communications biology

ARTICLE

https://doi.org/10.1038/s42003-023-05621-4 OPEN

The MetaInvert soil invertebrate genome resource provides insights into below-ground biodiversity and evolution

Gemma Collins[1,2], Clément Schneider[2,3], Ljudevit Luka Boštjančić[1,4,5], Ulrich Burkhardt[6], Axel Christian[3], Peter Decker[3], Ingo Ebersberger[1,2,7], Karin Hohberg[3], Odile Lecompte[4], Dominik Merges[8], Hannah Muelbaier[7], Juliane Romahn[1,2], Jörg Römbke[9], Christelle Rutz[4], Rüdiger Schmelz[10], Alexandra Schmidt[1,11], Kathrin Theissinger[1,2,5], Robert Veres[1,12], Ricarda Lehmitz[3], Markus Pfenninger[1,2,13] & Miklós Bálint[1,2,14]✉

Soil invertebrates are among the least understood metazoans on Earth. Thus far, the lack of taxonomically broad and dense genomic resources has made it hard to thoroughly investigate their evolution and ecology. With MetaInvert we provide draft genome assemblies for 232 soil invertebrate species, representing 14 common groups and 94 families. We show that this data substantially extends the taxonomic scope of DNA- or RNA-based taxonomic identification. Moreover, we confirm that theories of genome evolution cannot be generalised across evolutionarily distinct invertebrate groups. The soil invertebrate genomes presented here will support the management of soil biodiversity through molecular monitoring of community composition and function, and the discovery of evolutionary adaptations to the challenges of soil conditions.

[1] Senckenberg Biodiversity and Climate Research Centre, Frankfurt am Main, Germany. [2] LOEWE Centre for Translational Biodiversity Genomics, Frankfurt am Main, Germany. [3] Soil Zoology, Senckenberg Museum of Natural History, Görlitz, Germany. [4] Department of Computer Science, ICube, UMR 7357, University of Strasbourg, CNRS, Centre de Recherche en Biomédecine de Strasbourg, Strasbourg, France. [5] Department of Molecular Ecology, Institute for Environmental Sciences, Rhineland-Palatinate Technical University Kaiserslautern Landau, Landau, Germany. [6] Soil Organism Research, Görlitz, Germany. [7] Institute of Cell Biology and Neuroscience, Goethe University, Frankfurt am Main, Germany. [8] Department of Forest Mycology and Plant Pathology, Swedish University of Agricultural Sciences, Uppsala, Sweden. [9] ECT Oekotoxikologie GmbH, Flörsheim, Germany. [10] Freelance Biologist, A Coruña, Spain. [11] Limnological Institute, University of Konstanz, Konstanz, Germany. [12] Institute of Biology and Geology, Babeș-Bolyai University, Cluj-Napoca, Romania. [13] Johannes Gutenberg University, Mainz, Germany. [14] Department of Insect Biotechnology, Justus-Liebig University, Gießen, Germany. ✉email: miklos.balint@senckenberg.de

Soils and soil biodiversity are becoming increasingly valued and protected at the policy level[1]. Soil invertebrates are major components of soil biodiversity, and their activity is important for almost all soil ecosystem services[2]. For example, soil invertebrates are responsible for up to 50% of the litter decomposition[3]. They contribute to functional services crucial to humans, such as nutrient cycling, water storage and support above-ground food production through the integration of nutrients in food webs[4–6]. Furthermore, soil invertebrates play major roles in regulating microbial activity along the plant-soil continuum[7]. Consistent with their importance in soil ecosystems, they are actively promoted in soil biodiversity conservation frameworks[8].

However, soil invertebrates are inherently difficult to study morphologically due to their incredible diversity, huge abundances, and small body size with microscopic morphological details. Though generally tiny, they show a ~100-fold variation in body weight, which ranges from nanograms to grams[9]. There are potentially hundreds of thousands of undescribed species globally[10]. Moreover, taxonomic expertise is declining[11] and this is particularly problematic for groups where experts have always been rare.

DNA- and RNA-based methods are long promoted to support traditional taxonomy and ecological studies in difficult organism groups. Shotgun metagenomics randomly sequences DNA fragments from a sample, instead of relying on PCR-amplified taxonomic marker genes. Metagenomics is an increasingly feasible approach to record the presence of higher eukaryotes in a diverse range of samples[12–14]. Since metagenomics can utilise all genomic information for taxonomic identification, it has improved sensitivity and specificity compared to metabarcoding[15], and it promises superior quantification of species' biomass[16]. Metatranscriptomics in turn records genes which are actively transcribed into RNA and thus drive ongoing biological processes[17], informing about the metabolic activity of soil community members[18], and functional changes in these communities[17].

Comprehensive genome collections are the backbone for metagenomics and metatranscriptomics. If genome databases are available, shotgun metagenomics and metatranscriptomics have shown to provide unprecedented insights[17,19], e.g., into vegetation change over glacial cycles[15], historic population genomic processes[20,21], and kingdom-spanning processes of ecosystem functioning[22]. Large genome sequencing initiatives like the Earth Biogenome Project[23] will provide this data ultimately, but progress so far mainly focused on large, prominent organisms, such as mammals[24], birds[25], insects[26] and plants[15]. In addition to serving taxonomic identification, broad (many distinct groups) and dense (many species from a group) sequencing of genomes additionally allows identifying common patterns of gene evolution and test the taxonomic generality of hypotheses on genome evolution.

Results and discussion

A genome resource for soil invertebrates. Here, we have generated a large genomic resource to support insights into the structure, activity and functioning of soil invertebrate communities (Fig. 1). We had two aims. First, we wanted to provide a large number of soil invertebrate genomes to aid species identifications through metagenomics or metatranscriptomics. Second, we intended to explore patterns of genome evolution across taxa, which needs a taxonomically broad and dense sampling of genomes. We sequenced and assembled the genomes of 232 species, representing 14 common soil invertebrate groups (hereafter referred to as "groups") encompassing 94 families, most of which were lacking whole-genome data thus far (Fig. 2, Table 1), including Collembola ($n = 87$ species), Oribatida ($n = 62$), two

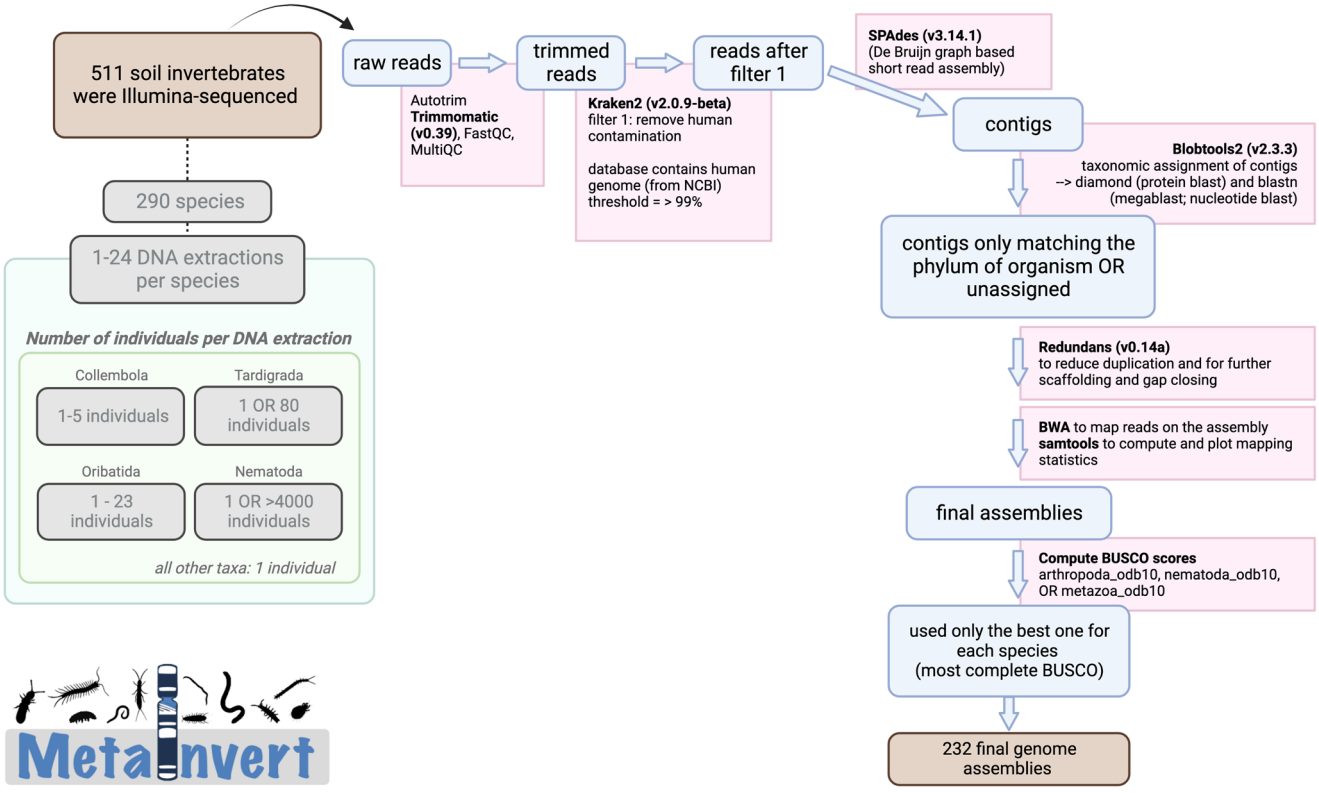

Fig. 1 Overview of the bioinformatic pipeline for genome assembly and quality control. The genome assembly pipeline consists of a read quality filtering step, short read assembly and several steps for removing non-target DNA reads, co-sequenced along the genomes of the targeted species. The MetaInvert logo was created by the first author. Animal silhouettes originate from phylopic.org, and they can be reused under Creative Common licences.

classes of Myriapoda ($n = 23$ Diplopoda; $n = 19$ Chilopoda) and Nematoda ($n = 18$). Genome completeness estimated with benchmarking universal single copy orthologs (BUSCO)[27] was 59.78% on average (median: 69.2%), with an average contig N50 of 6080 bases (median 4039), and with an average L50 of 28,375 (median: 11503, Supplementary Data 1).

Improved taxonomic assignment of metazoan environmental sequence data. To demonstrate the relevance of this genomic resource, we first used the 232 genomes to improve the taxonomic assignment metatranscriptomic sequences generated from a 2-year sampling of soil environmental RNA (eRNA) along an elevational gradient[28]. Such assignments of soil eRNA were previously limited in scope due to a general lack of soil invertebrate genome data. Briefly, we assigned eRNA reads with bacterial, fungal, plant and soil invertebrate genomes, with and without including the MetaInvert genomes presented here. We found that about 2.45% (854,409 reads) of the classified metatranscriptomic reads (40,265,768) could be assigned to soil invertebrates, in comparison to bacteria (77,1%, 31,063,088), fungi (20.1%, 8,078,679), and plants (0.33%, 134,852)[29]. Previous metatranscriptomic studies reported a similar microbial eukaryote to bacteria ratio[29,30]. The inclusion of the MetaInvert genomes significantly increased reads assigned to soil invertebrates (Kruskal-Wallis $X^2 = 9.14$, $df = 1$, $p = 0.002$, Fig. 3a). We recorded 11 soil invertebrate classes (Fig. 3b), of which the most abundant were nematodes of the class Chromadorea followed by clitellates (comprising both earthworms and enchytraeids). Linear regression showed a marked dip in soil invertebrate richness along the elevation gradient (ANOVA, $F_{elevation} = 0.22$, $p_{elevation} = 0.65$, $F_{elevation\text{^}2} = 9.1$, $p_{elevation\text{^}2} = 0.02$, Fig. 3c). This is in contrast with findings of hump-shaped elevation - richness relationships in soil invertebrates[31]. The pattern observed by us might be driven by distinct vegetation covers, although the confirmation of this needs a better sampling resolution. The community composition of soil invertebrates showed no statistically significant changes along the elevation gradient (analysis of deviance of multivariate generalised linear model fits, $df = 8$, $dev = 434.60$, $p = 0.13$), marginally significant differences across habitats study years ($df = 65$, $dev = 806.03929.87$, $p = 0.085$), and statistically significant differences between the two study years ($df = 5$, $dev = 1066.09$, $p = 0.04$).

No change in community composition along the elevation gradient is consistent with observed high abundances of soil invertebrates at high altitudes[30]. Differences in vegetation are known to influence soil invertebrate community composition, although our analysis may lack power to equivocally detect these. Differences in community composition between the study years may reflect year-specific environmental differences. However, we caution not to over interpret these results. The power of an analysis of drivers of community composition and richness on this gradient should be increased with more extensive sampling. The analyses nonetheless demonstrate the value of a dedicated soil invertebrate genome database for the identification of shotgun-sequenced environmental nucleotide samples from soils.

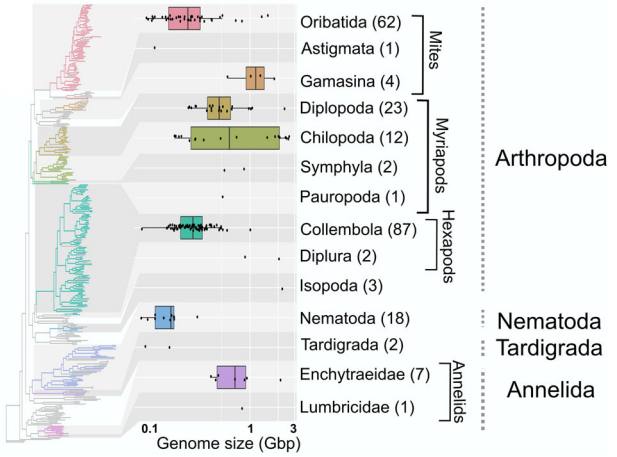

Fig. 2 Maximum likelihood phylogenetic tree of soil invertebrate genomes. The tree is based on an alignment of 141 metazoan BUSCO genes of the 232 soil invertebrates sequenced in this work (coloured branches), and 118 NCBI RefSeq (grey branches), representing four phyla. A high-resolution, annotated version of the tree is available as Supplementary Fig. 1. A more detailed tree and the alignment are available on FigShare[30]. Boxplots reflect the genome size distribution of the taxa subsumed in the corresponding clades in gigabases (Gb). Numbers of sequenced genomes with genome size estimates are indicated for each group. Genome size estimation was not possible for some of the assemblies. Genome size estimates can be found in Supplementary Data 1. Center line: median; box limits: upper and lower quartiles; whiskers: 1.5× interquartile range; points: outliers.

Table 1 Overview of 232 soil invertebrate genome assemblies.

Phylum	Taxon group [rank]	Common name	n known species (soil or terrrestrial)	n species (published genomes)	n species (genomes contributed here)
Annelida	Lumbricidae [family]	Earthworms	7000	2	1
Annelida	Enchytraeidae [family]	Potworms	700	1	7
Nematoda	Nematoda [phylum]	Nematodes	25000	73	18
Tardigrada	Tardigrada [phylum]	Tardigrades	1150	4	2
Arthropoda	Gamasina [infraorder]	Predatory mites	40000	1	4
Arthropoda	Astigmata [suborder]	Mites [not soil]		7	1
Arthropoda	Oribatida [suborder]	Box mites		7	62
Arthropoda	Chilopoda [class]	Centipedes	3000	2	19
Arthropoda	Diplopoda [class]	Millipedes	12000	3	23
Arthropoda	Symphyla [class]	Symphylans	200	0	2
Arthropoda	Pauropoda [class]	Pauropods	800	0	1
Arthropoda	Isopoda [order]	Pill bugs	3637	5	3
Arthropoda	Diplura [order]	Diplurans	1000	2	2
Arthropoda	Collembola [class]	Springtails	8500	35	87

For each taxonomic group we also list the number of species with publicly available genome assemblies (as of June 2022).

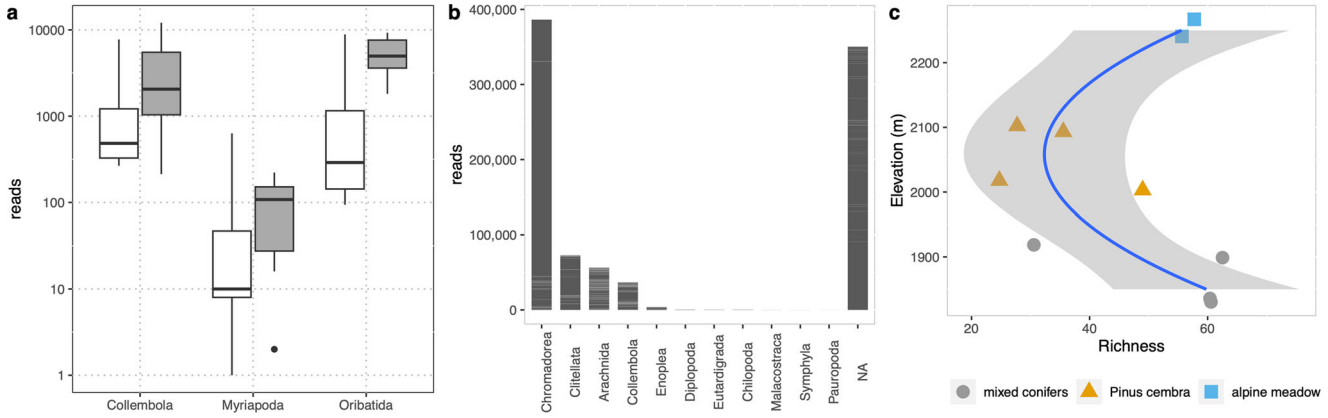

Fig. 3 Taxonomic assignments of soil metatranscriptomes using soil invertebrate genomes. a Assignment success of soil metatranscriptomic reads using genomes available in NCBI RefSeq (white) and MetaInvert genomes in addition to NCBI genomes (grey). Please note the log-scale of the y-axis (taxon observations with RefSeq genomes: Collembola $n = 290$, Myriapoda $n = 80$, Oribatida $n = 90$; independent taxon observations with RefSeq + MetaInvert genomes: Collembola $n = 810$, Myriapoda $n = 270$, Oribatida $n = 630$), center line: median; box limits: upper and lower quartiles; whiskers: 1.5× interquartile range; points: outliers; **b** reads assigned to common soil invertebrate classes, with NA marking metazoan reads not assigned to soil invertebrates at the class level; **c** soil invertebrate richness trend along an elevation gradient (grey area marks standard error of the trendline). Assignments are available as Supplementary Data 2.

Insights into genome size evolution. As a second example, we addressed hypotheses concerning genome size evolution. We estimated the genome size for 191 species using the assembly-based approach ModEst[31]. We found a 30-fold range of genome sizes across the groups (Fig. 2), from 79 Mb (the nematode *Discolaimus major*) to 2.9 Gb (the chilopod *Lithobius crassipesoides*). Nematoda and Tardigrada had typically small genomes, whereas the genomes of Enchytraeidae were remarkably larger. In addition to between-group variation, some groups also had a wide range of genome sizes among member species. For example, Chilopoda (centipedes) genomes ranged in size from 0.178 to 2.90 Gb, while Oribatida genomes ranged from 0.09 to 1.72 Gb. Repeat content and GC content also varied widely both within and between soil invertebrate groups (Supplementary Fig. 2).

Classic theory predicts that a few basic factors, in particular effective population size, should lead to causal relationships between genome properties and functional traits (Fig. 4a)[32–34]. However, recent studies have shown that taxon-specific processes might be more important for genome size than demography[35,36]. We used our taxonomically broad data set to test the classical hypothesis of a few factors generally influencing genome size evolution vs. a more lineage-specific view with a series of structural equation models (SEMs, Fig. 4). We used genomes with at least 50% BUSCO completeness and 8× mode coverage. To parametrize the SEM and connect the 143 new genome assemblies with ecological traits, we first gathered trait data from original literature. Information about habitat preferences was added from the Edaphobase data warehouse for soil biodiversity (https://portal.edaphobase.org/). We focussed on three traits: (a) body length as a proxy for body size (minimum female adult body length for nematodes, and mean adult body length for all other taxa), (b) reproduction mode, and c) the number of known habitat types where a species occurs, as a proxy of habitat generality (based on CORINE—Coordination of Information on the Environment[37]). We annotated repetitive elements with species-specific repeat libraries. We estimated effective population size (theta) directly from the genome data by making use of the genome-wide heterozygosity in the reference individual. This proxy measure of effective population size was calculated individually for each genome assembly with at least 8X coverage. Genomic and ecological traits are accessible in Supplementary Data 1.

The variables tested have complex interactions that need to be modelled in the SEMs. Effective population size should be influenced by habitat generalism, with the expectation that species able to thrive in a wide range of habitats should have larger population sizes and therefore also larger effective population sizes (N_e)[38]. N_e should be inversely related to body size, as larger populations of small-bodied organisms can be maintained by the same amount of resources in comparison to large-bodied species[32]. The reproductive mode is known to impact N_e, because the higher the degree of inbreeding, the smaller the expected N_e[39]. High N_e is frequently hypothesised to contribute to reducing repeats as evolutionary burdens from genomes, as selection is more efficient in larger populations[33,34]. Repeats are frequently considered to increase genome size[40,41]. If the repeats themselves are biassed in base composition, this should reflect in the overall GC content. Interestingly, GC content is also linked to resource availability[42], which may be linked to habitat generalism via higher metabolic flexibility[43]. Even though most of these observations originate from bacterial studies, ample evidence exists that the environment may influence base composition also in metazoans[42,44–46].

When modelling all soil invertebrate groups together, most hypothesised causal relationships were either statistically insignificant or pointed to the opposite directions than classical theory predicted (Supplementary Fig. 3). Most strikingly, high N_e size was linked to higher repeat content which in turn implies larger genome size. This suggests that efficient selection does not universally reduce the evolutionary burden of large genomes and repeat content[47]. The SEMs supported only two of the hypothesised causal relationships when these were modelled for all taxa together (Fig. 4b): a positive link between repeat content and genome size, and a negative link between repeat content and GC content. Genome size is frequently considered to be driven by repeat content[48,49], but with variation in the relationship among higher taxa of vertebrates[50]. Such variation might be due to epigenetic regulation via repetitive elements, maintenance of chromosome structure[51], and modification of gene expression and transcript diversification[52]. Higher GC content is linked to smaller genome size in many but not all eukaryotic groups[49]. This link might also originate from the expansion of repeats with low GC content.

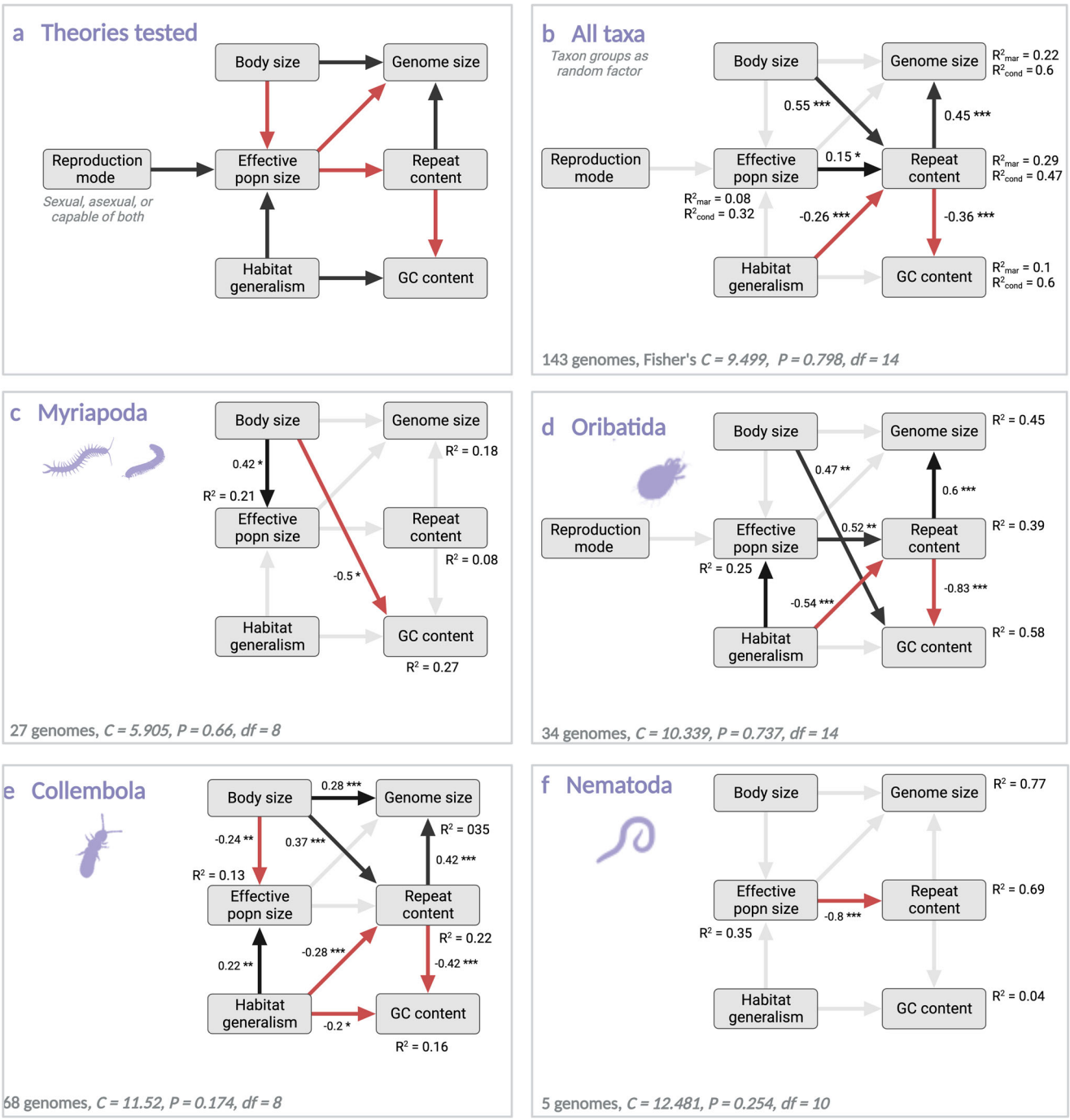

Fig. 4 Structural equation models (SEMs) of hypothesised causal relationships among genomic traits and their ecological drivers. a Initial SEM with hypothesised links; **b–f** SEMs fitted to all taxa, and to major taxonomic groups. Arrows indicate hypothesised or modelled relationships, positive (black) or negative (red). Links marked with grey arrows were not statistically significant in the SEM. Fisher's *C* evaluates conditional independence claims among nodes and indicates model fit, with *p*-values showing whether discrepancies between the model and the data are statistically significant. Degrees of freedom are marked with *df*. Values next to arrows show standardised estimates, with asterisk indicating the statistical significance of the relationship (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$). Animal silhouettes originate from phylopic.org, and they can be reused under Creative Common licences.

Our results confirm that the strength and direction of relationships among genome size, repeat content, GC content, and their ecological drivers vary among higher taxa of invertebrates (Supplementary Fig. 3). SEMs fitted separately to higher taxa (myriapods, oribatid mites, springtails, nematodes) showed marked group-specific differences in the support of causal hypotheses between genomic and ecological traits (Fig. 4c–f, Supplementary Fig. 3). The assumed positive link between body size and genome size[32] received statistical support only in Collembola, but with an opposite sign as predicted by the nucleotypic theory[32]. The effects of body size on genomes are often difficult to disentangle from other co-variables[53,54]. This indicates lineage-specific expansion or contraction of genomes, reported for diverse eukaryotes[50,55]. The expected negative relationship between N_e and repeat content[40,41] was confirmed only in nematodes. However, the relationship was positive in oribatid mites, and missing altogether in the other taxa. Habitat generalism was positively linked to effective population size only

in Collembola and oribatids, but not in myriapods and nematodes. This suggests that generalists may not be as fit as specialists in any particular habitats[56,57], and their evolution might depend on differential rates of population evolution compared to rates of environmental change[58]. Interestingly, models of oribatids and Collembola suggested that higher habitat generality might be linked to lower repeat contents. Altogether, our analysis supports a more nuanced, lineage-specific view of factors driving genome size evolution rather than the classical view of only a few general factors governing the C-value enigma.

Gene loss patterns in springtails and oribatid mites. As a third example, we explored whether shared gene loss might be related to repeated adaptations of phylogenetically distant metazoans to soil conditions. Gene loss is a key process in evolution[59,60]. Here, the dense taxon sampling for individual groups allowed to differentiate between consistent gene absence across several taxa, which likely indicates gene loss, and the sporadic absence of a gene in individual taxa, which likely represents noise introduced by assembly incompleteness. To further reduce the risk that incomplete gene annotations generate a spurious signal of gene loss, we used a targeted search for orthologs in the un-annotated genome assemblies to determine the presence/absence patterns of genes across taxa. We analysed the presence of 1482 core metazoan gene orthologs. Notably, this revealed that 50 core genes are missing in springtails ($n = 78$ species), and 97 core genes were not found in the oribatid mites ($n = 54$ species) (Fig. 5). Given the large number of investigated taxa in the two groups, it is unlikely that these genes have been accidentally missed. Instead, their absence indicates gene losses early during diversification of the respective groups, similar to what has been seen for other animals[61]. Overall, fifteen gene ontology terms were significantly enriched (testFisher < 0.05) among the genes lost involving biological processes such as tubulin metabolism and cellular and subcellular movement (Oribatida). There was a significant loss of genes involved in pyridine-containing compound metabolic processes in springtails (Fig. 5; Supplementary Data 3, 4). Pyridine-containing molecules have a considerable spectrum of antimicrobial and antiviral activities[62], and associated gene loss might be related to the gain of endogenous antibiotic synthesis ability by many springtail species[63]. We also manually screened the UniProt database (accessed on 28.6.2023) for putative gene functions associated with genes missing from Collembola and Oribatida assemblies. We aimed to identify functional or other relevant commonalities among the genes which might be missed by an algorithmic GO enrichment analysis. We could not detect patterns in gene functions. It was noteworthy that all existing annotations originated from only two species: *Drosophila melanogaster* or *Strigamia maritima*. This highlights the general difficulties with transferring annotations gained from a few model taxa to the breadth of biodiversity, with targeted annotation of specific genes being a solution.

In summary, our large collection of soil invertebrate genomes is a first major step towards a comprehensive DNA- or RNA-based identification of the entire soil biodiversity: they extend the scope of metagenomic or metatranscriptomic studies from microorganisms to metazoans. An important limitation of the study is the quality of the genomes, which precludes deeper analyses, such as structural comparisons. Genome quality is currently restrained by the qualitative and quantitative requirements of the current sequencing techniques with respect to genomic DNA. Although it is already possible to generate highly contiguous and complete genomes of soil invertebrates from single specimens[64], the minute amounts of genomic DNA (often fragmented because of field preservation) does not yet allow for the generation of better quality genomes on scale. Nonetheless, the genomes are of sufficiently high contiguity or completeness to considerably improve metagenomic and metatranscriptomic sequence assignments[65]. Further, the taxonomically broad and dense sampling of genomes provides unique insights into genome evolution, although clearly not into structural differences. Here we could show that no single theory of genome evolution fits all taxa: there are probably no simple overarching explanations for observed variations in genome properties, but interactions of multiple drivers result in divergent genome evolution patterns in different groups, reflecting their unique evolutionary history. Broad genome sampling allows for the identification of group-specific gene loss patterns, highlighting issues and future directions around the functional annotation of genomes from non-model taxa in diverse habitats. Overall, the 232 soil invertebrate genomes demonstrate the importance of genome sequencing efforts for understanding the ecology and evolution of the full scale of eukaryotic biodiversity, and project a future when maximum taxonomic and functional information will be gained from every environmental DNA or RNA fragment.

Methods

Specimen sampling and species-level identification. Specimens were collected in the field or obtained from cultures, supplemented with existing soil invertebrate specimens from Senckenberg museum collections (Supplementary Data 1). Sampling occurred between 2011 and 2020, mostly in Germany, but in some cases also from countries in Europe. Soil macrofauna was mainly collected by hand, whereas meso- and microfauna were obtained from soil samples with MacFadyen[66] or Baermann extraction[67]. DNA was extracted from over 500 single specimens, or occasionally from multiple individuals (single-species cultures of tardigrades and smaller-bodied nematodes, Supplementary Data 5). A non-destructive DNA extraction method[68] was preferred and used where possible. Otherwise, the MagAttract High Molecular Weight DNA Kit (Qiagen, Hilden, Germany) was used, mostly for cultured specimens. Voucher specimens are deposited in the Senckenberg museum collection in Görlitz.

For larger taxa such as Chilopoda, Diplopoda, Isopoda, Enchytraeidae and Lumbricidae, the species-level morphological identification was possible before DNA extraction, and only a single leg, a few body segments or musculature of mouthparts were used for DNA extraction, the rest of the body was kept as a voucher. For medium sized taxa like Acari and Collembola that normally would require clearing in lactic acid prior to species identification, the specimens were presorted on family or genus level, the whole specimens were used for non-destructive DNA extraction, and finally species-level identifications were carried out with recovered vouchers. In cases where non-destructive DNA extraction did not deliver sufficient amounts of DNA or the voucher was lost during extraction, identification was validated by aligning species markers (28 S, COI) from the whole-genome sequence data with existing species markers in GenBank or generated by us. For small, soft-skinned taxa (Nematoda, Tardigrada), where non-destructive DNA extraction is not possible, two different sources/techniques were used: (1) for most species, specimens were derived from own established cultures with known taxon and strain names, or (2) where such cultures did not exist, we freshly Baermann-extracted specimens from soil samples and identified morphospecies with at least 6 specimens at 400x magnification under an inverted microscope. We then extracted DNA from half of the specimens and prepared permanent slides of the other half (vouchers). We assigned species identity to the genome-sequenced specimens, if all vouchers were identified as the same species[69].

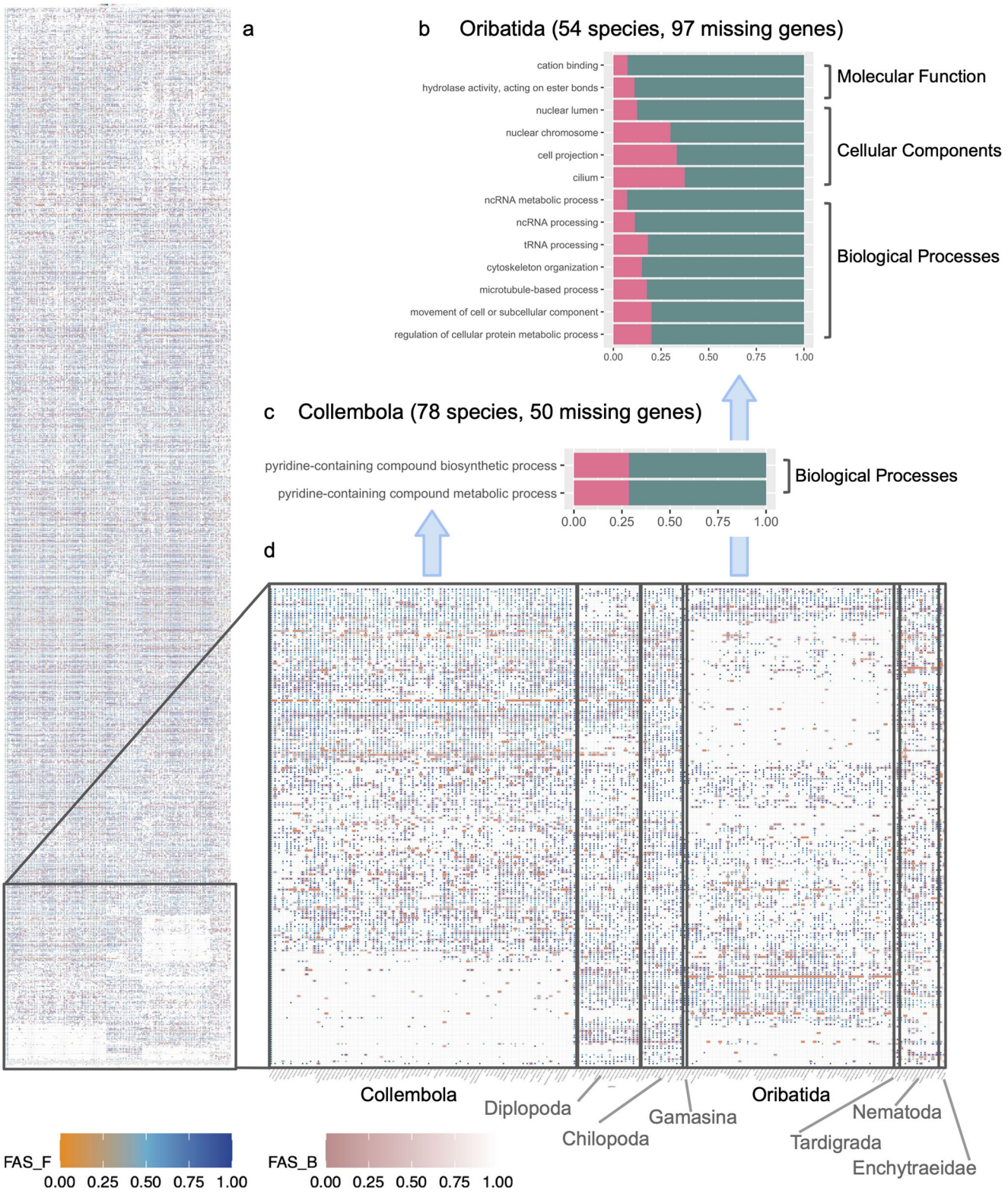

Fig. 5 Loss of metazoan core genes in soil invertebrate species. phylogenetic profiles of 1482 metazoan core genes across 177 soil invertebrate species; fraction of genes annotated with GO terms in the loss set (red) and in the background set (green) in oribatid mites **b** and springtails **c**; **d** genes consistently missing in springtails or oribatid mites. Colours in **a**, **d** represent feature architecture similarity among the identified orthologs and the reference gene, with a score between 1 (same architecture) and 0 (dissimilar architecture, or no features in the reference protein). The score is computed once by comparing the reference gene with the identified ortholog (FAS_F, dots on the graphic), and once by comparing the identified ortholog with the reference gene (FAS_B, background colour to dots). Data underlying the GO enrichment analysis are available as Supplementary Data 4.

Illumina sequencing. Sequencing libraries for each specimen, or pool of specimens, were prepared in-house at Senckenberg, Frankfurt, Germany with the BEST protocol[70] or with the NEBnext ULTRA II DNA Library Prep Kit, according to the manufacturer's protocol. Short-read Illumina sequencing (300-bp paired-end) was carried out at Novogene Europe (Cambridge, UK) using the NovaSeq 6000 platform, with unique dual indexing as the library tagging strategy for multiplexing on the lanes. Our central aim with the genome database was to improve species identifications. As this can be achieved with low sequencing coverage[65], our initial sequencing efforts targeted 2 gigabase (Gb) per species. We increased efforts to 10 Gb per species as sequencing became more affordable. For most of the reported genomes we obtained ~10 Gb per species.

Genome assembly pipeline. We established a pipeline to assemble reads into draft genomes (Fig. 1). First, the sequencing adapters were trimmed using Trimmomatic (v0.39; paramaters: ILLUMINACLIP:adapters.fasta:2:30:10:8:true SLIDINGWINDOW:4:20 MINLEN:50 TOPHRED33[71]). The trimmed reads were queried against the human genome (GRCh38 assembly on NCBI) using Kraken2 (v2.0.9-beta; --confidence set to 0.2, other parameters default[72]), and all 'human' positive reads, if any, were discarded. The remaining reads were then assembled using SPAdes (v3.14.1; default settings[73]). The resulting contigs were then queried against the NCBI non-redundant nucleotide database using blastn (megablast mode, -max_target_seqs 10, -max_hsps 1, -evalue 1e-25), and against the NCBI non-redundant protein database using Diamond (blastx mode, --sensitive --max-target-seqs 1, --evalue 1e-25[74]). NCBI databases were downloaded on 27-Oct-2020. Blobtools2 (v2.3.3[7]) was used to perform a taxonomic assignment based on the Blast and Diamond results, using the 'bestsumorder' rule. The contigs assigned to the phylum of the target organism as well as the unassigned contigs were kept (i.e., contigs assigned to other phyla were considered obvious contaminants and removed). Redundans (v0.14a[75]) was used to reduce the amount of duplication in the retained contigs, as well as further scaffolding and gap closing (default parameters were used). The resulting scaffolds were used as the final assembly draft for subsequent analyses. The Burrows-Wheeler Aligner (BWA) was then used to map the reads on the assembly and samtools[76] (v1.11-2-g26d7c73) to compute and plot the mapping statistics (e.g., GC content).

Quality assessment of assemblies using BUSCO. Benchmarking Universal Single-Copy Orthologs (BUSCO) databases[27] are sets of genes for specific taxon groups, where every gene in the BUSCO set is expected to be present once in each member species. We searched for BUSCO genes in our final assemblies as a quality indicator of genome assembly completeness, we used the most specific BUSCO database that was available for each of the invertebrate groups (nematoda_odb10 BUSCO genes for nematode assemblies, arthropoda_odb10 for arthropods, metazoa_odb10 for tardigrades, enchytraeids and earthworms). We selected the genome assembly with the highest percentage of complete BUSCO genes as the species representative if more than a single replicate per species was available. This resulted in a total of 232 genome assembly drafts used for downstream analyses.

Improving metatranscriptomic assignments. Metatranscriptomic reads were generated from soil samples collected along an elevation gradient spanning 400 m of elevation in the Alps[28,77,78]. Briefly, short soil cores were taken and preserved in LifeGuard (Qiagen, Hilden, Germany) in 2015 and 2017. RNA was extracted with an RNeasy PowerSoil Total RNA Kit (Qiagen) from ten cores. RNA sequencing libraries were prepared of each RNA extracts with a NEBNext Ultra RNA Library Prep Kit (Frankfurt am Main, Germany), and 8 gigabases of each library were sequenced at Novogene (UK) on an Illumina NovaSeq6000 sequencer in a 150 bp paired-end reaction. Reads were trimmed of adapters with Trimmomatic[71]. Reads were taxonomically assigned with kraken2[72] in a three-step process. First, we screened the metatranscriptomes against the human genome for eventual human contamination. Second, we assigned remaining reads with a custom database containing all bacterial, plant and fungal reference genomes from NCBI (accessed on 15.1.2023). Third, we then tested the impact of a dedicated genome database for soil invertebrate detection: unassigned reads from the second step were mapped against all springtail (57), oribatid mite (9) and myriapod genomes (8) available in NCBI RefSeq as of 20.6.2023, with and without including the 232 MetaInvert genomes (Supplementary Data 2). We visualised the richness of soil invertebrates along the elevation gradient at the genus level. As nucleotide sequence counts are not normally distributed and they are frequently overdispersed[79], we evaluated differences in community composition among the study years and habitats, and along the elevation with a model-based analysis of multivariate abundance data[80]. Community analyses were performed in R v4.2.2[81].

Building the phylogeny using metazoan BUSCO genes. We searched for BUSCO genes with the metazoan_obd10 database (v4.1.4) to generate a single phylogeny of the 232 soil invertebrate genomes and a selection of 118 publicly available invertebrate RefSeq genomes from NCBI (downloaded on 16.09.2021). The RefSeq genomes were included if they a) were from the same taxon group as our specimens, b) served to shorten the evolutionary distance between taxa in the tree. More specifically, we included any chromosome-level Protostomia genomes (excluding Insecta), genomes of any assembly quality for species within our 14 taxonomic groups of interest, and some additional specific outgroups (two Echinodermata, three Rotifera, a Priapulida, *Machilis hrabei* and *Drosophila albomicans*). We found 141 metazoan BUSCO genes which were present in at least 75% of the genome assemblies (Supplementary Data 1, 6). The phylogenetic approach is based on the https://github.com/mag-wolf/BUSCO-to-Phylogeny pipeline. We aligned these with Mafft (v7.481[82]) with 1000 iterative refinements. These gene alignments were then concatenated into a supermatrix using FASCONCAT (v1.04[83]) and trimmed using clipkit (v1.1.5[84]), keeping only parsimony-informative and completely conserved sites. We used IQ-TREE (v2.0.3[85]) to build four separate maximum likelihood trees (each with 1000 bootstrap replicates), selecting the best one based on the -log Likelihood value closest to zero[86]. We used R to visualise the phylogeny, using the packages ggtree (v3.1.5.900[87]), tidyverse, treeio (v1.17.2[88]) and colorspace[89].

We note the placement of Tardigrada in our phylogeny is next to Nematoda which is in disagreement with the currently accepted view that they should be closer to Arthropoda[90]. This is likely an artefact due to lack of public outgroup data[91,92], and has no downstream consequences for our analyses.

Estimating genome size. To estimate genome size, we used ModEst[31] which yields results comparable in accuracy to flow cytometry, the main non-sequencing method of genome size estimation, even from incomplete genomes. Briefly, we first plotted the distribution of sequencing coverage across each genome and visually inspected each plot for the mode coverage (the highest point of the peak). If a genome assembly did not have a clearly discernible peak in sequencing coverage then genome size was not estimated for this species. Otherwise, genome size was

estimated by dividing the total mapped bases by the mode coverage.

Estimating effective population size. Using mlRho (v2.9) we estimated theta directly from the genome data by making use of the genome-wide heterozygosity in the reference individual. This proxy measure of effective population size was calculated individually for each genome assembly with at least 8X coverage, twice as high as recommended by Haubold et al.[93].

Annotating repeat content. In addition to investigating several genome properties (i.e., GC content, BUSCO gene content, genome size and effective population size), and because repeat content is particularly relevant for explaining genome size variation among species, we also annotated the repetitive elements. Species-specific repeat libraries were constructed using the automated RepeatModeler (v2.0.1) pipeline with LTR Structural discovery pipeline activated[94]. For each genome, the resulting repeat libraries were merged with the RepBase (v26.05) Arthropoda-specific section[95] and subsequently used for the annotation and estimation of proportion of repetitive elements with RepeatMasker (v4.1.2-P1[96]).

Ecological trait data. To connect the 232 new genome assemblies with ecological traits of the respective species, we first gathered existing functional trait data from Edaphobase (https://portal.edaphobase.org/) and from literature. We focussed on a) body length (minimum female adult body length for nematodes, and mean body length for all other taxa) as a proxy for body size, b) reproduction mode, and c) known occurrences in different soil habitat types (based on level 2 hierarchies described by the Coordination of information on the environment (CORINE)[37]). We provide this collected information as an additional database resource in Supplementary Data 1.

Structural equation models. We tested established or hypothesised causal relations among genomic, life-cycle and ecological variables through a series of structural equation models, with the aim of resolving multivariate relationships from the many interrelated variables. We selected only genomes with at least 50% BUSCO completeness and 8X mode coverage. Log transformations were applied to body size variables (due to non-normal distribution as determined by a two-sided Kolmogorov-Smirnov test ($p < 0.01$). We fitted the SEMs with piecewiseSEM (v2.1.0[97]). We performed the path analyses for all taxa together (linear mixed effect models, with soil invertebrate groups as random variable), and separately for each of the more densely sampled taxa (Collembola, Oribatida, combined Chilopoda and Diplopoda, and Nematoda, linear models). Reproduction mode was included only into the models of all taxa and of oribatids, as this data were limited in the other groups.

Searching for core metazoan genes. As a first-look into the functional capacities of the soil invertebrates in our study, we searched the genomes for potential loss of protein-coding genes. To make this analysis robust, we decided to focus on evolutionarily old genes that were present already in the last common ancestor of the animals. Using 11 species from across the Metazoa tree of life (Supplementary Data 7) which were part of the Orthologous MAtrix database (OMA[98]), we computed a list of 1482 core metazoan genes which were common to at least 9 of these 11 species using DCC (https://github.com/BIONF/dcc2) and pre-computed ortholog groups from the OMA DB. Given the evolutionary age of these genes and their conserved presence throughout the animal evolution, it appears likely that their loss has a substantial functional impact. We preferred to use a custom core gene set over the standard BUSCO Metazoa ODB10 data set mainly for two reasons First, the BUSCO set with only 954 core genes is considerably smaller than the set computed by us. This gives us more power to detect differences in the presence/absence pattern of genes in the analysed taxa. Second, OMA groups represent cliques of orthologous proteins, i.e., all members within a group identify each other as pair-wise orthologs. As a consequence, OMA groups reconstruct orthologous relationships across proteins from many taxa with the highest precision among all available tools[98]. We then searched for orthologs of these 1482 metazoan core genes among the more complete (>50% BUSCO completeness) soil invertebrate genomes ($n = 177$) with fDOG-Assembly (https://github.com/BIONF/fDOG/tree/fdog_goes_assembly). fDOG-Assembly performs targeted, feature-aware ortholog search without the need for annotated genomes as the starting point. Due to the taxonomic breadth of our dataset, six separate ortholog searches were performed, each using the three most closely-related reference species with protein annotations available (Supplementary Data 8). Genes without orthologs in all investigated species were excluded from the following analyses. The resulting phylogenetic ortholog profiles were visualised with PhyloProfile (v1.8.6[99]) and clustered according to the euclidean distance of the presence and absence patterns of the ortholog groups. Hence, after visual inspection of the ortholog profiles, we were able to identify patches of core metazoan genes which were missing from certain groups.

We tested for gene ontology (GO) enrichment of the potentially missing genes using the InterProScan database[100] and the function runTest from the topGO package (v2.42.0[101]). For this GO-enrichment analysis, 1482 core metazoan genes were assigned to their ontology group(s), where GO annotation data were available. Using this list as a comparison, the two gene lists of interest (50 genes missing from the 78 Collembola species; 97 genes missing from the 54 Oribatida species) were separately tested for any significant enrichment of genes belonging to any of the three gene ontology groups (biological process, cellular component, or molecular function). Significant enrichment of a gene ontology term in the missing genes was stated when the category was represented by more than five genes in the list of 1482 core metazoan genes and with a significant over-representation in a Fisher's exact test ($p < 0.05$). Further, we manually screened putative functions associated with genes missing from Collembola and Oribatida assemblies in the UniProt database (accessed on 28.6.2023), aiming to identify functional or other relevant commonalities which might be missed by an algorithmic GO enrichment analysis.

The mean empirical probability of not being able to detect a particular gene in a taxon was 0.22 for all OMA genes, excluding those missing in springtails and oribatids. So this is also the probability of not finding a particular OMA gene in the genome of a new taxon. The probability that it is actually present in the majority of the taxa if it is also not found in the second sequenced species drops to 0.05; already in the third species in which the gene is not found, the probability that the gene is actually present in the majority of the species of the taxon is below the significance level.

Statistics and reproducibility. Genome analyses are based on Illumina genomes of 232 soil invertebrate species. Genome sizes could be estimated for 191 species. Metatranscripomic assignment was performed on 10 soil RNA samples. Structural equation models were fitted on genome properties of 143 taxa, including 27 myriapods, 34 oribatids, 68 springtails, 5 nematodes. Genomes of 177 species were assessed for the presence of core metazoan genes.Tests of normal distribution were performed to ensure that

assumptions of regression are fulfilled for the structural equation models.

Reporting summary. Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Vouchers are deposited in the collections of the Senckenberg Museum of Natural History Görlitz (SMNG), Germany. Raw sequence files and draft assemblies accessible through the ENA/NCBI project PRJNA758215. 28 S and COI barcodes are publicly available at dx.doi.org/10.5883/DS-TBGMI. Genome metadata can be accessed at the Genomes on a Tree (https://goat.genomehubs.org/projects/METAINVERT). Repeat elements can be accessed in the Dfam database (https://www.dfam.org/). Alignment of BUSCO genes and the resulting phylogenetic tree are available in FigShare (https://doi.org/10.6084/m9.figshare.24435052)[29]. Source data for Fig. 2 are part of Supplementary Data 1. Source data for Fig. 3 are provided in Supplementary Data 2. Source data for Fig. 5c, d are provided as Supplementary Data 4.

Code availability

No custom code or mathematical algorithms are central for the conclusions of the paper. R commands for metatranscriptome analysis and structural equation models are deposited in FigShare[29]. A list of used software with versions are deposited in FigShare[29].

Received: 16 August 2023; Accepted: 21 November 2023;
Published online: 08 December 2023

References

1. FAO, ITPS, GSBI, CBD & EC. *State of knowledge of soil biodiversity - Status, challenges and potentialities, Report 2020*. (FAO). https://doi.org/10.4060/cb1928en. 2020.
2. Potapov, A. M. et al. Feeding habits and multifunctional classification of soil-associated consumers from protists to vertebrates. *Biol. Rev.* **97**, 1057–1117 (2022).
3. García-Palacios, P., Maestre, F. T., Kattge, J. & Wall, D. H. Climate and litter quality differently modulate the effects of soil fauna on litter decomposition across biomes. *Ecol. Lett.* **16**, 1045–1053 (2013).
4. Bardgett, R. D. & van der Putten, W. H. Belowground biodiversity and ecosystem functioning. *Nature* **515**, 505–511 (2014).
5. de Vries, F. T. & Wallenstein, M. D. Below-ground connections underlying above-ground food production: a framework for optimising ecological connections in the rhizosphere. *J. Ecol.* **105**, 913–920 (2017).
6. Lavelle, P. et al. Soil invertebrates and ecosystem services. *Eur. J. Soil Biol.* **42**, S3–S15 (2006).
7. Challis, R., Richards, E., Rajan, J., Cochrane, G. & Blaxter, M. BlobToolKit – interactive quality assessment of genome assemblies. *G3 Genes Genomes Genet.* **10**, 1361–1374 (2020).
8. Guerra, C. A. et al. Tracking, targeting, and conserving soil biodiversity. *Science* **371**, 239–241 (2021).
9. Potapov, A. M. et al. Size compartmentalization of energy channeling in terrestrial belowground food webs. *Ecology* **102**, e03421 (2021).
10. Stork, N. E. How many species of insects and other terrestrial arthropods are there on earth? *Annu. Rev. Entomol.* **63**, 31–45 (2018).
11. Pearson, D. L., Hamilton, A. L. & Erwin, T. L. Recovery plan for the endangered taxonomy profession. *BioScience* **61**, 58–63 (2011).
12. Greshake Tzovaras, B. et al. What is in umbilicaria pustulata? A metagenomic approach to reconstruct the holo-genome of a lichen. *Genome Biol. Evol.* **12**, 309–324 (2020).
13. Pedersen, M. W. et al. Supplement: postglacial viability and colonization in North America's ice-free corridor. *Nature* **537**, 45–49 (2016).
14. Schmidt, A. et al. Shotgun metagenomics of soil invertebrate communities reflects taxonomy, biomass, and reference genome properties. *Ecol. Evol.* **12**, e8991 (2022).
15. Wang, Y. et al. Late Quaternary dynamics of Arctic biota from ancient environmental genomics. *Nature* **600**, 86–92 (2021).
16. Bista, I. et al. Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples. *Mol. Ecol. Resour.* **18**, 1020–1034 (2018).
17. Yates, M. C., Derry, A. M. & Cristescu, M. E. Environmental RNA: a revolution in ecological resolution? *Trends Ecol. Evol.* **36**, 601–609 (2021).
18. Shakya, M., Lo, C.-C. & Chain, P. S. G. Advances and challenges in metatranscriptomic analysis. *Front. Genet.* **10**, 904 (2019).
19. Seeber, P. A. & Epp, L. S. Environmental DNA and metagenomics of terrestrial mammals as keystone taxa of recent and past ecosystems. *Mammal. Rev.* **52**, 538–553 (2022).
20. Bálint, M. et al. Environmental DNA time series in ecology. *Trends Ecol. Evol.* **33**, 945–957 (2018).
21. Pedersen, M. W. et al. Environmental genomics of Late Pleistocene black bears and giant short-faced bears. *Curr. Biol.* https://doi.org/10.1016/j.cub.2021.04.027. (2021).
22. Law, S. R. et al. Metatranscriptomics captures dynamic shifts in mycorrhizal coordination in boreal forests. *Proc. Natl Acad. Sci. USA* **119**, e2118852119 (2022).
23. Lewin, H. A. et al. The Earth BioGenome Project 2020: starting the clock. *Proc. Natl Acad. Sci. USA* **119**, e2115635118 (2022).
24. Genereux, D. P. et al. A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245 (2020).
25. Feng, S. et al. Dense sampling of bird diversity increases power of comparative genomics. *Nature* **587**, 252–257 (2020).
26. Hotaling, S. et al. Long reads are revolutionizing 20 years of insect genome sequencing. *Genome Biol. Evol.* **13**, evab138 (2021).
27. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
28. Merges, D. et al. Metatranscriptomics reveals contrasting effects of elevation on the activity of bacteria and bacterial viruses in soil. *Mol. Ecol.* https://doi.org/10.1111/mec.16756. (2022).
29. Collins, G. et al. Supplementary Data to MetaInvert. https://doi.org/10.6084/m9.figshare.24435052.v1. (2023).
30. Winkler, M. et al. Side by side? Vascular plant, invertebrate, and microorganism distribution patterns along an alpine to nival elevation gradient. *Arct. Antarct. Alp. Res.* **50**, e1475951 (2018).
31. Pfenninger, M., Schönnenbeck, P. & Schell, T. ModEst: accurate estimation of genome size from next generation sequencing data. *Mol. Ecol. Resour.* **22**, 1454–1464 (2022).
32. Gregory, T. R. Coincidence, coevolution, or causation? DNA content, cellsize, and the C-value enigma. *Biol. Rev.* **76**, 65–101 (2001).
33. Lynch, M. Evolution of the mutation rate. *Trends Genet.* **26**, 345–352 (2010).
34. Sung, W., Ackerman, M. S., Miller, S. F., Doak, T. G. & Lynch, M. Drift-barrier hypothesis and mutation-rate evolution. *Proc. Natl Acad. Sci. USA* **109**, 18488–18492 (2012).
35. Blommaert, J. Genome size evolution: towards new model systems for old questions. *Proc. R. Soc. B Biol. Sci.* **287**, 20201441 (2020).
36. Pasquesi, G. I. M. et al. Squamate reptiles challenge paradigms of genomic repeat element evolution set by birds and mammals. *Nat. Commun.* **9**, 2774 (2018).
37. Steemans, C. Coordination of Information on the Environment (CORINE). *Encycl. Geogr. Inf. Sci. Ed. Kemp K Sage Publ. Inc Thousand Oaks CA* 49–50 (2008).
38. Hubbell, S. P. *The Unified Neutral Theory of Biodiversity and Biogeography*. (Princeton University Press, 2001).
39. Wang, J., Santiago, E. & Caballero, A. Prediction and estimation of effective population size. *Heredity* **117**, 193–206 (2016).
40. Chénais, B., Caruso, A., Hiard, S. & Casse, N. The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. *Gene* **509**, 7–15 (2012).
41. Hawkins, J. S., Grover, C. E. & Wendel, J. F. Repeated big bangs and the expanding universe: directionality in plant genome size evolution. *Plant Sci.* **174**, 557–562 (2008).
42. Rocha, E. P. C. & Danchin, A. Base composition bias might result from competition for metabolic resources. *Trends Genet.* **18**, 291–294 (2002).
43. Chen, Y.-J. et al. Metabolic flexibility allows bacterial habitat generalists to become dominant in a frequently disturbed ecosystem. *ISME J.* **15**, 2986–3004 (2021).
44. Chaurasia, A., Uliano, E., Berná, L., Agnisola, C. & D'Onofrio, G. Does Habitat Affect the Genomic GC Content? A Lesson from Teleostean Fish: A Mini Review. In *Fish Ecology* 61–80 (Nova Science Publishers, 2011).
45. Foerstner, K. U., von Mering, C., Hooper, S. D. & Bork, P. Environments shape the nucleotide composition of genomes. *EMBO Rep.* **6**, 1208–1213 (2005).
46. Moura, A., Savageau, M. A. & Alves, R. Relative amino acid composition signatures of organisms and environments. *PLoS ONE* **8**, e77319 (2013).
47. Charlesworth, B. & Barton, N. Genome size: does bigger mean worse? *Curr. Biol.* **14**, R233–R235 (2004).
48. Canapa, A., Barucca, M., Biscotti, M. A., Forconi, M. & Olmo, E. Transposons, genome size, and evolutionary insights in animals. *Cytogenet. Genome Res.* **147**, 217–239 (2015).
49. Elliott, T. A. & Gregory, T. R. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philos. Trans. R. Soc. B Biol. Sci.* **370**, 20140331 (2015).
50. Kapusta, A., Suh, A. & Feschotte, C. Dynamics of genome size evolution in birds and mammals. *Proc. Natl Acad. Sci. USA* **114**, E1460–E1469 (2017).

51. Plohl, M., Luchetti, A., Meštrović, N. & Mantovani, B. Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. *Gene* **409**, 72–82 (2008).
52. Meštrović, N. et al. Structural and functional liaisons between transposable elements and satellite DNAs. *Chromosome Res.* **23**, 583–596 (2015).
53. Hultgren, K. M., Jeffery, N. W., Moran, A. & Gregory, T. R. Latitudinal variation in genome size in crustaceans. *Biol. J. Linn. Soc.* **123**, 348–359 (2018).
54. Yu, J. P., Liu, W., Mai, C. L. & Liao, W. B. Genome size variation is associated with life-history traits in birds. *J. Zool.* **310**, 255–260 (2020).
55. Raffaele, S. & Kamoun, S. Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat. Rev. Microbiol.* **10**, 417–430 (2012).
56. Bono, L. M., Draghi, J. A. & Turner, P. E. Evolvability costs of niche expansion. *Trends Genet. TIG* **36**, 14–23 (2020).
57. MacArthur, R. H. *Geographical Ecology*. (Harper & Row Publishers Inc., 1972).
58. Sachdeva, V., Husain, K., Sheng, J., Wang, S. & Murugan, A. Tuning environmental timescales to evolve and maintain generalists. *Proc. Natl Acad. Sci. USA* **117**, 12693–12699 (2020).
59. Albalat, R. & Cañestro, C. Evolution by gene loss. *Nat. Rev. Genet.* **17**, 379–391 (2016).
60. Sharma, V. et al. A genomics approach reveals insights into the importance of gene losses for mammalian adaptations. *Nat. Commun.* **9**, 1215 (2018).
61. Guijarro-Clarke, C., Holland, P. W. H. & Paps, J. Widespread patterns of gene loss in the evolution of the animal kingdom. *Nat. Ecol. Evol.* **4**, 519–523 (2020).
62. De, S. et al. Pyridine: the scaffolds with significant clinical diversity. *RSC Adv.* **12**, 15385–15406 (2022).
63. Suring, W. et al. Evolutionary ecology of beta-lactam gene clusters in animals. *Mol. Ecol.* **26**, 3217–3229 (2017).
64. Schneider, C. et al. Two high-quality de novo genomes from single ethanol-preserved specimens of tiny metazoans (Collembola). *GigaScience* **10**, 5 (2021).
65. Bohmann, K., Mirarab, S., Bafna, V. & Gilbert, M. T. P. Beyond DNA barcoding: the unrealized potential of genome skim data in sample identification. *Mol. Ecol.* **29**, 2521–2534 (2020).
66. Macfadyen, A. Improved funnel-type extractors for soil arthropods. *J. Anim. Ecol.* **30**, 171–184 (1961).
67. Decker, H. *Phytonematologie*. (Deutscher Landwirtschftsverlag, 1969).
68. Gilbert, M. T. P., Moore, W., Melchior, L. & Worobey, M. DNA extraction from dry museum beetles without conferring external morphological damage. *PLoS ONE* **2**, e272 (2007).
69. Schenk, J., Hohberg, K., Helder, J., Ristau, K. & Traunspurger, W. The D3-D5 region of large subunit ribosomal DNA provides good resolution of German limnic and terrestrial nematode communities. *Nematology* **19**, 821–837 (2017).
70. Carøe, C. et al. Single-tube library preparation for degraded DNA. *Methods Ecol. Evol.* **9**, 410–419 (2017).
71. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
72. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
73. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
74. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
75. Pryszcz, L. P. & Gabaldón, T. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res.* **44**, e113 (2016).
76. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
77. Merges, D., Bálint, M., Schmitt, I., Böhning-Gaese, K. & Neuschulz, E. L. Spatial patterns of pathogenic and mutualistic fungi across the elevational range of a host plant. *J. Ecol.* **106**, 1545–1557 (2018).
78. Merges, D., Bálint, M., Schmitt, I., Manning, P. & Neuschulz, E. L. High throughput sequencing combined with null model tests reveals specific plant-fungi associations linked to seedling establishment and survival. *J. Ecol.* **108**, 574–585 (2020).
79. Bálint, M. et al. Millions of reads, thousands of taxa: microbial community structure and associations analyzed via marker genes. *FEMS Microbiol. Rev.* **40**, 686–700 (2016).
80. Wang, Y., Naumann, U., Wright, S. T. & Warton, D. I. mvabund– an R package for model-based analysis of multivariate abundance data. *Methods Ecol. Evol.* **3**, 471–474 (2012).
81. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2022).
82. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
83. Kück, P. & Meusemann, K. FASconCAT: convenient handling of data matrices. *Mol. Phylogenet. Evol.* **56**, 1115–1118 (2010).
84. Steenwyk, J. L., Iii, T. J. B., Li, Y., Shen, X.-X. & Rokas, A. ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. *PLoS Biol.* **18**, e3001007 (2020).
85. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
86. Misof, B. et al. Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**, 763–767 (2014).
87. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
88. Wang, L.-G. et al. Treeio: an R package for phylogenetic tree input and output with richly annotated and associated data. *Mol. Biol. Evol.* **37**, 599–603 (2020).
89. Zeileis, A. et al. Colorspace: a toolbox for manipulating and assessing colors and palettes. *J. Stat. Softw.* **96**, 1–49 (2020).
90. Treffkorn, S., Mayer, G. & Janssen, R. Review of extra-embryonic tissues in the closest arthropod relatives, onychophorans and tardigrades. *Philos. Trans. R. Soc. B Biol. Sci.* **377**, 20210270 (2022).
91. Giribet, G. & Edgecombe, G. D. The phylogeny and evolutionary history of arthropods. *Curr. Biol.* **29**, R592–R602 (2019).
92. Telford, M., Rota-Stabelli, O. & Pisani, D. Phylo-evo-devo, tardigrades and insights into the evolution of segmentation. in (Padova University Press, 2018).
93. Haubold, B., Pfaffelhuber, P. & Lynch, M. mlRho - a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. *Mol. Ecol.* **19**, 277–284 (2010).
94. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl Acad. Sci. USA* **117**, 9451–9457 (2020).
95. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
96. Smit, A., Hubbley, R. & Green, P. RepeatMasker Open 4.0. http://www.repeatmasker.org (2015).
97. Lefcheck, J. S. piecewiseSEM: piecewise structural equation modelling in r for ecology, evolution, and systematics. *Methods Ecol. Evol.* **7**, 573–579 (2016).
98. Altenhoff, A. M. et al. OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. *Nucleic Acids Res.* **49**, D373–D379 (2021).
99. Tran, N.-V., Greshake Tzovaras, B. & Ebersberger, I. PhyloProfile: dynamic visualization and exploration of multi-layered phylogenetic profiles. *Bioinf. Oxf. Engl.* **34**, 3041–3043 (2018).
100. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
101. Alexa, A. & Rahnenfuhrer, J. topGO: Enrichment Analysis for Gene Ontology v2. https://doi.org/10.18129/B9.bioc.topGO (2022).

Acknowledgements

This work is a result of the LOEWE Centre for Translational Biodiversity Genomics funded by the Hessen State Ministry of Higher Education, Research and the Arts (HMWK). This project contributes to the Soil Invertebrate Genome Initiative (https://tbg.senckenberg.de/sigi/), affiliated with the Earth BioGenome Project. Special thanks to Damian Baranski, Jürgen Otte and Jörg Müller for DNA extractions, to Astrid König for extraction and preparation of nematodes and tardigrades from Senckenberg cultures and soils, to Lena Bonassin and Jade Tessien for help with organising the datasets and barcodes, to Prof. Dr. Florian Grundler (INRES Molekulare Phytomedizin, Bonn University, Germany) for thousands of J2 juveniles of *Heterodera schachtii* and *Meloidogyne incognita* in ethanol from their cultures, and to Magnus Wolf for the BUSCO-to-phylogeny pipeline. Animal silhouettes originate from PhyloPic and they can be reused under Creative Commons licenses (http://www.phylopic.org).

Author contributions

P.D., I.E., K.H., O.L., R.L., M.P., M.B. conceived and designed the experiments. P.D., K.H., H.M., R.L., performed the experiments. G.C., C.S., L.B., I.E., O.L., H.M., J.u.R., C.R., R.V., M.P., M.B. analysed the data. C.S., L.B., U.B., A.C., P.D., I.E., K.H., D.M., H.M., J.ö.R., J.u.R., C.R., R.S., A.S., K.T. contributed materials/analysis tools. G.C., M.P., M.B. wrote the paper.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-023-05621-4.

Correspondence and requests for materials should be addressed to Miklós. Bálint.

Peer review information This manuscript was previously reviewed at another Nature Portfolio journal. *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: George Inglis. A peer review file is available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.