Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2016, **7,** 1071–1081 doi: 10.1111/2041-210X.12557

Metabarcoding and mitochondrial metagenomics of endogean arthropods to unveil the mesofauna of the soil

Paula Arribas[1,2]*, Carmelo Andújar[1,2], Kevin Hopkins[1], Matthew Shepherd[3] and Alfried P. Vogler[1,2]

[1]Department of Life Sciences, Natural History Museum, Cromwell Road, London SW7 5BD, UK; [2]Department of Life Sciences, Imperial College London, Silwood Park Campus, Ascot SL5 7PY, UK; and [3]Natural England, Renslade House, Bonhay Road, Exeter EX4 3AW, UK

Summary

1. Biological communities inhabiting the soil are among the most diversified, complex and yet most poorly studied terrestrial ecosystems. The greatest knowledge gaps apply to the arthropod mesofauna (0·1–2 mm body size) because conventional morphological and molecular approaches are in many cases insufficient for the characterisation of these complex communities. The development of high-throughput sequencing (HTS) methodologies is required to solve current impediments and to further advance our understanding of below-ground biodiversity.

2. We propose a flotation–Berlese–flotation (FBF) protocol for sampling and specimen processing to obtain 'clean' DNA extractions of arthropod mesofauna from the soil. In addition, we developed and tested HTS protocols for the characterisation of arthropod communities from these bulk DNA extractions using *cox1* metabarcoding and shotgun metagenomic sequencing on the MiSeq Illumina platform.

3. The FBF protocol provided DNA of soil arthropods from sufficiently large volumes of soil and free from contaminating bacteria and inhibitors. Metabarcoding and metagenomic sequencing on two deep soil samples from Iberian grasslands revealed > 100 species of Acari and Collembola from 28 families. Genome assembly straight from shotgun sequencing of bulk specimens produced partial and full mitogenomes for 54 species with average length of > 6000 bp. Metabarcoding and metagenomic sequencing resulted in closely congruent OTUs, but species numbers were highest with metabarcoding, while ~73% of species were confirmed by matching shotgun sequence reads and ~48% by contig assembly from those shotgun reads.

4. In combination, the FBF protocol together with the PCR-based and shotgun sequencing pipelines addressed most of the challenges of studying soil arthropod mesofauna on the MiSeq Illumina platform. They are powerful, cost-efficient tools for characterising soil diversity in a phylogenetic and community ecology context. These methodological developments of HTS approaches for the study of mesofauna will accelerate ecological and evolutionary studies, biomonitoring of soil arthropods, and progress in both theoretical and applied soil science.

Key-words: Acari, biomonitoring, Collembola, community structure, deep soil, high-throughput sequencing, mesofauna extraction, metagenome skimming, phylodiversity, soil biodiversity

Introduction

The fauna of the soil is considered a 'biotic frontier' (André, Noti & Lebrun 1994) that may comprise 25% of all multicellular species on Earth (Decaëns *et al*. 2006) with important roles in ecosystem processes (Bardgett & van der Putten 2014). These communities include species-rich lineages of arthropods that affect the physicochemical and biological properties of the soil and leaf litter through complex interactions of detritivores, primary producers and predators (Ponge 2013). However, poor knowledge of species diversity and community composition means that soil communities largely remain a 'black box' for biology (Decaëns 2010; Orgiazzi *et al*. 2015).

Several methodological and logistical issues have hindered our understanding of soil biodiversity. An important part of soil mesofauna (invertebrates ranging from 100–150 μm to 2–3 mm) is composed of arthropods whose study is difficult due to their great species richness, high abundance, small body size, local-scale heterogeneity of communities and poor taxonomic background knowledge (Bardgett 2002; Decaëns 2010). In addition, high levels of cryptic diversity were discovered with sequencing data, questioning the validity of morphology-based species circumscriptions (Cicconardi *et al*. 2010; Cicconardi, Fanciulli & Emerson 2013). Molecular data for endogeic groups are equally scarce in the NCBI data base; for example, only a single mitochondrial genome sequence is available for Oribatida, a large group of soil-dwelling mites. This paucity of knowledge has hampered the study of soil biodiversity even in relatively well-known regions such as Europe, leaving great uncertainties about total species richness and turnover, phylogenetic diversity, geographical structure, temporal dynamics and the role of arthropod communities on soil

*Correspondence author. E-mail: pauarribas@um.es ecosystem function (Fierer *et al.* 2009; Wu *et al.* 2011; Bardgett & van der Putten 2014).

The characterisation of previously unmeasurable diversity has seen substantial progress thanks to the implementation of high-throughput sequencing (HTS) for metabarcoding and metagenomics performed on bulk environmental samples (Zepeda Mendoza, Sicheritz-Ponten & Gilbert 2015). To date, these approaches mainly have been applied to the study of microbial communities of the soil, mostly for metabarcoding of bacterial and archaeal biodiversity (Orgiazzi *et al.* 2015), which is usually performed by DNA extractions directly on the soil matrix and amplification of the 16S rRNA gene. Only recently HTS approaches were extended to the analysis of macroscopic diversity (e.g. Ji *et al.* 2013), and metabarcoding of soil mesofauna has been performed using universal primers for Metazoa (e.g. Wu *et al.* 2011; Yang *et al.* 2013, 2014) or targeting particular lineages, such as nematodes (e.g. Griffiths *et al.* 2006). HTS now offers unprecedented possibilities to overcome past constraints to the study of soil mesofauna.

Protocols for HTS-based characterisation of metazoan soil communities currently in use have been adopted from studies of soil microbes and marine meiofauna (e.g. Fierer & Jackson 2006; Fonseca *et al.* 2010). In these procedures, the amount of processed soil was small (around 200 g per sample), but this volume is likely to be insufficient for capturing mesofaunal (and particularly arthropod) diversity. DNA extraction was usually directly from the soil matrix, which results in high proportions of bacterial and fungal DNA (e.g. Yang *et al.* 2013), together with the high levels of PCR inhibitors from the humus. In addition, the widely used 18S (SSU) rRNA gene has been used as the target marker for soil mesofauna despite its insufficient variability for molecular species delimitations (see Tang *et al.* 2012). Metabarcoding studies using the 'universal barcode' *cox1* on soil mesofauna have been limited to bulk samples of springtails using the now-obsolete 454 platform (Ramirez-Gonzalez *et al.* 2013).

Besides PCR-based approaches, the shotgun sequencing of total genomic DNA from bulk samples presents an alternative for the characterisation of complex communities. Genome assembly from mixtures of shotgun reads favours the contig formation from the (high-copy) mitochondrial fraction, which constitutes roughly 1% of all reads and produces complete or partial mitochondrial genome assemblies for individual species in the sample (Crampton-Platt *et al.* 2015). In a recent study of soil beetle communities, this 'mitochondrial metagenomics' (MMG) approach produced mitochondrial genome sequences for entire species assemblages, permitting the concurrent analysis of species diversity, phylogenetic structure and drivers of diversification (Andújar *et al.* 2015). To date, MMG has been applied exclusively to insects (e.g. Gillett *et al.* 2014; Crampton-Platt *et al.* 2015; Tang *et al.* 2015) but not yet to the extremely diverse and minute Acari and Collembola.

The methodological development of HTS approaches for the study of mesofauna has been identified as a priority for the field of soil biology (Orgiazzi *et al.* 2015) and was among the major challenges highlighted during the First Global Soil Biodiversity Initiative Conference 2014 (http://www.gsbiconference.elsevier.com). Here, we contribute towards this goal by applying both *cox1* metabarcoding and MMG to unveil the diversity of soil arthropods. First, we adapted classical soil sampling and sample processing protocols to obtain 'clean' DNA extractions of mesofauna from large volumes of soil as the starting material for efficient HTS analyses. Secondly, we address the methodological challenges of *cox1* metabarcoding on the MiSeq Illumina platform, which is hampered by the large size of the standard *cox1* barcode fragment (>650 bp) relative to the read length achievable with the Illumina platform (<300 bp). In addition, the large number of anticipated soil samples requires an accurate and efficient system for multi-library sequencing, which we achieve by adapting a dual-tagging PCR system to *cox1* metabarcoding. Thirdly, we extend metabarcoding and metagenomic approaches to Acari and Collembola, the major arthropod groups of soil ecosystems, including (i) validation of *cox1* primers on different lineages, (ii) evaluation of multiple options for the OTU delimitation and identification and (iii) comparison of metabarcoding results against the MMG approach, that also provides the first *de novo* assembly of mitochondrial genomes form bulk mesofauna samples. Using bulk soil samples from Mediterranean grassland habitats, we estimate the local species diversity and compare species recovery with various approaches. The combination of MMG and *cox1* metabarcoding allows the characterisation of soil arthropod diversity from a phylogenetic community ecology perspective and offers promising avenues to fill the gaps in our knowledge of these highly diverse communities.

Materials and methods

DEEP SOIL SAMPLING AND MESOFAUNA EXTRACTION

Two soil samples were collected from the Southern Iberian Peninsula at Sierra de Grazalema, Cádiz (CAD, 36.707424°/−5.456676°), and Sierra de Cabra, Córdoba (COR, 37.481117°/−4.388536°), both located in similar wet grassland habitats. After removing the superficial layer (up to 5 cm deep), we sampled a 9·5-cm diameter core to 35 cm depth, comprising *c.* 2500 cm^3 of soil (Fig. 1a). Deep soil samples were processed following a flotation–Berlese–flotation protocol (FBF) that allows for the 'clean' extraction of arthropod mesofauna (see Data S1, Supporting information, for details). Briefly, the FBF protocol is based on soil flotation in water, which allows the extraction of the organic matter and soil mesofauna from raw soil samples (Fig. 1b). Subsequently, the organic portion is placed in a modified Berlese apparatus to capture specimens alive and preserve them in absolute ethanol (Fig. 1c). The last step of the FBF protocol is an additional flotation of the ethanol-preserved arthropods, resulting in 'clean' bulk specimen samples ready for DNA extraction (Fig. 1d).

DNA EXTRACTION, PCR AMPLIFICATION, ILLUMINA SEQUENCING AND BIOINFORMATICS

Prior to bulk DNA extraction of each sample, 'voucher' specimens representing the main lineages of Acari and Collembola were picked from each sample and subjected to non-destructive individual DNA extractions. The *vouchers* were identified morphologically to family level.

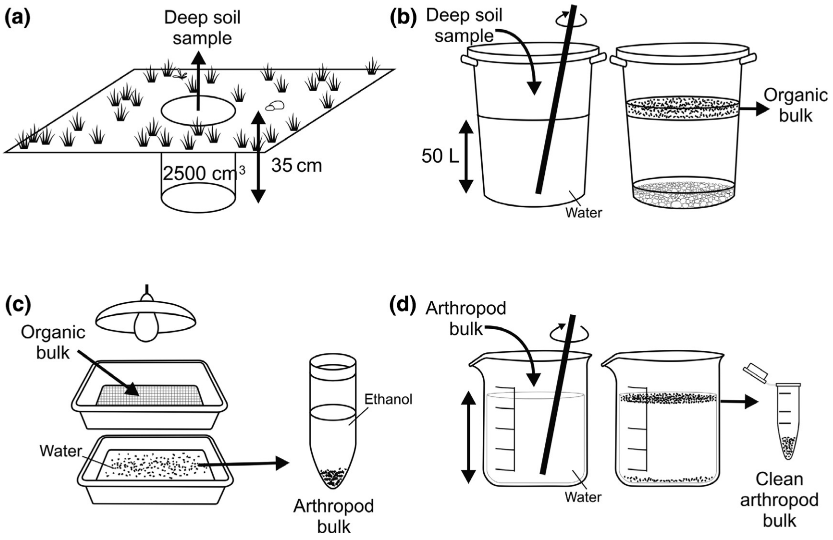

Fig. 1. Flotation–Berlese–flotation (FBF) protocol for the extraction of soil arthropod mesofauna.

DNA extractions were performed with the BioSprint 96 DNA Kit and a BioSprint 96 Workstation (Qiagen, Venlo, Netherlands). The 5′ end of the *cox1* gene (barcode fragment) was PCR-amplified using three primer pairs modified from the original barcode primers of Folmer *et al.* (1994) (see Table S1 for primer sequences). The best PCR product for each specimen was Sanger-sequenced with ABI technology. This set of *vouchers* reference sequences was further used (i) to evaluate the performance of the *cox1* amplification for different lineages of Acari and Collembola, (ii) to improve the taxonomic assignment for the obtained OTUs and (iii) to evaluate the performance of both *cox1* metabarcoding and MMG.

cox1 *metabarcoding of soil arthropod mesofauna*

The remaining specimens in each sample were homogenised, and one bulk DNA extraction per sample was performed using the DNeasy Blood and Tissue Spin-Column Kit (Qiagen). For metabarcoding, the 5′-*cox1* region of ~650 bp was amplified using the *FoldF-FoldR* and *LCO1490_short-HCO2198_short* primer pairs which target the same 'universal barcode' binding sites but differ in their primary sequence for increased breadth of sequencing success (see Table S1). Primers were modified to include an overhang adapter sequence for subsequent nested PCR, as an extension of the Illumina protocol for the 16S rRNA gene sequencing in microbial samples (16S Library Preparation Protocol at http://support.illumina.com). For each sample, six independent reactions for each pair of the two primers were performed (i.e. 12 independent PCR replicates per sample). All information regarding PCR reagents and conditions is included in Data S2. PCR *cox1* amplicons from each replicate for each sample were pooled and cleaned, after which these primary amplicons were used as template for a limited-cycle PCR amplification to add dual-index barcodes and the P5 and P7 Illumina sequencing adapters (Nextera XT Index Kit; Illumina, San Diego, CA, USA) (Fig. 2). The resulting metabarcoding libraries were each sequenced on an Illumina MiSeq sequencer (2 × 300 bp paired-end reads) on 2·5% of the flow cell, to produce paired of reads (R1 and R2) with a given dual combination of tags per sample (Fig. 2).

Raw reads from each metabarcoding library were filtered against a reference data base including all *cox1* sequences for Arthropoda from NCBI (February 2015) using BLAST (e-value = 10^{-5}). Retained reads were quality-filtered using TRIMMOMATIC v0.30 (Lohse, Bolger & Nagel 2012) and processed with *'fastx_barcode_splitter'* to split different primer reads and with *'fastx_trimmer'* to trim the primer sequence from the 5′ end. The main methodological challenge of *cox1* metabarcoding on the MiSeq Illumina platform results from the large size of the standard *cox1* barcode fragment (>650 bp) relative to the read length achievable with the Illumina platform (<300 bp), which causes a gap of ~100 bp between each pair of R1 and R2 read sequences (see Fig. 2). We therefore only used R1 reads for the clustering steps in the OTU delimitation, to which the corresponding paired-end R2 fragments were added in a subsequent step (see below). R1 primer-trimmed reads were trimmed at the 3′ end for a uniform length of 270 bp, discarding all shorter reads. The resulting data set was de-replicated, sorted according to decreasing abundance and subjected to *de novo* chimera detection and deletion following the UPARSE pipeline (Edgar 2013). OTU clustering was performed with three commonly used clustering algorithms: USEARCH v7 (greedy heuristic approach, Edgar 2010), CROP (Bayesian approach, Hao, Jiang & Chen 2011) and SWARM (agglomerative approach, Mahé *et al.* 2014), with two pre-clustering filtering options: maxeee = 1 and non-maxee (UPARSE pipeline; Edgar 2013) and under 20 similarity thresholds from 92·6 to 99·6%. After the comparison of the results, OTUs delimited for both CAD and COR samples (USEARCH algorithm, 97% similarity threshold) were combined and aligned with transAlign (Bininda-Emonds 2005) to generate the *metabarcode-cox1* data set for further analyses.

Mitochondrial metagenomics of soil arthropod mesofauna

For metagenomic sequencing, concentrations of total DNA extracts were measured using the Qubit dsDNA HS Assay Kit (Invitrogen, Waltham, MA, USA) and equimolar pooled aliquots were used to prepare two TruSeq DNA PCR-free metagenomic libraries to be further sequenced on a MiSeq sequencer (2 × 300 bp paired-end reads) using 50% of a flow cell (Fig. 2). Raw reads were quality-filtered, trimmed and assembled following the MMG pipeline (Andújar *et al.* 2015; Crampton-Platt *et al.* 2015, see Data S3 for details). The resulting mitochondrial contigs were assembled against a set of 12 mitochondrial genomes obtained from NCBI (November 2014) covering the main groups of Acari and Collembola using the 'Map to reference' function (minimum overlap = 200 bp; maximum mismatches per read = 50%) in Geneious. Protein-coding gene sequences extracted from the contigs were aligned with transAlign, edited and re-concatenated to produce partial and complete

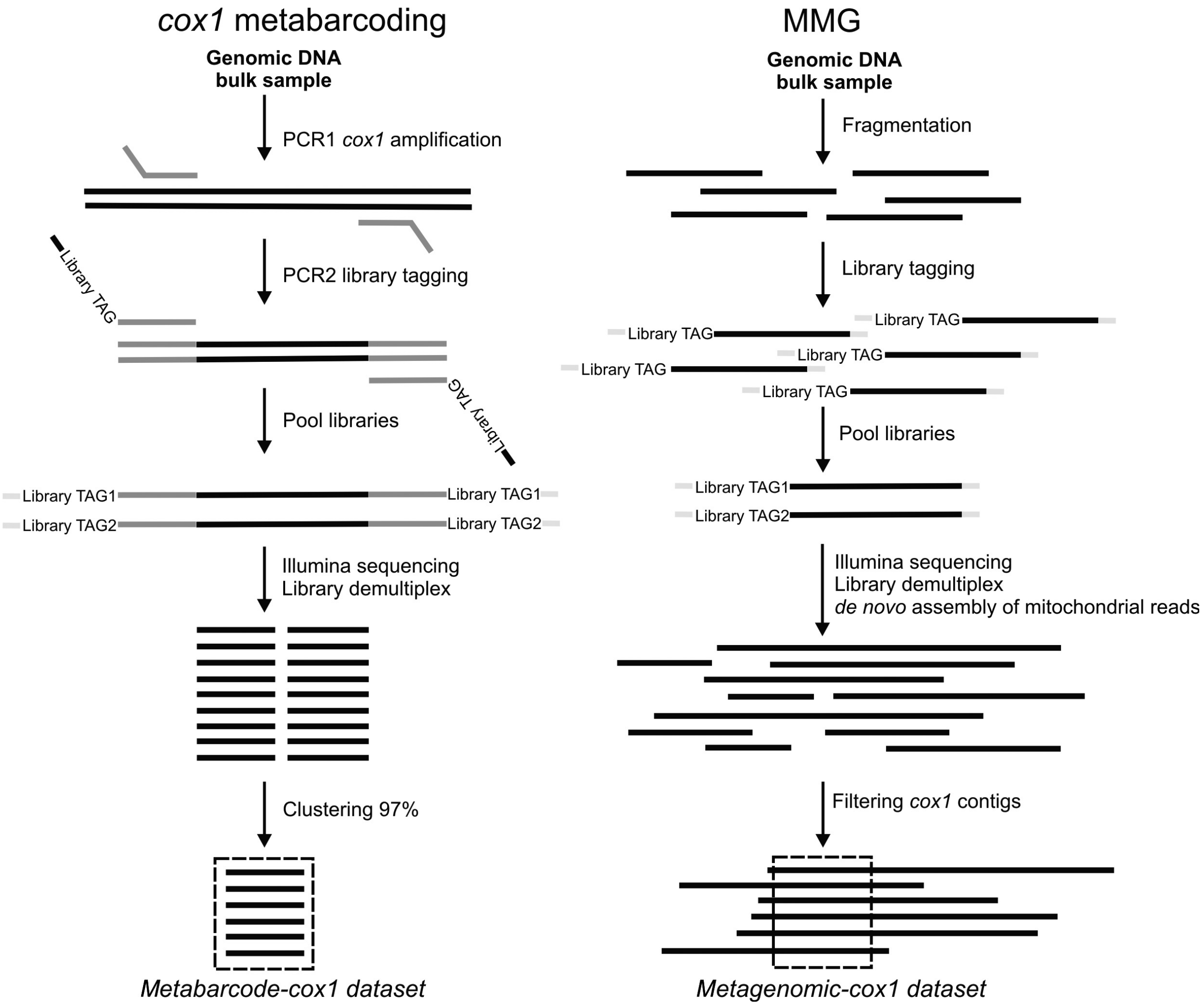

Fig. 2. Metabarcoding and mitochondrial metagenomics (MMG) pipelines for the study of soil arthropod mesofauna from bulk samples. For metabarcoding, the initial PCR is conducted with the locus-specific *cox1* primers to which an 'overhang adapter' is added as a template in the secondary PCR. The primers targeting these overhang regions contain the dual indices for variable tags (TAG1, TAG2) and the Illumina P5 and P7 sequencing adapters. For metagenomics, ligation is used to add the variable tags (together with the sequencing adapters) directly to the shotgun fragments. Note that in the MMG approach the great majority of tagged fragments corresponding to non-mitochondrial DNA are removed during the assembly process that favours the high-copy mitogenomes.

mitochondrial contigs (mitogenomes). Subsequently, the first 270 bp of the *cox1* gene (barcode fragment) was extracted from these mitogenomes to generate the *metagenomic-cox1* data set.

DIVERSITY ESTIMATION, SPECIES IDENTIFICATION AND DATA SET PERFORMANCE

The *metagenomic-cox1* and *metabarcode-cox1* data sets were combined and aligned together using MAFFT v6.240 (Katoh *et al.* 2002) and transAlign, and a final OTU definition (USEARCH at 97% similarity and maxee = 1) was performed to generate the *Combined OTUs* data set including a single representative sequence per OTU corresponding to the first 270 bp of the barcode *cox1* fragment (R1 reads). Additionally, a *Combined full-length OTUs* data set was generated by adding the corresponding R2 metabarcoding paired-end read or metagenomic contig to each OTU's representative sequence, to generate a full-length *cox1* barcode (~540 bp up to 650 bp). Subsequent analyses were performed with both the *Combined OTUs* and *Combined full-length OTUs* data sets to check for the effect of sequence length on OTU identification.

The taxonomic assignment of the OTUs (both *Combined* and *Combined full-length data sets*) was by matches to (i) the overall NCBI *nt* data base (November 2014) and (ii) *vouchers* sequences using BLAST (97% similarity over ≥ 150 bp). To test for improvements of OTU identification from adding these *vouchers* sequences to the NCBI nt data base, both NCBI nt + *vouchers* BLAST matches and only-NCBI nt BLAST matches were fed into MEGAN v5 (Huson *et al.* 2011) to compute the taxonomical affinity of each OTU with the lowest common ancestor algorithm (Huson *et al.* 2007). We followed the taxonomic ranks in the NCBI Taxonomy data base (February 2015) that does not implement recent changes in the assignment of some suborders.

The results of both HTS Illumina approaches were compared by assessing the consistent recovery of OTUs in the *metagenomic-cox1* and *metabarcode-cox1* data sets. Similarly, OTU recovery using directly the reads from the metagenomic libraries was evaluated using

BLAST (97% similarity over ≥ 150 bp). Finally, the proportion of *vouchers* sequences matching the delimited OTUs in the combined data sets was also assessed using BLAST (97% similarity over ≥ 150 bp).

Results

COX1 METABARCODING AND MITOCONDRIAL METAGENOMICS OF SOIL ARTHROPOD MESOFAUNA

A total of 384 Acari and 112 Collembola specimens were obtained in the CAD sample and 200 Acari and 104 Collembola specimens in the COR sample. The morphological identification of 79 *vouchers* showed a taxonomically broad range of Acari and Collembola, including representatives of the orders Oribatida, Astigmata (currently Sarcoptiformes including suborder Oribatida and its cohort Astigmatina), Trombidiformes, Mesostigmata, Entomobryomorpha and Poduromorpha (see Table S2 including GenBank voucher numbers).

The two metabarcoding libraries included a total of ≈960 000 paired reads. The number of OTUs delimited at different similarity thresholds showed a rapid asymptotic decrease from very high numbers to a largely stable value from *c*. 97·5% similarity and below. These final numbers of OTUs were very similar for the three clustering methods and pre -clustering filtering options used (Fig. 3). For instance, for a similarity threshold of 97%, the number of OTUs from the different algorithms and filtering treatments resulting in only small differences ranges from 106 to 114 OTUs for CAD and from 75 to 79 for COR (see Table S3). The USEARCH algorithm produced the data set most compatible with the other two analyses, recovering all OTUs defined by either of those, and thus, the sets of OTUs obtained with this algorithm using a 97% similarity threshold were combined and aligned to generate the *metabarcode-cox1* data set consisting of 180 sequences of 270 bp representing the OTUs from the two metabarcoding libraries.

Two metagenomic shotgun libraries generated from the same bulk samples included a total of >16 million paired reads whose assembly produced 43 and 38 contigs for the CAD and COR libraries covering the same *cox1* region of 270 bp. Both libraries were combined to generate the *metagenomic-cox1* data set consisting of 81 sequences. The alignment and OTU clustering of both combined *metagenomic-cox1* and *metabar-*

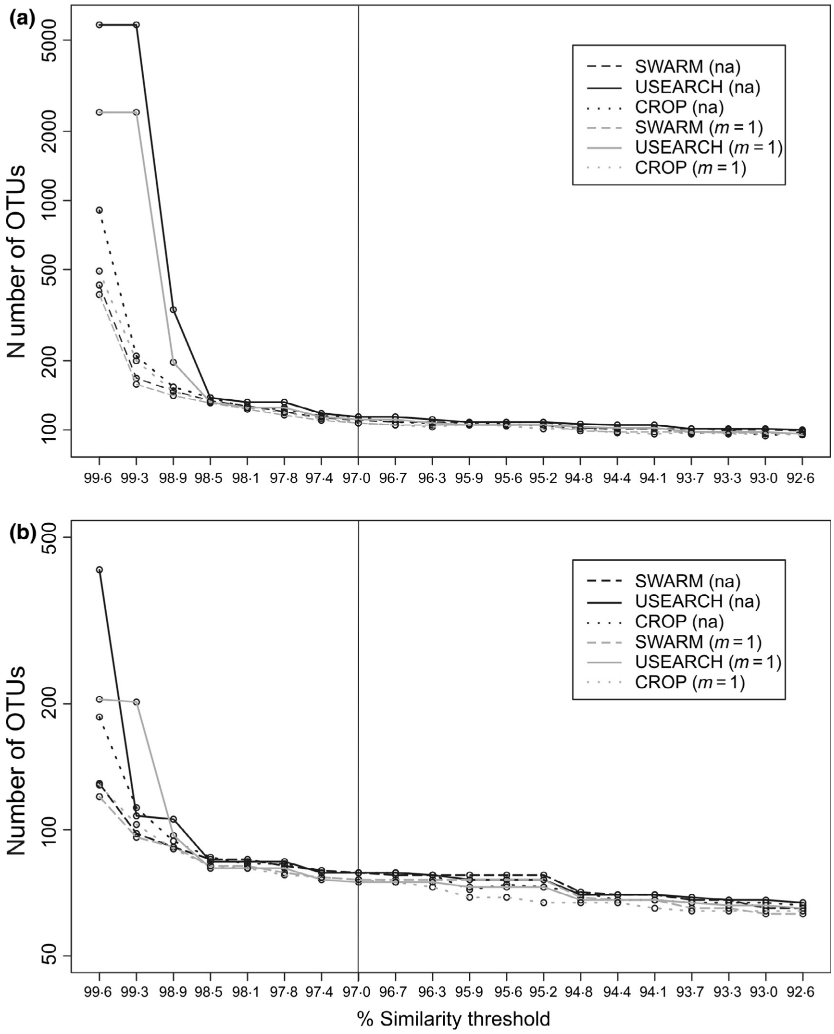

Fig. 3. Number of OTUs resulting from the *cox1* metabarcoding of (a) CAD and (b) COR soil mesofauna samples using: three clustering algorithms (USEARCH V7, CROP and SWARM); two pre-clustering filtering options [maxeee = 1 (*m* = 1) and non-maxee (na)] and 20% similarity thresholds. Vertical line marks the 97% similarity threshold used to generate the *metabarcode-cox1* data set.

code-cox1 data sets resulted in a final number of 177 entities, to produce the *Combined OTUs* data set and its extended *Combined full-length OTUs* data set after incorporating the R2 reads (Appendix S2).

DIVERSITY ESTIMATION, SPECIES IDENTIFICATION AND DATA SET PERFORMANCE

Using the BLAST matches to both NCBI nt + *vouchers* sequences and the lowest common ancestor assignment in MEGAN, 88% of the 177 OTUs from the *Combined full-length OTUs* data set were identified as arthropods, and of these, 70% were identified as Acari and Collembola (Table 1, Fig. S1). The remaining arthropod OTUs were assigned to Diptera, Coleoptera or Isoptera or could only be identified to higher taxonomic levels, such as Endopterigota or Arthropoda (Fig. S1). For Acari, 22 different families were present and included the common soil-inhabiting lineages Oribatida, Mesostigmata and Trombidiformes, but also the family Proctophyllodidae (Astigmata) usually considered a feather mite that might dwell in the soil during development or was the result of by-catch (Fig. 4). Six different families of Collembola from the orders Entomobryomorpha and Poduromorpha were recognised (Fig. 4). Overall, only three OTUs and nine families were shared between both samples. OTU identification was much poorer when using only the NCBI nt data base (i.e. not adding *vouchers* sequences) in MEGAN and resulted in ambiguous assignment of entities identified as Acari and Collembola using NCBI nt + *vouchers* as reference data base (see Figs S1 and S2 for details).

Table 1. Number of specimens, OTUs delimited (*Combined full-length OTUs* data set) and recovery of OTUs by the metabarcoding and metagenomic libraries (mitochondrial reads and contigs) for the Cádiz (CAD) and Córdoba (COR) samples

Sample	Taxa	Specimens	OTUs	Metabarcoding recovery	MMG reads recovery	MMG contigs recovery
CAD	Acari	384	48	48	41	24
	Collembola	112	17	17	10	6
	Other arthropods	–	29	29	16	8
	Non-arthropods	–	14	14	2	0
COR	Acari	200	29	27	24	19
	Collembola	104	17	17	11	7
	Other arthropods	–	19	19	10	8
	Non-arthropods	–	9	9	2	0

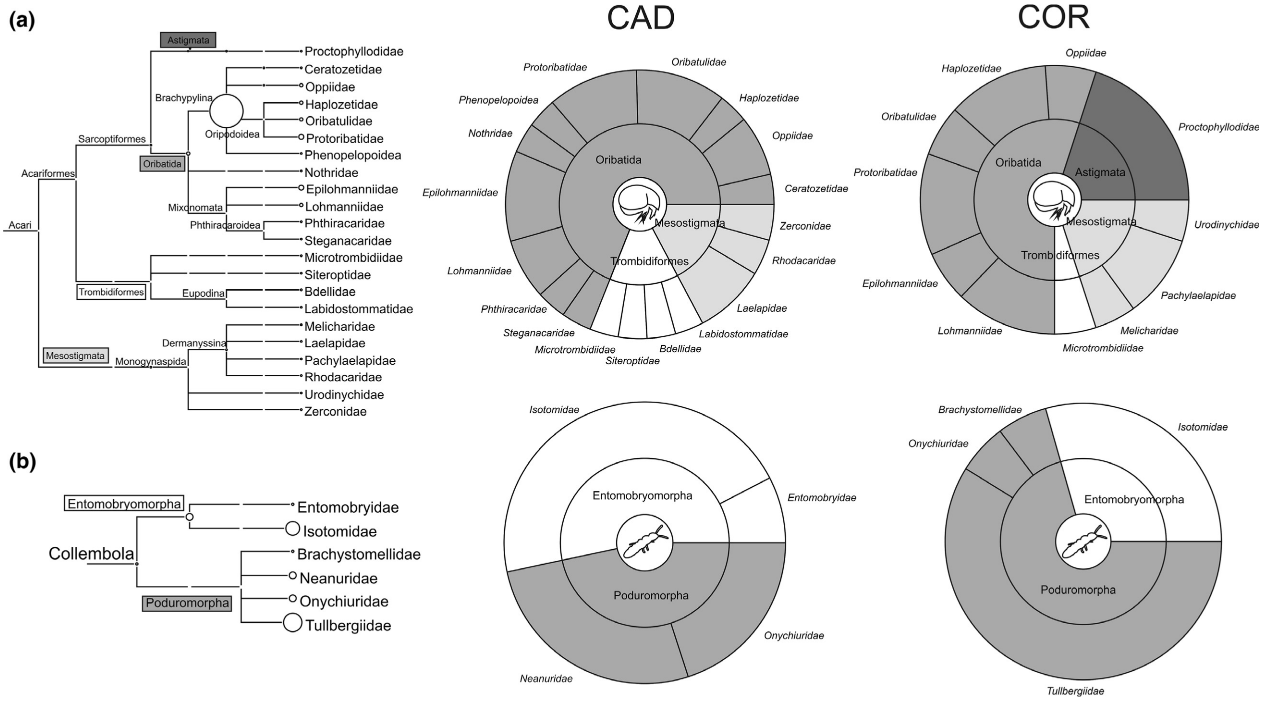

Fig. 4. Composition of Acari (a) and Collembola (b) families in the studied deep soil samples. The size of circles in the taxonomic trees is proportional to the number of OTUs assigned to each level by the lowest common ancestor algorithm on the *Combined full-length OTUs* data set against the overall NCBI nt data base and *vouchers* sequences. Taxonomic ranks are based on the NCBI Taxonomy data base.

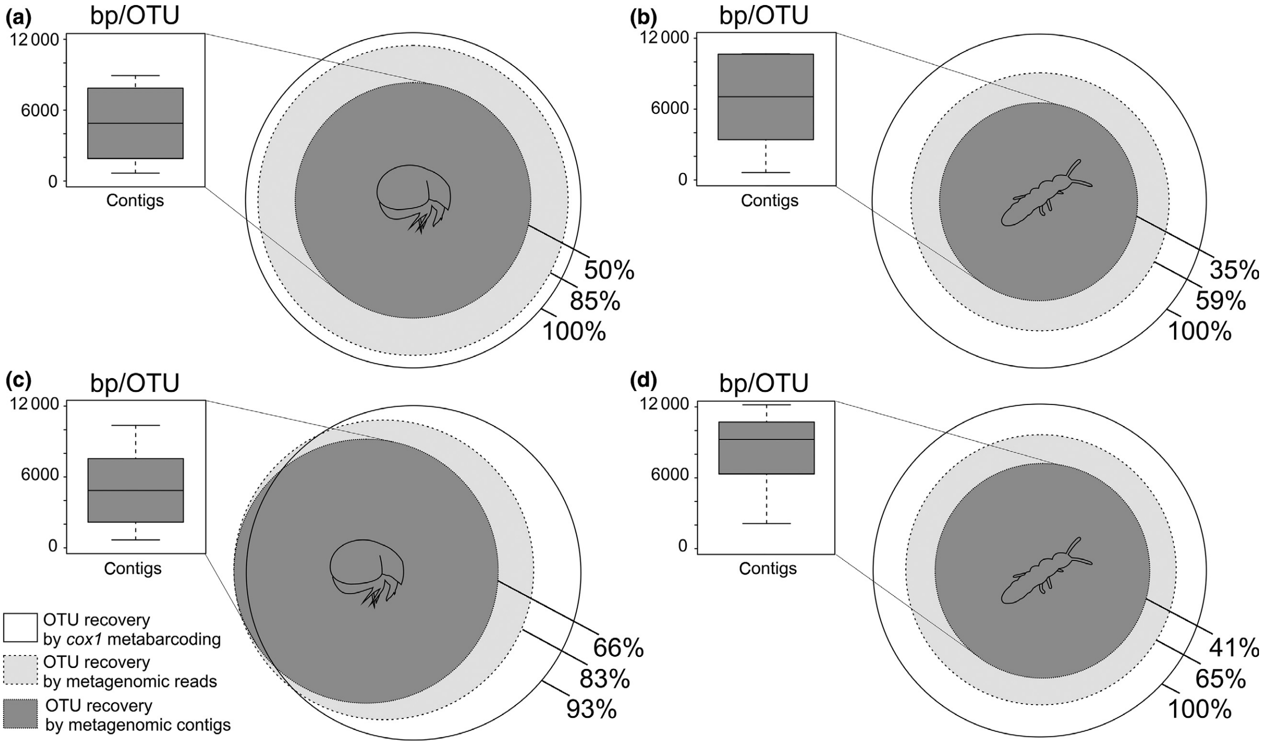

Fig. 5. OTU recovery in the *metabarcoding-cox1* (outer circle), the metagenomic reads (pale grey circle) and the assembled contigs from the *metagenomic-cox1* data set (dark grey circle) for Acari (a, c) and Collembola (b, d) from CAD and COR samples, respectively. The small panels to the left of each circle indicate the length of mitochondrial contigs (mean and lower and upper quartiles).

The recovery of entities from metabarcoding and metagenomic libraries was largely congruent, but the latter recovered fewer OTUs and only a few of the non-arthropod OTUs from the *Combined full-length OTUs* (Table 1, Fig. 5). Nearly all of the OTUs identified as Acari and Collembola in the *Combined full-length OTUs* data set were present in the *metabarcode-cox1* data set, while around 73% were detected in the reads from the metagenomic libraries and 48% were detected in the assembled contigs of the *metagenomic-cox1* data set. The recovery of OTUs from the metagenomic libraries was higher for the Acari (around 84% detected in the reads and 58% in the contigs) than for the Collembola (62% and 38%, respectively; Table 1, Fig. 5). Mitogenome assemblies corresponding to the OTUs had an average length of >6000 bp, including almost complete mitochondrial genomes for representatives of the main families of endogean Acari (e.g. Oribatulidae, Pachylaelapidae, Oppiidae, Epilohmanniidae) and Collembola (e.g. Isotomidae, Brachystomellidae, Onychiuridae; Fig. 5, Appendix S2).

Finally, evaluating the recovery of *vouchers* in the libraries, almost all barcode Sanger sequences from *vouchers* (75/79) matched one of the OTUs from the *Combined full-length OTUs* data set with a similarity ≥97%. When considering the recovery of these *voucher* sequences in the metabarcoding and metagenomic libraries, of the 75 OTUs matched to the *vouchers* sequences, all (100%) were found in the *metabarcode-cox1* data set, 95% in the metagenomic reads and 80% in the metagenomic contigs (Appendix S2). The four *vouchers* barcode sequences not recovered were two Oribatida from CAD (families Lohmanniidae and Mycobatidae) and two other Oribatida from COR (Galumnoidea and an unidentified Brachypylina specimen; Table S2). These sequences may have been singleton representatives of their species without further individuals in the bulk sample after removing the vouchers. In some cases, sequences from multiple *vouchers* matched a single OTU. These cases were in agreement with the morphological identifications that also indicated the same species but usually in different developmental instars (Table S2).

We repeated the OTU identification and analysis of OTU recovery with metabarcoding, shotgun read matching and read assembly, using the shorter *Combined OTUs* sequences, that is the uniform-length 270-bp fragment of the 3′ end of *cox1*. Results from these analyses were almost the same as those obtained with the *Combined full-length OTUs* data set (see Table S4, Figs S1, S2, S3, Appendix S2), demonstrating the power of the *cox1* marker even based on a fraction of the full barcode sequence.

Discussion

THE FBF PROTOCOL FOR THE HTS OF SOIL ARTHROPOD MESOFAUNA

This study addresses two major challenges for characterising the unknown diversity of soil invertebrates: (i) the isolation of minute specimens and clean DNA from large volumes of soil and (ii) the subsequent community characterisation of taxonomically intractable specimens through HTS methodologies. Soil flotation (also known as soil washing) is a classical method for the extraction of soil arthropod mesofauna initially proposed by Normand (1909), which takes advantage of the differential sedimentation of minerals vs. organic material and the high survival of soil mesofauna to short-term inundation. Different studies have shown the effectiveness of soil flotation for different groups of arthropod macrofauna (e.g. Lemagnen 2009; Sandler *et al.* 2010; Andújar *et al.* 2015). We found a broad range of Acari and Collembola families typically associated with deep soil, suggesting a largely unbiased recovery of these groups and pointing to a high phylogenetic diversity in the mesofauna of the poorly studied deep soil layers (but see Andújar *et al.* 2015). The total number of specimens recovered here appears low, but the density of arthropods in soil layers below 3–5 cm is generally much lower than in the leaf litter layers (e.g. Ponge 2000). When applied to the top 0–5 cm of soil from the same sites, the FBF protocol resulted in an up to 25× greater number of specimens (unpublished data). The FBF draws on existing methods modified here to minimise mortality while maximising specimen capture, for example recovering both suspended and floating material or facilitating vertical migration in the modified Berleses. The approach is unlikely to introduce any additional biases of species recovery beyond those obtained with simple Berlese protocols (without floating) but permits the extraction of 'clean' specimen bulks from a much larger starting volume of soil.

Conventional DNA extractions from the soil matrix generally are limited to 200 g per sample and, despite being suitable for assessing microbial diversity, this amount is not sufficient for the spatial dimension of soil arthropod communities. FBF allows the processing of 20 L or more of soil and thus the magnitude required for a meaningful capture of soil arthropod communities. In addition, the physical separation of the specimens from the soil effectively removes the high bacterial component, which is co-amplified by universal arthropod primers and has caused some studies to abandon *cox1* metabarcoding altogether (see Yang *et al.* 2013). The protocol also removes possible inhibitors of DNA polymerases present in humus that interfere with PCR amplification and library construction. Standardisation of the FBF approach, as those already established for other fauna extraction methods (e.g. ISO 23611 part 1–6), would provide a quantitative procedure for measuring soil arthropod communities using HTS-based biomonitoring.

COX1 METABARCODING AND MITOCONDRIAL METAGENOMICS OF SOIL ARTHROPOD MESOFAUNA

Metabarcoding and metagenomics protocols for soil arthropod mesofauna require further optimisation and adaption to the new sequencing platforms. Previous studies have emphasised the importance of using *cox1* as the target gene for metabarcoding of eukaryotes (Tang *et al.* 2012), particularly in the case of soil arthropods (Ramirez-Gonzalez *et al.* 2013). We show that the two sets of primers and PCR conditions on the bulk samples resulted in the consistent amplification across divergent lineages of Acari and Collembola. In addition, the *cox1* metabarcoding data closely match the *voucher*s set, showing no apparent taxonomic biases in the amplification. Several OTUs from other soil-inhabiting arthropod orders, including Coleoptera, Isoptera or Diptera, were also detected that likely result from the amplification of additional specimens or tissue in the initial bulk sample. Thus, the primers and PCR conditions used for *cox1* metabarcoding seem to reveal a largely complete arthropod fauna inhabiting soils.

A key step in our *cox1* metabarcoding approach is the use of two nested PCRs for library tagging, which permits to processing high numbers of samples in a single flow cell as potentially required for the study of biodiversity patterns and long-term monitoring of soil arthropods. The Illumina Nextera XT Index Kit is widely used for amplification of the 16S rRNA gene in bacterial metabarcoding and relies on the secondary PCR protocol with Illumina sequencing primers that also include dual indexes by which each of the samples can be recognised. Using the current kit, up to 96 amplicon tagging combinations can be created and sequenced together. This system was applied here to *cox1* barcodes for the first time and constitutes an accurate and simple way to process numerous community samples simultaneously. The tagging with the secondary PCR reduces the problems of differential primer efficiency and tag jumping of conventional single-PCR approaches for metabarcoding (e.g. Schnell, Bohmann & Gilbert 2015). A further advantage is that at current prices, the approach is about an order of magnitude cheaper per sample than the alternative protocols of using ligation for adding the library tag-sequencing primers.

In addition, we devised a protocol that uses the MiSeq Illumina output of maximally 300 bp to characterise the ~650-bp *cox1* fragment. The high-quality portion of 270 bp of the R1 reads were used for OTU delimitation, to which the R2 reads were linked via the paired-end information and the mitochondrial genomes from shotgun sequence assemblies to obtain an almost full-length barcode for each OTU. The OTUs from both *Combined OTUs* (R1, 270 bp) and *Combined full-length OTUs* (R1 + R2, up to 650 bp) data sets resulted in almost the same biodiversity profile. Likewise, the delimitation of OTUs performed on the metabarcoding R1 reads under various parameter settings for read quality filtering, clustering algorithms and similarity thresholds showed a very consistent result across the three methods and largely constant number of OTUs below a similarity cut-off of 97·5% and less. Thus, the widely used 97% threshold was supported here as a universal cut-off for OTU delimitation in *cox1* metabarcoding, although the effects of parameter settings and clustering method sometimes have an impact on diversity estimations (e.g. Flynn *et al.* 2015). Our results might suggest that the metabarcoding methodology is robust, even based on the single read length, and efforts for longer fragments may be not needed, even though recent approaches for full-length sequencing of *cox1* barcodes using internal primers are becoming available (e.g. Shokralla *et al.* 2015). Longer barcodes could improve phylogenetic placement and species-level identifications, but full-length mitochondrial genomes, for example from MMG, are more desirable for this purpose.

In parallel to metabarcoding, we therefore conducted mitogenome assembly in Acari and Collembola from bulk samples and demonstrated the potential of the MMG approach for studying the extremely diverse soil mesofauna and possibly any complex metazoan community. Approximately 4% of all reads in the shotgun sequencing from bulk extraction were identified as mitochondrial, which was much higher than the proportion found in existing studies on beetles (e.g. Andújar *et al.* 2015; Crampton-Platt *et al.* 2015). We obtained contigs of >8000 bp for around 30 species each of Acari and Collembola from a range of major lineages and thus greatly increased the available mitogenomes in the NCBI data base currently limited to one mitogenome for oribatid Acari and 10 for all of Collembola. The success of mitogenome assembly is greatest if DNA concentrations of individual DNA extractions in the pool are equimolar (e.g. Gillett *et al.* 2014), and assembly success was shown to be greatly reduced by variation in DNA concentrations and intraspecific genetic variation in the bulk samples (Gómez-Rodríguez *et al.* 2015). Working with complex mixtures of soil arthropod presumably created similar problems, but we still obtained a considerable number of long mitogenomes from these bulk samples. The results could likely be improved further by higher sequencing depth beyond the 25% of a MiSeq flow cell allocated per sample and in particular could recover the low biomass components of communities (see below).

'ILLUMINA-TING' SOIL ARTHROPOD MESOFAUNA

The combined metabarcoding and metagenomic libraries included more than 100 species of Acari and Collembola from only two soil samples. Maximum diversity of soil arthropods is thought to be located in superficial layers (epigeal and hemiedaphic zones; Burges & Raw 1967), which has been the focus of most studies of soil mesofauna. Here, taking advantage of the FBF protocol and sampling at a depth of 5–40 cm, we show a broad diversity of families present in the euedaphic zone (Fig. 4). The two samples used were from a similar habitat, albeit 100 km apart, and give a glimpse of the high alpha and beta and phylo-beta diversity of Acari and Collembola present in deep soil layers, whose turnover rate possibly is greater than in superficial layers (Andújar *et al.* 2015) where the levels of diversity are slightly better known (e.g. Erdmann, Scheu & Maraun 2012; Shaw, Faria & Emerson 2013). Only nine families (of 28 families in total) were common to both samples and turnover was even higher at the species level with only three of 108 OTUs in common (Fig. 4). The application of the FBF protocol to further sites and both deep and superficial layers will produce a comparative framework for the vertical profiles and differential processes driving these soil arthropod communities.

A recurrent problem when working with soil mesofauna is the 'molecular deficit' in reference data bases, that is the limited molecular resources available for highly diverse groups, which is affecting the identification of entities resulting from HTS approaches (Bik *et al.* 2012; Zepeda Mendoza, Sicheritz-Ponten & Gilbert 2015). Here, we generated *vouchers* as additional references sequenced for the entire *cox1* barcode with conventional methods, providing the link to morphological identifications (see Figs S1 and S2 for details). These *vouchers* can be seen as type specimens at the sequence level, in analogy to conventional types with full morphological descriptions. They are the starting point for the development of curated phylogenetic reference data bases for Acari and Collembola. MMG provides this information through the *de novo* assembly of complete and partial mitogenomes directly from the mixtures and thus will be quick to solve the lack of molecular resources for soil mesofauna. The link to the mitogenomes assembled from bulk samples can provide the evolutionary context for each of the OTUs that become terminals of a phylogenetic tree onto which morphological and ecological traits are mapped for comparative biology.

Metabarcoding and metagenomics thus work together for the characterisation of soil communities and their component species. Metabarcoding can be conducted *de novo*, as the read reliability seems to be sufficiently high for biodiversity estimation even to low taxonomical levels. This was also evident from the fact that most *vouchers* sequences are close matches to the OTU clusters from metabarcoding but also from mitogenome assemblies. Yet, using metabarcodes without careful filtering or mapping to established reference sequences (PCR-based *cox1* barcodes or full mitochondrial genomes) holds the risk of contamination and/or misidentification of the defined entities. Metabarcoding produced the largest number of species from our samples, of which only 75% or 85% could be fully validated by the metagenomic read mapping for the Collembola and Acari, respectively. The entities only represented by metabarcoding may be species that contribute very low biomass to the pool, but they also may be spurious, for example resulting from PCR artefacts or nuclear mitochondrial pseudogenes. Similarly, almost all non-Acari and Collembola OTUs were obtained with metabarcoding only, which suggests that they are low biomass members of the community, including remaining specimen traces or free DNA, diet items or possible internal parasites and symbionts of the target specimens, while laboratory contaminants also cannot be excluded. The metagenomic approach (both read mapping and particularly contig assembly) is more resilient to the recovery of these non-targeted groups and/or contamination. It is also worth noting that all *vouchers* sequences matching the OTUs correspond to entities recovered by all three sequencing approaches, including metagenomic assembly. This may again indicate that some entities detected with metabarcoding alone are rare or small-bodied components of the specimen mix, but the great majority is detected consistently across approaches.

Conclusions

In a recent review, Bardgett & van der Putten (2014) identified the main avenues for further research in our understanding of below-ground biodiversity and ecosystem functioning. They emphasised the need to advance understanding in (i) the spatial patterns of soil biodiversity, (ii) the temporal patterns and community dynamics, (iii) the linkages between soil biodiversity and ecosystem processes and (iv) the understanding of eco-evolutionary dynamics and environmental change. Paradoxically, the greatest knowledge gaps in these topics apply to soil mesofauna, rather than the even more complex soil microbial diversity, mainly because of the problems associated with their characterisation using conventional morphological and molecular approaches and the slow adoption of HTS to the study of mesofauna. The methodological development of HTS approaches using both *cox1* metabarcoding and MMG, together with the novel application of the FBF, can serve as a catalyst for a revolution in soil biodiversity research. First, previous limitations related to the time and resources needed for the characterisation of global and regional diversity patterns of soil mesofauna can be easily overcome (including the number of samples and sample size), and this work can now be extended to the study of phylogenetic community ecology by the incorporation of the phylogenetic information associated with the *cox1* gene via mitogenomics. Secondly, these approaches facilitate biomonitoring programmes based on complete arthropod communities, to study responses to environmental change. The taxonomic characterisation of community composition is achieved by *cox1* metabarcoding, while MMG in addition provides information on the relative abundances (biomass) from counts of shotgun reads for the study of community dynamics. Finally, the MMG approach also contributes the phylogenetic guide trees for evolutionary studies that link ecosystem processes and functional diversity. The successful application of *cox1* metabarcoding and MMG to arthropod bulk communities demonstrates a powerful approach to address new questions and revisit previous research on the diversity and distribution of animals in the soil, and so to understand a large and important portion of global biodiversity under our feet.

Acknowledgements

This research was funded by the Leverhulme Trust (grant F/00696/P to APV) and the NHM Biodiversity Initiative. PA was supported by a postdoctoral grant from the Royal Society (Newton International Program, UK). CA received additional support of a Synthesys grant (GB-TAF- 2966) and a postdoctoral NERC grant (NE/L013134/1). We thank to the regional government of Andalucía (Spain) for the collecting permissions and to the staff of the Sierra de Grazalema Natural Park and Sierras Subbéticas Natural Park for all the support received during field sampling. Thanks are due to Alex Aitken, Stephen Russell, Andie Hall, Peter Foster (all NHM) and Shilo Dickens (University of Cambridge) for their technical assistance and to Jesús Arribas who assisted with field sampling and contributed drawings included in the figures. Thanks to Benjamin Linard and Kirsten Miller (NHM), Brent Emerson (CSIC), the Associate Editor and an anonymous referee for their useful comments on the manuscript.

Data accessibility

GenBank Accessions numbers are given in Table S2; primers and PCR conditions in Table S1 and Data S2; *Combined OTUs*, *Combined full-length OTUs* and *metagenomic-cox1* data sets deposited in the Dryad repository: http://datadryad.org/resource/doi:10.5061/dryad.5t5s1; all supplementary tables and figures cited in the main text have been uploaded as online Supporting Information.

References

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

Andújar, C., Arribas, P., Ruzicka, F., Platt, A.C., Timmermans, M.J.T.N. & Vogler, A.P. (2015) Phylogenetic community ecology of soil biodiversity using mitochondrial metagenomics. *Molecular Ecology*, **24**, 3603–3617.

Arribas, P., Andújar, C., Hopkins, K., Shepherd, M. & Vogler, A.P. (2016) Data from: Metabarcoding and mitochondrial metagenomics of endogean arthropods to unveil the mesofauna of the soil. *Methods in Ecology and Evolution*, http://dx.doi.org/105061/dryad.5t5s1.

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

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

Bik, H.M., Porazinska, D.L., Creer, S., Caporaso, J.G., Knight, R. & Thomas, W.K. (2012) Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology & Evolution*, **27**, 233–243.

Bininda-Emonds, O.R.P. (2005) transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. *BMC Bioinformatics*, **6**, 156.

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

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

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

Crampton-Platt, A., Timmermans, M.J.T.N., Gimmel, M.L., Kutty, S.N., Cockerill, T.D., Vun Khen, C. & Vogler, A.P. (2015) Soup to tree: the phylogeny of beetles inferred by mitochondrial metagenomics of a Bornean rainforest sample. *Molecular Biology and Evolution*, **32**, 2302–2316.

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

Decaëns, T., Jiménez, J.J., Gioia, C., Measey, G.J. & Lavelle, P. (2006) The values of soil animals for conservation biology. *European Journal of Soil Biology*, **42**, S23–S38.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics (Oxford, England)*, **26**, 2460–2461.

Edgar, R.C. (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, **10**, 996–998.

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.

Fierer, N. & Jackson, R.B. (2006) The diversity and biogeography of soil bacterial communities. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 626–631.

Fierer, N., Strickland, M.S., Liptzin, D., Bradford, M.A. & Cleveland, C.C. (2009) Global patterns in belowground communities. *Ecology Letters*, **12**, 1238–1249.

Flynn, J.M., Brown, E.A., Chain, F.J.J., MacIsaac, H.J. & Cristescu, M.E. (2015) Toward accurate molecular identification of species in complex environmental samples: testing the performance of sequence filtering and clustering methods. *Ecology and Evolution*, **5**, 2252–2266.

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

Fonseca, V.G., Carvalho, G.R., Sung, W., Johnson, H.F., Power, D.M., Neill, S.P. *et al.* (2010) Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nature Communications*, **1**, 98.

Gillett, C.P.D.T., Crampton-Platt, A., Timmermans, M.J.T.N., Jordal, B.H., Emerson, B.C. & Vogler, A.P. (2014) Bulk *de novo* mitogenome assembly from pooled total DNA elucidates the phylogeny of weevils (coleoptera: curculionoidea). *Molecular Biology and Evolution*, **31**, 2223–2237.

Gómez-Rodríguez, C., Crampton-Platt, A., Timmermans, M.J.T.N., Baselga, A. & Vogler, A.P. (2015) Validating the power of mitochondrial metagenomics for community ecology and phylogenetics of complex assemblages. *Methods in Ecology and Evolution*, **6**, 883–894.

Griffiths, B.S., Donn, S., Neilson, R. & Daniell, T.J. (2006) Molecular sequencing and morphological analysis of a nematode community. *Applied Soil Ecology*, **32**, 325–337.

Hao, X., Jiang, R. & Chen, T. (2011) Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics*, **27**, 611–618.

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

Huson, D., Mitra, S., Ruscheweyh, H., Weber, N. & Schuster, S.C. (2011) Integrative analysis of environmental sequences using MEGAN4. *Genome Research*, **21**, 1552–1560.

Ji, Y., Ashton, L., Pedley, S.M., Edwards, D.P., Tang, Y., Nakamura, A. *et al.* (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, **16**, 1245–1257.

Katoh, K., Misawa, K., Kuma, K. & Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, **30**, 3059–3066.

Lemagnen, R. (2009) Techniques: tamisages, lavages et autres berlèses. *Invertébrés Armoricains*, **1**, 19–22.

Lohse, M., Bolger, A. & Nagel, A. (2012) RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research*, **40**, W622–W627.

Mahé, F., Rognes, T., Quince, C., de Vargas, C. & Dunthorn, M. (2014) Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ*, **2**, e593.

Normand, D.H. (1909) Description d'un nouveau procedé de capture des Coléoptères hypogés. *Bulletin de la Société entomologique de France*, **19**, 329.

Orgiazzi, A., Dunbar, M.B., Panagos, P., de Groot, G.A. & Lemanceau, P. (2015) Soil biodiversity and DNA barcodes: opportunities and challenges. *Soil Biology and Biochemistry*, **80**, 244–250.

Ponge, J. (2000) Vertical distribution of Collembola (Hexapoda) and their food resources in organic horizons of beech forests. *Biology and Fertility of Soils*, **32**, 508–522.

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

Ramirez-Gonzalez, R., Yu, D.W., Bruce, C., Heavens, D., Caccamo, M. & Emerson, B.C. (2013) PyroClean: denoising pyrosequences from protein-coding amplicons for the recovery of interspecific and intraspecific genetic variation. *PLoS One*, **8**, e57615.

Sandler, R., Falco, L., Ciocco, C.Di., de Luca, R. & Coviella, C.E. (2010) Eficiencia del embudo Berlese-Tullgren para extracción de artrópodos edáficos en suelos argiudoles típicos de la provincia de Buenos Aires. *CI. Suelo (Argentina)*, **28**, 1–7.

Schnell, I.B., Bohmann, K. & Gilbert, M.T.P. (2015) Tag jumps illuminated – reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources*, **15**, 883–894.

Shaw, P., Faria, C. & Emerson, B. (2013) Updating taxonomic biogeography in the light of new methods–examples from Collembola. *Soil Organisms*, **85**, 161–170.

Shokralla, S., Porter, T.M., Gibson, J.F., Dobosz, R., Janzen, D.H., Hallwachs, W., Golding, G.B. & Hajibabaei, M. (2015) Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific Reports*, **5**, 9687.

Tang, C.Q., Leasi, F., Obertegger, U., Kieneke, A., Barraclough, T.G. & Fontaneto, D. (2012) The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 16208–16212.

Tang, M., Hardman, C.J., Ji, Y., Meng, G., Liu, S., Tan, M. *et al.* (2015) High-throughput monitoring of wild bee diversity and abundance via mitogenomics (M Gilbert, Ed.). *Methods in Ecology and Evolution*, **6**, 1034–1043.

Wu, T., Ayres, E., Bardgett, R., Wall, D.H. & Garey, J.R. (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.

Yang, C., Ji, Y., Wang, X., Yang, C. & Yu, D.W. (2013) Testing three pipelines for 18S rDNA-based metabarcoding of soil faunal diversity. *Science China. Life Sciences*, **56**, 73–81.

Yang, C., Wang, X., Miller, J.A., de Blécourt, M., Ji, Y., Yang, C., Harrison, R.D. & Yu, D.W. (2014) Using metabarcoding to ask if easily collected soil and leaf-litter samples can be used as a general biodiversity indicator. *Ecological Indicators*, **46**, 379–389.

Zepeda Mendoza, M.L., Sicheritz-Ponten, T. & Gilbert, M.T.P. (2015) Environmental genes and genomes: understanding the differences and challenges in the approaches and software for their analyses. *Briefings in Bioinformatics*, 1–14.

Received 26 December 2015; accepted 22 February 2016
Handling Editor: Douglas Yu

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Appendix S1.

Data S1. Flotation-Berlese-flotation protocol (FBF).

Data S2. *cox1* metabarcoding PCR reagents and conditions.

Data S3. Mitochondrial metagenomics pipeline (MMG).

Table S1. Primers and PCR conditions used.

Table S2. *Vouchers* specimens Sanger sequenced.

Table S3. Number of delimited OTUs from the different clustering algorithms, pre-clustering filtering options and similarity thresholds applied to the metabarcoding libraries.

Table S4. Number of specimens, OTUs delimited and recovery of OTUs by the metabarcoding and metagenomic libraries (*Combined OTUs* data set).

Fig. S1. Taxonomic assignment of the OTUs (*Combined full-length OTUs* data set).

Fig. S2. Taxonomic assignment of the OTUs (*Combined OTUs* data set).

Fig. S3. Composition of Acari and Collembola families in the studied deep soil samples.

Appendix S2.

Table S5. Combined_full_length_OTUs_data.

Table S6. Combined_OTUs_data set.