofauna communities at smaller scales than for other animal groups (Caruso *et al.* 2012). The topic also reopens the debate about the commonalities between diversity patterns above and below ground (Decaëns 2008). Our study used natural habitats of the Iberian Peninsula expected to harbour ancient and heterogeneous soils, which predicts greater species diversity than in more recently formed soil ecosystems (Zaitsev *et al.* 2012), a correlation that deserves further investigation. In addition, the study of codistributed invertebrates beyond the Coleoptera is required, to establish whether these findings hold generally for soil mesofauna. A denser scale of sampling sites is also needed to obtain greater precision on the magnitude and spatial scale of beta diversity. Finally, our study captured species diversity at the three sites by combining representative soil samples from the main forest and grassland habitat types, but this may overlook some degree of ecological turnover due to landscape heterogeneity (e.g. Kounda-Kiki *et al.* 2009).

Third, the analysis revealed clear differences between beetle communities from deep and superficial soil layers. Deep-soil communities showed both greater species diversity and PD than superficial-soil communities (Table S6, Supporting information). The high compositional and phylogenetic beta diversity between deep and superficial layers (Fig. 3, Table S7, Supporting information) points to a strong vertical stratification of soil beetle composition. This is true also for phylogenetic beta diversity between deep and soil layers which was even higher than what is expected from their species (compositional) turnover, in particular for the most complete matrix with the Sanger barcodes included that shows the vertical differentiation most clearly (Fig. 3; Table 2). Likewise, using metrics from phylogenetic community ecology, we find that the exclusive deep-soil taxa are phylogenetically clustered, while superficial assemblages are stochastically distributed across the tree of Coleoptera (Table 3). These findings reveal the existence of deep-soil specialist lineages, which were identified as typical endogean lineages consistently found in the deep soil only in our study, including Anillina, Leptotyphlini, Osoriini, Torneumatini, Anommatini, Pselaphidae and Scydmaenidae (Fig. 2). The phylogenetic trees revealed replacement among the three regions mainly at the tip level, where species are unique to a single region, resulting in the high regional turnover among the deep-soil communities (Table S7, Supporting information). Hence, at a regional scale, strictly deep-soil communities appear highly affected by geographic speciation. Yet, they show great phylogenetic distance from other lineages, indicating tight association with the deep soil over extended evolutionary periods, which results in the greater than expected phylogenetic turnover against lineages in superficial layers (Table 2).

Taken together, our results suggest that past and current geographic isolation is a plausible mechanism driving diversity turnover in soil beetle communities at regional scales. The intensity of these processes is mediated by the soil layers and, ultimately, by niche conservatism that maintains these processes (Wiens *et al.* 2010). For other groups of soil arthropods, mainly Collembola, a strong trade-off has been shown between the adaptation to the deep soil conditions and the species dispersal capacity and/or physiological tolerances (Ponge *et al.* 2006), and such specialization increases species vulnerability under climate change (Bokhorst *et al.* 2012). Our results on soil beetles support this idea and show that soil layer specialization could be a major driver of species diversity, structure and spatial assembly of soil communities.

Mitochondrial metagenomics to study the phylogenetic assembly of communities

The metagenomics approach overcomes the taxonomic impediments and main challenges of community phylogenetics to the study of complex, hyperdiverse and poorly known communities, arising from the difficulties of species circumscription, phylogenetic placement and community delineation (Graham & Fine 2008; Emerson *et al.* 2011). We obtained 95 complete or nearly mitogenomes (>10 kb of protein-coding genes), which contribute to firm estimates of relationships in Coleoptera and place the members of a community relative to known lineages. Contigs from independent libraries were highly similar for the length of the mitogenomes, which demonstrates that the assembly from mixtures of specimens is reliable and repeatable. The method also detected intraspecific variation used for establishing GMYC groups. For instance, in the *BC* data set, across the six libraries a total of 112 (of the 264 total) contigs were grouped into 44 GMYC species, and this variation was confirmed by Sanger sequences which revealed very close matches of the local variants (Fig. S2, Supporting information). However, the metagenomic analysis mainly revealed interlibrary variation, because closely related haplotypes from a single site are incorporated into a given contig, which masks the intrapopulation variation that was clearly evident in the Sanger barcodes.

The alternative criteria for selecting the *BC* and *3KB* data sets provide interesting insights into the efficiency of building phylogenetic matrices from the *de novo* assemblies. The *3KB* data maximized the number of long mitogenomes >10 kb (95 vs. 67 in *BC*), mainly by combining nonoverlapping contigs of presumed partial mitogenome sequences into a single terminal based on their placement and distances in the tree (see Material and methods). The longer sequences generally resulted in a better-supported phylogeny. The *BC* data set included a greater total number of contigs (266 vs. 214 in *3KB*), but many of them were short and their placement in the tree was less certain. However, the use of contigs centred on the *cox1* segment simplifies the matrix construction and allows direct comparisons with standard barcodes or metabarcoding data sets. At the sequencing depth used here, neither data set came close to incorporating all of the 324 species that were obtained by contig assembly and Sanger barcoding combined. The lower species coverage of contigs (188 species; 58%) compared to that of barcoding alone (288 species; 89%) may be a disadvantage for certain applications. However, the biodiversity patterns obtained from the metagenomic data sets closely match those obtained from the combination of contigs and barcodes sequences, as shown by the significant correlation (Mantel test) found for the compositional and phylogenetic dissimilarity matrixes for the *BC* vs. *BC+Sanger* data sets ($r^2 = 0.84$, $P = 0.001$ and $r^2 = 0.83$, $P = 0.004$, respectively) and *3KB* vs. *BC+Sanger* ($r^2 = 0.90$, $P = 0.004$ and $r^2 = 0.66$, $P = 0.007$). In addition, the placement of short *cox1* fragments is greatly improved in the presence of related, long mitogenome sequences without which the phylogenetic community patterns would not have emerged.

The mitogenome library allows for improved phylogenetic analysis for the target communities, unlike existing metabarcoding approaches, but in addition mitochondrial metagenomics provides a sensitive test for the presence of the corresponding species directly from extractions performed on complex bulk samples or even directly from the soil by read mapping against reference sequences. The Illumina sequencing output from our bulk samples contained reads that correspond to 90% (733 of 813) of *cox1* barcodes and 84% of the GMYC species obtained with Sanger sequencing. Thus, the cumulative sequence data will allow for phylogenetic community studies to be performed using species presence, and potentially abundance and intraspecific variation, solely generated by the direct mapping of high-throughput sequencing reads. Determining the species presence in the six libraries (Fig. 3) added many records that were missed in the assembled contigs. Hence, we need to distinguish the straightforward step of matching reads against a reference, from the more difficult task of building this reference set. Currently, the assembly step is a critical bottleneck in the described *de novo* metagenomics protocol that could potentially be overcome by greater sequencing depth, enrichment of mitochondrial DNA, longer reads or improved assemblers. Once a reference sequence exists, read mapping can be performed with much lower sequence coverage than is required for the initial assembly and may become an equally cost-effective way as metabarcoding for the characterization of communities. This will permit the community analysis of numerous samples, for example in the current study separating those combined at the landscape level, to test more specifically the link of community composition to factors such as soil type, soil age, stage of succession, aboveground vegetation and others. We envision a system by which numerous species can be monitored simultaneously as indicators of soil type or soil diversity, as a powerful tool for soil management and biodiversity conservation.

Acknowledgements

This research was founded by the Leverhulme Trust (grant F/00696/P to APV) and the NHM Biodiversity Initiative. CA received additional support of a Synthesys grant (GB-TAF-2966). PA was supported by a postdoctoral grant from the Royal Society (Newton International Program, UK). MTJN was a supported by a NERC Postdoctoral Fellowship (NE/I021578/1). Thanks are due to Alex Aitken, Stephen Russell, Andie Hall, Julia Llewellyn-Hughes, Peter Foster (NHM) and Shilo Dickens (University of Cambridge) for their technical assistance. Carles Hernando (IBE, Barcelona) helped with the taxonomic identifications, and Jesús Arribas collaborated during the sampling and did the drawings. Benjamin Linard and Carola Gómez-Rodríguez (NHM), two anonymous referees and the subject editor helped with constructive comments and advice.

References

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

Bardgett RD (2002) Causes and consequences of biological diversity in soil. *Zoology*, **105**, 367–374.

Baselga A, Orme CDL (2012) betapart: an R package for the study of beta diversity. *Methods in Ecology and Evolution*, **3**, 808–812.

Bates ST, Clemente JC, Flores GE *et al.* (2013) Global biogeography of highly diverse protistan communities in soil. *ISME Journal*, **7**, 652–659.

Bokhorst S, Phoenix GK, Bjerke JW *et al.* (2012) Extreme winter warming events more negatively impact small rather than large soil fauna: shift in community composition explained by traits not taxa. *Global Change Biology*, **18**, 1152–1162.

Bryant JA, Lamanna C, Morlon H *et al.* (2008) Microbes on mountainsides: contrasting elevational patterns of bacterial and plant diversity. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 11505–11511.

Burges A, Raw F (1967) *Soil Biology*. Academic Press, London.

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. *Journal of Animal Ecology*, **81**, 214–221.

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.

Cicconardi F, Fanciulli P, Emerson B (2013) Collembola, the biological species concept and the underestimation of global species richness. *Molecular Ecology*, **2**, 5382–5396.

Cornell HV, Harrison SP (2013) What Are Species Pools and When Are They Important? *Annual Review of Ecology, Evolution, and Systematics*, **45**, 45–67.

Crampton-Platt A, Timmermans M, 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*, in press.

Curtis TP, Sloan WT, Scannell JW (2002) Estimating prokaryotic diversity and its limits. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 10494–10499.

Decaëns T (2008) Priorities for conservation of soil animals. *CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources*, **3**, 014.

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

Dettai A, Gallut C, Brouillet S *et al.* (2012) Conveniently pre-tagged and pre-packaged: extended molecular identification and metagenomics using complete metazoan mitochondrial genomes. *PLoS ONE*, **7**, e51263.

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, **29**, 1969–1973.

Ducarme X, André HM, Wauthy G, Lebrun P (2004) Are there real endogeic species in temperate forest mites? *Pedobiologia*, **48**, 139–147.

Eddy SR, Durbin R (1994) RNA analysis using covariance models. *Nucleic Acids Research*, **22**, 2079–2088.

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 of London. Series B, Biological Sciences*, **366**, 2391–2402.

Erdmann G, Scheu S, Maraun M (2012) Regional factors rather than forest type drive the community structure of soil living oribatid mites (Acari, Oribatida). *Experimental & Applied Acarology*, **57**, 157–169.

Faith DP, Lozupone CA, Nipperess D, Knight R (2009) The cladistic basis for the phylogenetic diversity (PD) measure links evolutionary features to environmental gradients and supports broad applications of microbial ecology's "phylogenetic beta diversity" framework. *International Journal of Molecular Sciences*, **10**, 4723–4741.

Fierer N, Strickland MS, Liptzin D, Bradford MA, Cleveland CC (2009) Global patterns in belowground communities. *Ecology Letters*, **12**, 1238–1249.

Gaston KJ (2000) *Pattern and Process in Macroecology* (eds Gaston KJ, Blackburn TM). Blackwell Science Ltd, Malden, Massachusetts.

Gillett CPDT, Crampton-Platt A, Timmermans MJTN *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.

Gotelli N, Colwell R (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, **4**, 379–391.

Graham CH, Fine PVA (2008) Phylogenetic beta diversity: linking ecological and evolutionary processes across space in time. *Ecology Letters*, **11**, 1265–1277.

Graham CH, Parra JL, Rahbek C, McGuire JA (2009) Phylogenetic structure in tropical hummingbird communities. *Pro-*

ceedings of the National Academy of Sciences of the United States of America, **106**, 19673–19678.

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

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

Jeannel R (1963) *Monographie des Anillini Bembidiides Endoges: Coleoptera Trechidae*. Memoires du Museum national d'histoire naturelle, Paris.

Kembel SW, Cowan PD, Helmus MR *et al.* (2010) Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*, **26**, 1463–1464.

Kounda-Kiki C, Celini L, Ponge JF, Mora P, Sarthou C (2009) Nested variation of soil arthropod communities in isolated patches of vegetation on a rocky outcrop. *Soil Biology and Biochemistry*, **41**, 323–329.

Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, **21**, 1095–1109.

Leprieur F, Albouy C, De Bortoli J *et al.* (2012) Quantifying phylogenetic beta diversity: distinguishing between "true" turnover of lineages and phylogenetic diversity gradients. *PLoS ONE*, **7**, e42760.

Malé P-JG, Bardon L, Besnard G *et al.* (2014) Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. *Molecular Ecology Resources*, **14**, 966–975.

Martin J, Sykes S, Young S *et al.* (2012) Optimizing read mapping to reference genomes to determine composition and species prevalence in microbial communities. *PLoS ONE*, **7**, e36427.

Miller M, Pfeiffer W, Schwartz T (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: *Proceedings of the Gateway Computing Environments Workshop (GCE)*, pp. 1–8. New Orleans, Louisiana.

Myers EW (2000) A whole-genome assembly of *Drosophila*. *Science*, **287**, 2196–2204.

Nielsen UN, Ayres E, Wall DH, Bardgett RD (2011) Soil biodiversity and carbon cycling: a review and synthesis of studies examining diversity-function relationships. *European Journal of Soil Science*, **62**, 105–116.

Nielsen UN, Ayres E, Wall DH *et al.* (2014) Global-scale patterns of assemblage structure of soil nematodes in relation to climate and ecosystem properties. *Global Ecology and Biogeography*, **23**, 968–978.

Nipperess D, Matsen F (2013) The mean and variance of phylogenetic diversity under rarefaction. *Methods in Ecology and Evolution*, **4**, 566–572.

Oksanen J, Blanchet G, Kindt R *et al.* (2015) vegan: Community Ecology Package. R package version 2.2-1. http://CRAN.R-project.org/package=vegan

Ponge J-F (2013) Plant–soil feedbacks mediated by humus forms: a review. *Soil Biology and Biochemistry*, **57**, 1048–1060.

Ponge J, Dubs F, Gillet S, Sousa J, Lavelle P (2006) Decreased biodiversity in soil springtail communities: the importance of dispersal and landuse history in heterogeneous landscapes. *Soil Biology and Biochemistry*, **38**, 1158–1161.

Pons J, Barraclough T, Gomez-Zurita J *et al.* (2006) Sequence-Based Species Delimitation for the DNA Taxonomy of Undescribed Insects. *Systematic Biology*, **55**, 595–609.

Ranjard L, Dequiedt S, Chemidlin Prévost-Bouré N *et al.* (2013) Turnover of soil bacterial diversity driven by wide-scale environmental heterogeneity. *Nature Communications*, **4**, 1434.

Riesenfeld SJ, Pollard KS (2013) Beyond classification: gene-family phylogenies from shotgun metagenomic reads enable accurate community analysis. *BMC Genomics*, **14**, 419.

Stamatakis A, Hoover P, Rougemont J (2008) A rapid bootstrap algorithm for the RAxML Web servers. *Systematic Biology*, **57**, 758–771.

Straub SCK, Parks M, Weitemier K *et al.* (2012) Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. *American Journal of Botany*, **99**, 349–364.

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.

Thioulouse J, Chessel D, Dolédec S, Olivier J (1997) ADE-4: a multivariate analysis and graphical display software. *Statistics and Computing*, **7**, 75–83.

Wardle D (2002) *Communities and Ecosystems: Linking the Aboveground and Belowground Components*. Princeton University Press, Princeton.

Webb C, Ackerly D, McPeek M, Donoghue M (2002) Phylogenies and community ecology. *Annual Review of Ecology and Systematics*, **33**, 475–505.

Wernersson R (2005) FeatureExtract–extraction of sequence annotation made easy. *Nucleic Acids Research*, **33** (Web Server issue): W567–W569.

Wiens JJ, Ackerly DD, Allen AP *et al.* (2010) Niche conservatism as an emerging principle in ecology and conservation biology. *Ecology Letters*, **13**, 1310–1324.

Wilson EO (2002) *The Future of Life*. Knopf, New York.

Wu T, Ayres E, Bardgett RD, Wall DH, Garey JR (2011) Molecular study of worldwide distribution and diversity of soil animals. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 17720–17725.

Yu DDW, Ji Y, Emerson BBC *et al.* (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, **3**, 613–623.

Zaitsev AS, van Straalen NM, Berg MP (2012) Landscape geological age explains large scale spatial trends in oribatid mite diversity. *Landscape Ecology*, **28**, 285–296.

Zhou X, Li Y, Liu S *et al.* (2013) Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *Giga Science*, **2**, 1–12.

C.A., P.A. and A.P.V. conceived the work; C.A. and P.A. led the sample collection; C.A. and F.R. obtained the molecular data; C.A. and P.A. analysed the data; C.A., P.A. and A.P.V. led the writing, and all authors contributed to the discussion of results and the writing.

Data accessibility

RAW read data for the 6 Illumina libraries have been uploaded into SRA GenBank. BioProject Accession SRP056403.

Final DNA and amino acid sequence alignments, including *BC* and *3KB* data sets, Sanger *cox1* sequences and mitogenome references have been uploaded in Dryad doi:10.5061/dryad.f61bp.

Phylogenetic trees in Figs 1 and S3 (Supporting information) have been uploaded in Dryad doi:10.5061/dryad.f61bp.

Supplementary tables, figures and text have been uploaded as online Supporting Information.

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 Sampling regions in Southern Spain.

Fig. S2 Framework for the application of mitochondrial metagenomics to the study of the phylogenetic structure of communities.

Fig. S3 Maximum likelihood phylogenetic tree including all the sequences used in the current study.

Table S1 Sampling localities.

Table S2 Data on the contigs in the *3KB* dataset.

Table S3 Data on the contigs in the *BC* dataset.

Table S4 Data on the reference mitogenome sequences.

Table S5 Data on the Sanger *cox1* barcode sequences.

Table S6 Species richness and phylogenetic diversity (PD) of each community in the *3KB*, *BC* and *BC+Sanger* datasets.

Table S7 Compositional and phylogenetic dissimilarities among communities in the *3KB*, *BC* and *BC+Sanger* datasets.

Table S8 Comparison of 3KB, BC and BC+Sanger datasets based on Mantel correlations for the compositional and phylogenetic dissimilarity matrixes

Data S1 Extended details on the bioinformatic pipeline and primer information.