Genomic evidence for the wide-spread presence of lignocellulases among soil invertebrates

Hannah Muelbaier[1], Freya Arthen[1], Gemma Collins[1], Thomas Hickler[1], Karin Hohberg[1], Ricarda Lehmitz[1], Yannick Pauchet[2], Markus Pfenninger[3], Anton Potapov[1], Juliane Romahn[1], Ina Schaefer[1], Stefan Scheu[4], Clement Schneider[5], Ingo Ebersberger[6], and Miklós Bálint[7]

[1]Affiliation not available
[2]Max-Planck-Institute for Chemical Ecology
[3]Senckenberg Biodiversität & Klima Forschungszentrum
[4]University of Goettingen
[5]Senckenberg Gesellschaft fur Naturforschung
[6]Goethe University
[7]Senckenberg Biodiversity and Climate Research Centre

September 6, 2023

Abstract

Lignocellulose is a major component of plant biomass. Its decomposition is crucial for the terrestrial carbon cycle. Microorganisms are considered as primary decomposers and evidence increases that some invertebrates may also decompose lignocellulose. We investigated the taxonomic distribution and evolutionary origins of GH45 cellulases in a collection of soil invertebrate genomes and found that these genes are common in springtails and oribatid mites. Phylogenetic analysis revealed that cellulase genes were acquired early in the evolutionary history of these groups. Domain architectures and predicted 3D enzyme structures indicate that these cellulases are functional. Patterns of presence and absence of these genes across different lineages prompt further investigation into their evolutionary and ecological benefits. The ubiquity of cellulase genes suggests that soil invertebrates may play a role in lignocellulose decomposition, independently from microorganisms. Understanding the ecological and evolutionary implications might be crucial for understanding soil food webs and the carbon cycle.

Genomic evidence for the wide-spread presence of lignocellulases among soil invertebrates

Hannah Muelbaier[1,2], Freya Arthen[1,2], Gemma Collins,[2,3] Thomas Hickler[4,5], Karin Hohberg[6], Ricarda Lehmitz[2,6], Yannick Pauchet[7], Markus Pfenninger[2,4,8], Anton Potapov[9,10,11], Juliane Romahn[2,4], Ina Schaefer[2,4,11], Stefan Scheu[12], Clément Schneider[2,6], Ingo Ebersberger[1,2,4]†, Miklós Bálint[2,4,13]†*

† These authors contributed equally to this work

Affiliations:

[1] Applied Bioinformatics Group, Inst. of Cell Biology and Neuroscience, Goethe University, Max-von-Laue Str. 13, 60438 Frankfurt am Main, Germany

[2] LOEWE Centre for Translational Biodiversity Genomics, Senckenberganlage 25, 60325 Frankfurt am Main, Germany

[3] Manaaki Whenua – Landcare Research, Auckland, New Zealand

[4] Senckenberg Biodiversity and Climate Research Centre, Senckenberganlage 25, 60325 Frankfurt am Main, Germany

[5] Department of Physical Geography, Goethe University, Altenhöferallee 1, 60438 Frankfurt/Main, Germany

[6] Senckenberg Museum of Natural History Görlitz, Am Museum 1, 02826 Görlitz, Germany

[7] Insect symbiosis, Max Planck Institute for Chemical Ecology, Beutenberg campus, Hans-Knoell-Str. 8, 07745 Jena

[8] Institute for Molecular and Organismic Evolution, Johannes Gutenberg University, Mainz, Germany

[9] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Puschstrasse 4, 04103, Leipzig, Germany

[10] Institute of Biology, University of Leipzig, Puschstrasse 4, 04103, Leipzig, Germany

[11] Animal Ecology, University of Goettingen, Untere Karspüle 2, 37073, Goettingen, Germany

[12] "J.F. Blumenbach Institute of Zoology and Anthropology, University of Goettingen, Berliner Str. 28 37073 Goettingen, Germany

[13] Institute of Insect Biotechnology, Justus-Liebig University, Heinrich-Buff-Ring 26, 35392, Giessen, Germany

* miklos.balint@senckenberg.de

Abstract

Lignocellulose is a major component of plant biomass. Its decomposition is crucial for the terrestrial carbon cycle. Microorganisms are considered as primary decomposers and evidence increases that some invertebrates may also decompose lignocellulose. We investigated the taxonomic distribution and evolutionary origins of GH45 cellulases in a collection of soil invertebrate genomes and found that these genes are common in springtails and oribatid mites. Phylogenetic analysis revealed that cellulase genes were acquired early in the evolutionary history of these groups. Domain architectures and predicted 3D enzyme structures indicate that these cellulases are functional. Patterns of presence and absence of these genes across different lineages prompt further investigation into their evolutionary and ecological benefits. The ubiquity of cellulase genes suggests that soil invertebrates may play a role in lignocellulose decomposition, independently from microorganisms. Understanding the ecological and evolutionary implications might be crucial for understanding soil food webs and the carbon cycle.

Introduction

Most photosynthetically bound carbon on land ends up in woody plants as lignocellulose, a composite of several polysaccharides. The decomposition of lignocellulose occurs predominantly in soils, which returns most of this carbon into the atmosphere (Post et al., 1990). Nevertheless, terrestrial ecosystems currently sequester about 29% of the anthropogenic carbon emissions, which implies an important but not fully understood role of terrestrial carbon cycling for climate regulation (Cragg et al., 2015). Microorganisms, especially bacteria and fungi, encode glycoside hydrolase cocktails for lignocellulose degradation in their genomes (Cragg et al., 2015), and are considered the main actors of decomposition (Bradford et al., 2017; Crowther et al., 2019; Pausas and Bond, 2020). The contribution of animals to decomposition of lignocellulose - beyond purely mechanical shredding - remains less understood. Experiments have shown that the presence of soil invertebrates can increase litter mass loss by up to 50 percent (García-Palacios et al., 2013). It is estimated that they decompose more deadwood in tropical forests than free-living microorganisms (Griffiths et al., 2019). Nevertheless, the mechanisms behind decomposition performed by soil invertebrates remains obscure and the ability of soil animals to degrade composite polysaccharides without relying on gut symbionts remains a long-standing debate in soil ecology (Berg et al., 2004; Cragg et al., 2015).

It was originally assumed that lignocellulose degradation by animals was entirely ‘outsourced’ to the gut microbiome (Briones, 2018; García-Palacios et al., 2013). However, evidence is emerging that at least some invertebrates, such as molluscs, crustaceans and phytophagous insects can synthesize cellulase enzymes themselves (Busch et al., 2019; Chang and Lai, 2018; Cragg et al., 2015; Griffiths et al., 2021; Han et al., 2022; Kern et al., 2013; King et al., 2010; Shelomi et al., 2014; Watanabe et al., 1998). Scattered evidence also exists for the expression of active endogenous cellulases by distantly-related soil invertebrates, e.g. the earthworm *Pheretima hilgendorfi* (Nozaki et al., 2009), the Antarctic springtail *Cryptopygus antarcticus* (Hong et al., 2014), as well as few oribatid mites and other springtails (Busch et al., 2019). Based on these individual findings, we hypothesize that a larger fraction of soil invertebrates than previously thought may be directly contributing to lignocellulose decomposition of dead plant matter in soils, without necessarily relying on a

specialized microbiome. Given their global abundance and diversity in many soil ecosystems (FAO et al., 2020; Phillips et al., 2019; Potapov et al., 2023; van den Hoogen et al., 2019), soil invertebrates could therefore have an important but so far overlooked role in the terrestrial carbon cycle which is distinct from the lignocellulose decomposition ability of microorganisms. To evaluate whether endogenous decomposition ability is a common feature shared by the main groups of soil invertebrates, we screened a diverse set of newly sequenced genomes of Collembola, Enchytraeidae, Gamasina, Myriapoda, Nematoda, Oribatida, Tardigrada as representatives of dominating and ubiquitous soil invertebrates, for the presence and origin of cellulase genes.

Results

We used a fungal sequence as a starting point to obtain a comprehensive overview of the taxonomic distribution of GH45-type cellulases across all species represented by a genome in NCBI RefSeq. We identified orthologs to fungal GH45 cellulases in 16,401 bacteria, 910 archaea and 1,101 eukaryotes (Table S1). The resulting phylogenetic profile revealed that GH45 cellulases are abundant only in fungi, where they are, however, limited to individual systematic groups (Fig. 1A). In bacteria, only 31 species were found to harbour a GH45-type cellulase. Next to fungi, orthologs to fungal cellulases were largely confined to animals (Fig. 1A). We found no evidence that these comprise fungal contaminations (see Taxonomic assignment and contaminant detection in Materials and Methods). In metazoans, most orthologs were found in arthropods, and they were completely absent in vertebrates.

The analysis of publicly available genomes shed light on the general distribution of GH45-type cellulases across the Tree of Life. However, it lacked the resolution to investigate the occurrence of this cellulase in soil invertebrates. We extended the analysis by including novel genome assemblies for an additional 176 species representing a diverse selection of soil invertebrates (NCBI BioProject PRJNA758215) into the analysis. This revealed a high occurrence of GH45-type cellulases in springtails (70%; 56 out of 78 analyzed species) and in oribatid mites (60%; 33 out of 54 species, Table S2). Additionally, we detected cellulases in Coleoptera and Thysanoptera (2 out of 2 species, Table S1). In three out of nine nematode species we also found GH45 cellulases, whereas none were found in 30 representatives of Chilopoda and Diplopoda (Table S2).

We found GH45-type cellulases in three of four known main springtail lineages (Poduromorha, Entomobryomorpha, Symphypleona), missing only from the earliest branching Neelipleona. Within Symphypleona and Entomobryomorpha, cellulases are consistently absent in one clade each (Supp. Fig S1).

GH45-type cellulases are present in almost all representatives of the basal clades in Oribatida (Enarthronota, Mixonomata, Holosomata; Table S2). In contrast, they were missing in half of the investigated species from the later-branching Brachypylina (Supp. Fig S2). All of the species investigated here that did not have GH45 cellulase genes were sexually reproducing, while the GH45 cellulase genes were present in other sexually-reproducing species and parthenogens. Besides Oribatida, we investigated a second mite taxon, the Gamasina. Gamasina is represented by two species in the present data set (Table S2) and GH45 cellulase genes were absent in both.

We reconstructed the phylogenetic relationships of the sequences identified both in RefSeq assemblies and in soil invertebrate genomes (Fig. 1B) to better understand the evolutionary trajectory resulting in the present-day distribution of soil invertebrate GH45 cellulases. Animals are paraphyletic in this tree, but most of the invertebrate GH45 orthologs (94 out of 107) were placed in only four phylogenetically largely homogeneous clades, one each for the Collembola, Oribatida, Thysanoptera, and Coleoptera. On a larger scale, we found that the invertebrate cellulase clades are embedded into an evolutionary background formed by fungal and bacterial sequences. We noted that oribatids are placed in a single clade together with a few interspersed bacterial sequences. We compared the phylogeny from oribatid mites and springtails with our reconstructed gene tree (Fig. 1C). A reconstruction of GH45 phylogeny with all found cellulase co-orthologs shows a highly dynamic and complex evolutionary history with lineage-specific gene duplications and losses (Supp. Fig.S3).

We investigated the GH45 domain architecture and 3D protein structure and found that domain architecture in fungi, antarctic springtail and mustard leaf beetle were similar. The enzymes in these species also had similar predicted 3D structures with the cellulases found by us (Fig. 1D).

Discussion

Most cellulases discovered to date in metazoan genomes belong to the GH45 family (Busch et al., 2019), endo-β-1,4,-glucanases (EC 3.2.1.4) which catalyze the decomposition of complex cellulose into more accessible oligosaccharides (Davies et al., 1993). It was hypothesized that invertebrate GH45 cellulases were repeatedly obtained via horizontal gene transfer (HGT) from fungal donors (Busch et al., 2019). Our search of 18,412 RefSeq genomes revealed that only 31 bacterial species harbour a GH45-type cellulase. This confirms the earlier hypothesis about the evolutionary roots of this enzyme within fungi (Busch et al., 2019). The taxonomic distribution of GH45 presence was non-uniform in metazoans, with most orthologs being found in arthropods. Our results confirm previous research which did not find evidence for GH45 presence in vertebrates (Chang and Lai, 2018).

The screening of new soil invertebrate genomes uncovered novel GH45 cellulase presence patterns. GH45 genes were found in well over half of investigated springtail and oribatid mite species. Springtails form a basal hexapod group abundant across the globe, especially in cold regions (Potapov et al., 2023). They are known as fungal feeders, but also consume detritus and fresh plant materials (Potapov et al., 2020). While the ortholog search showed the presence of GH45-type cellulase in springtail lineages Poduromorha, Entomobryomorpha and Symphypleona it is missing from the earliest branching Neelipleona. Neelipleona feed on detritus colonized by fungi and bacteria rather than on plant remains (Potapov et al., 2022), and thus may not directly require lignocellulase activity. A single acquisition event post-dating the divergence of the Neelipleona may explain this observation. However, since Neelipleona are represented only by a few taxa, the conclusion on cellulase absence should be taken cautiously. Within Symphypleona and Entomobryomorpha, cellulases are consistently absent in one clade each. The absence of cellulase genes is unexpected in the lucerne flea *Sminthurus viridis* that feeds on live leaf tissue (Greenslade and Ireson, 1986), and in the families of Dicyrtomidae, Bourletiellidae and Sminthuridae which consume mainly fresh plant materials (Potapov et al., 2020, 2022). The lack of

cellulase genes suggests that these taxa either outsource lignocellulose decomposition to their midgut microbiome, rely on other GH families, or do not digest cellulose. However, the presence of GH45-type cellulase genes seems to be a common trait among most springtails.

Mites are the most numerous arthropods on land (Rosenberg et al., 2023), with most representatives in soil ecosystems belonging to Oribatida. Oribatid mites can have diverse feeding strategies, but they mostly feed on leaf litter at different decomposition stages and on microorganisms (Maraun et al., 2023). Our results showed that GH45-type cellulases are present all over the basal clades of Oribatida while they were missing in half of the Branchypylina, which are later-branching. One interesting finding is that GH45 cellulase is missing especially in sexually reproducing mites, while they are present in other sexually reproducing taxa. This fits to the pattern that parthenogenetic oribatid mites tend to occupy lower trophic positions and typically function as primary decomposers, opposed to secondary decomposers feeding predominantly on microorganisms (Fischer et al., 2014). However, ecological interpretation of these patterns is difficult since we do not know if species without GH45 cellulase genes contain other classes of cellulases, digest cellulose with the help of their microbiome or are indeed incapable of cellulose digestion. In general, the complex pattern of GH45 presence is similar to the low phylogenetic conservatism of ecological traits in oribatids, such as feeding mode (Potapov et al., 2022).

As expected, we detected cellulases in Coleoptera (Kirsch et al., 2014). GH45-type cellulases were completely absent in Chilopoda and Diplopoda. The latter was surprising as Diplopoda are a key litter-feeding soil invertebrate group (Joly et al., 2020; Potapov et al., 2022). GH45 cellulases were also absent in Gamasina mites which are predators and therefore might not benefit from cellulose degradation. The first report of endogenous cellulases in Thysanoptera suggests that our analysis uncovers only the tip of the iceberg. We expect that taxonomically broad genome sequencing of eukaryotes promoted e.g. by the Earth BioGenome Project (Formenti et al., 2022; Lewin et al., 2022) will recover further animal groups in possession of enzymes targeting lignocellulose decomposition.

Taken together, our data suggests an early acquisition of a GH45-type cellulase during the diversifications of both springtails and oribatids, instead of repeated horizontal transfer events. This implies that the possession of a GH45-type cellulase is an ancestral trait in these groups. Similar to our results, cellulase acquisition was shown to be important for the diversification of herbivorous beetles (Kirsch et al., 2014). Differences in the GH45 cellulase gene tree from the oribatid and springtail species trees likely result from a highly dynamic evolution of the GH45 cellulase repertoire. Lineage-specific gene duplications and losses have partially disconnected the evolutionary history of the contemporary cellulase genes from the phylogeny of the species they are found in (Supp. Fig. S3). Lineage-specific duplications have been described for other cellulases (Shelomi et al., 2016; Shin et al., 2022), and differential duplicate loss has been shown to result in gene tree - species tree incongruencies (Parey et al., 2020). The presence of cellulases detected in thrips suggests that similar processes might have been important also during the evolution of other arthropod groups.

Although cellulase presence in genomes is not a proof of function, several lines of evidence point toward functionality. First, domain architecture of GH45 cellulases in fungi, antarctic springtail and mustard leaf beetle are similar, and the enzymes themselves also have similar predicted 3D structures with the cellulases found by us (Fig. 1D). Second, fungal (Cragg et

al., 2015), beetle (Busch et al., 2019) and Antarctic springtail (Hong et al., 2014) GH45 cellulases were all shown to be functional. Finally, orthologs with conserved domain architectures were retained over hundreds of millions of years of evolution in springtails and oribatids. This suggests little change has occurred in the trophic niche and position of springtails and oribatid mites in soil food webs since their origin, further supported by the presence of both taxonomic groups in the first fossil soils (Schaefer and Caruso, 2019; Shear et al., 1984). Taken together, these are strong indications that GH45-type cellulases in springtails and oribatids perform cellulose decomposition. Future work to experimentally evaluate the functional properties of soil invertebrate cellulases (Song et al., 2017) should consider all glycoside hydrolase genes, as gene duplication events may have led to substrate diversification (Busch et al., 2019; Shin et al., 2022), with duplicates being able to break down other polysaccharides like xyloglucan, mannans or xylan. While orthologs can be identified bioinformatically, functional properties need to be confirmed experimentally by expressing these enzymes heterologously, and test their substrate specificity on cellulose and hemicellulose polysaccharides.

The ability for repeated HGT of GH45-type cellulases, and long term conservation in new recipient taxa, may be related to several properties. First, cellulases, as secreted, gut-acting enzymes (Fischer et al., 2013) do not depend on existing physiological pathways and their regulation for proper functioning in the recipient organism. Second, such enzymes likely reach their correct extracellular destination directly after the successful transfer of the cellulase gene, its incorporation into the genome and its translation, because the signals of protein export generally work independent of origin in most other taxa, even over vast evolutionary distances (Clérico et al., 2008). Third, the reaction catalyzed by cellulases yields products that can serve as a beneficial fitness-relevant resource in any organism, because the necessary downstream pathways are ubiquitously present. Supporting these arguments, genes transferred horizontally are often secreted proteins (Savory et al., 2015; Undheim and Jenner, 2021).

Horizontal gene transfer of cellulases, among other plant cell wall-degrading enzymes, is a key process in the evolution of herbivory in arthropods (Wybouw et al., 2016). It resulted, for example, in the massive radiation of Phytophaga, the most species-rich clade of beetles (Busch et al., 2019), and in adaptation to lignocellulose-rich diets in crustaceans (King et al., 2010). The long-term evolutionary preservation of GH45 genes suggests that cellulases likely confer fitness benefits to soil invertebrates. These benefits may come from a direct use of plant carbohydrate resources, although some theories imply that soil invertebrates are limited rather by access to proteins, but not by access to carbohydrates. The ability to degrade complex polysaccharides may also provide access to more nutritious, protein-rich cytosols or microorganisms colonizing the inside of plant cells, such as saprotrophic fungi, which are considered as major dietary components of both collembolans and oribatids (Pollierer and Scheu, 2021). We expect that a taxonomically broad comparison of the presence of cellulase genes with traits and trophic niches (Maraun et al., 2023; Potapov et al., 2022) will provide insights into the functional ecology and evolution of soil invertebrates.

The wide-spread presence of cellulases in springtails and oribatid mites suggests that invertebrates are independently capable of enzymatic lignocellulose decomposition, forming a third evolutionarily and ecologically distinct group with such capability, in addition to bacteria and fungi. This has important consequences for our understanding of soil food webs and the soil carbon cycle. Fungi compared to bacteria are known to react differently to

environmental change such as experimental warming (Melillo et al., 2017) or habitat degradation (Zhou et al., 2018). This results from key differences in life history strategies, e.g. growth rates or nutrient use (Jansson and Hofmockel, 2020). Their differential reaction to environmental change influences decomposition as distinct taxa determine the rate and biochemical pathways of organic matter processing (Crowther et al., 2019). For example, fungal-based, slow energy channels are more resistant to drought, limiting C and N losses from agricultural soils (de Vries et al., 2012). Fungi accordingly contribute more to litter decomposition than bacteria under drought conditions (Ullah et al., 2023). Soil invertebrates react differently to environmental change compared to microorganisms (Sünnemann et al., 2021). We hypothesize that global change has a more detrimental impact on decomposition performed by soil invertebrates, given their lower effective population sizes and adaptive elasticity (Lanfear et al., 2014; Pauls et al., 2013). It might be essential to consider these differences for a better integration of below-ground processes into ecosystems models (Chertov et al., 2017; Deckmyn et al., 2020; Filser et al., 2016) including global carbon models (Friedlingstein et al., 2022), and for better predictions of soil carbon and nutrient cycling.

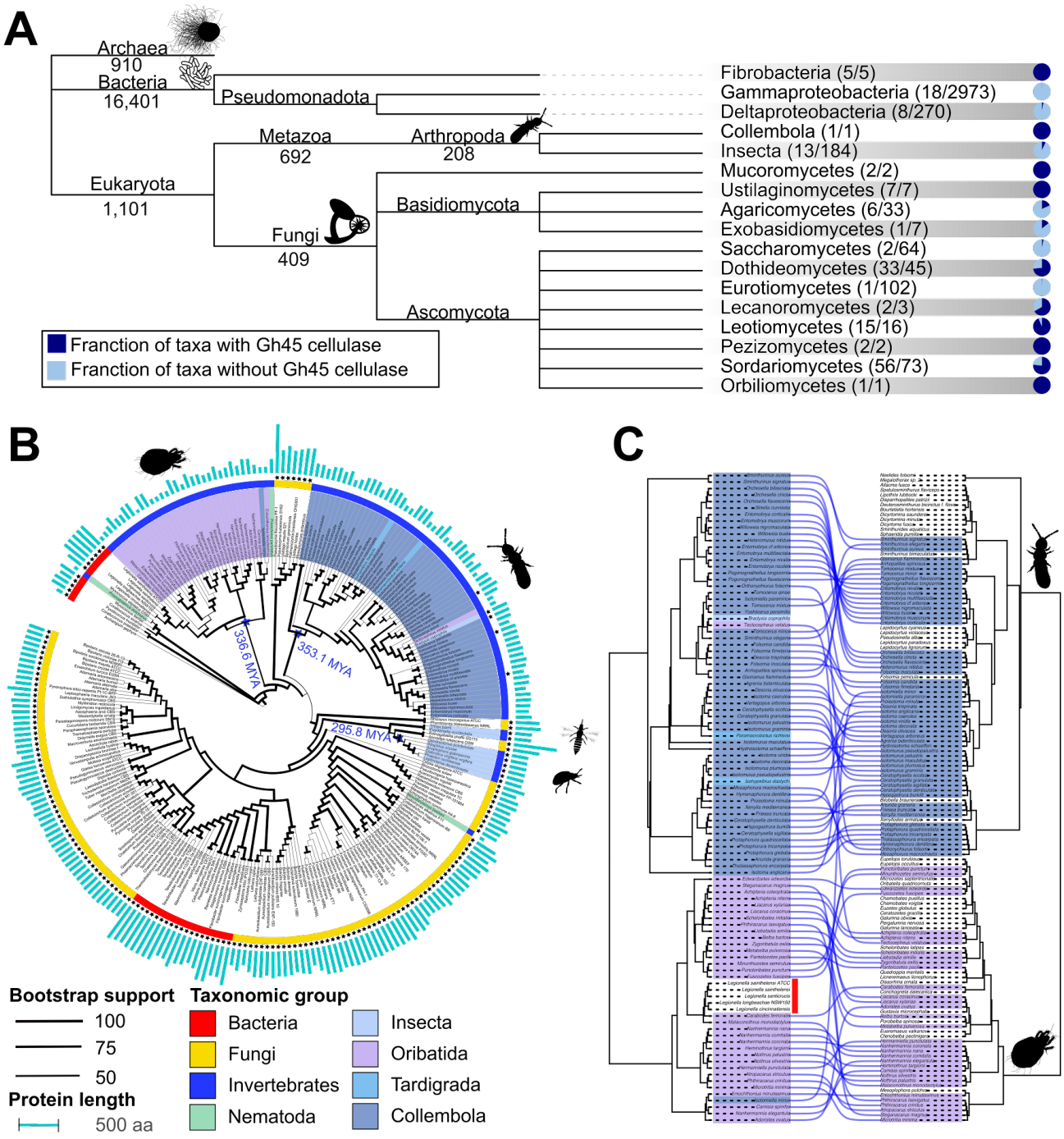

A
Archaea
910
Bacteria
16,401
Eukaryota
1,101
Pseudomonadota
Metazoa
692
Arthropoda
208
Fungi
409
Basidiomycota
Ascomycota
Franction of taxa with Gh45 cellulase
Franction of taxa without Gh45 cellulase
Fibrobacteria (5/5)
Gammaproteobacteria (18/2973)
Deltaproteobacteria (8/270)
Collembola (1/1)
Insecta (13/184)
Mucoromycetes (2/2)
Ustilaginomycetes (7/7)
Agaricomycetes (6/33)
Exobasidiomycetes (1/7)
Saccharomycetes (2/64)
Dothideomycetes (33/45)
Eurotiomycetes (1/102)
Lecanoromycetes (2/3)
Leotiomycetes (15/16)
Pezizomycetes (2/2)
Sordariomycetes (56/73)
Orbiliomycetes (1/1)
B
336.6 MYA
353.1 MYA
295.8 MYA
Bootstrap support
100
75
50
Protein length
500 aa
Taxonomic group
Bacteria
Fungi
Invertebrates
Nematoda
Insecta
Oribatida
Tardigrada
Collembola
C
D
Domain architecture
Rhizoctonia solani
lcr
Gh45
Phaedon cochleariae
lcr
Gh45
3D structure comparison

Fig. 1. A) Abundance of GH45 cellulases in the three domains of life in NCBI RefSeq genomes; B) Maximum likelihood phylogeny of the GH45-cellulase family. Branch lengths are not drawn to scale, line weights indicate percent bootstrap support. Species represented by a genome assembly in the NCBI RefSeq or GenBank databases are indicated by an asterisk. Pictograms identify the four main soil invertebrate clades: oribatid mites, springtails, thrips and beetles (clockwise). Bars mark GH45 protein length. Internal node labels provide age estimates of the respective clades (Kumar et al., 2022); C) correspondence of oribatid and springtail GH45 gene trees (left) and of a phylogenomic reconstruction of the species phylogenies (right). D) Comparison of protein domain architecture of *F. candida* and *R. solani* (left) and a 3D structure alignment of both proteins (right). The color indicates the similarity of structures in the aligned proteins, with blue marking a high, and red marking low correspondence. Lcr: low complexity region; Gh45: Pfam glyco-hydro 45 domain (PF02015).

Materials and Methods

Domain architectures of invertebrate GH45-type cellulases

Reviewed evidence exists for the presence of GH45-type cellulases in the antarctic springtail (*Cryptopygus antarcticus*; Collembola, UniprotID D3GDK4) and the mustard beetle (*Phaedon cochleariae*; Insecta, UniprotID O97401). Since these experimentally confirmed cellulases harbor a Glyco-hydro 45 Pfam domain (PF02015) (Bankevich et al., 2012), we restricted our analysis to cellulase orthologs that carry this Pfam domain. Pfam domains were annotated with hmmscan from the HMMER package using Pfam v.32 using the default e-value cutoff of 0.01.

Genome Assembly pipeline

The genome assemblies provided by the MetaInvert Project (Bioproject ID: PRJNA758215) cover a phylogenetically diverse set of soil-living invertebrates collected from the field or obtained from cultures. Short read Illumina sequencing (300bp paired-end) with the NovaSeq 6000 platform was done at Novogene Europe (Cambridge, UK), reads were trimmed with Trimmomatic, human contaminating reads were filtered with Kraken2 (Wood et al., 2019) and contig assembly was done with SPAdes (Bankevich et al., 2012). The resulting contigs were taxonomically assigned with Blobtools2 (Challis et al., 2020) using the NCBI non-redundant protein database as a reference, and only contigs with an assignment to the phylum of the target species together with unassigned contigs were kept. Redundancy reduction and scaffolding was done with Redundans (Pryszcz and Gabaldón, 2016), and genome assembly completeness was assessed with BUSCO (v 4.1.4) using the precomputed metazoan (obd10) reference set. For our ortholog search, we selected genomes with at least 50% BUSCO completeness (176 assemblies, Table S2).

RefSeq Gene Set collection

We downloaded gene sets for all 18,412 taxa represented in the NCBI RefSeq Genome release 207 (O'Leary et al., 2016). The resulting taxon collections comprised 16,401 bacteria, 910 archaea., 409 fungi, 262 invertebrates and 430 vertebrates. The taxon list together with the accession numbers are provided in Table S1.

Taxonomic assignment and contaminant detection

To rule out that fungal or bacterial contaminations of the underlying genome assemblies are responsible for the animal cellulase orthologs (Steinegger and Salzberg, 2020), we taxonomically classified each of the detected invertebrate orthologs. In brief, we used the sequence as a query for a Diamond v.2.0.13 (Buchfink et al., 2015) search against the NCBI non-redundant protein database (downloaded January 2022). From the resulting hit list, we excluded the trivial hit against itself and then assigned the query sequence to the last common ancestor of the taxa within a 10% bit score margin of the best hit (Huson et al., 2016). A sequence was flagged as a putative contaminant if its taxonomic assignment was not placed on the lineage from the species whose genome was analyzed to the root of the tree of cellular life. This workflow is implemented into the software package taXaminer (https://github.com/BIONF/taxaminer).This provided no evidence for a foreign origin of these sequences (Table S3).

Orthology-based phylogenetic profiles of fungal Gh45 cellulase

Profile-based targeted ortholog searches in annotated gene sets were performed with fDOG (Birikmen et al., 2021) using the GH45 cellulase of the fungus *Rhizoctonia solani* (NCBI Accession XP_043186467.1) as the seed. For the training of the initial profile Hidden Markov model we used the parameter *--minDist genus* and *--maxDist phylum* limiting the number of training sequences to 6 (see Table S4 for more information). Candidate orthologs were filtered for the presence of the Pfam glyco-hydro 45 domain (PF02015; see Table S5 for a list of discarded orthologs). Ortholog search in the unannotated MetaInvert genome assemblies were performed with the fDOG extension fDOG-Assembly. In brief, genomic regions likely containing a GH45-type cellulase were identified with a tBLASTn search using the consensus sequence included in the initial core gh45 core group from fDOG as query. The hit region was extended by 500 nucleotides on either side and genes in the resulting candidate genomic region were annotated with MetaEuk v5.34c21f2 (Levy Karin et al., 2020) using the OMA database (release December 2021) (Nguyen et al., 2015) as the reference database for the gene prediction. The corresponding protein sequences were then tested for orthology using the routines of fDOG and afterwards features were annotated with FAS (Dosch et al., 2023). The fDOG-assembly workflow is available from https://github.com/BIONF/fDOG/tree/fdog_goes_assembly. The results from fDOG and fDOG-Assembly were merged and visualized with PhyloProfile (Tran et al., 2018).

Gh45 cellulase gene tree reconstruction

To investigate the evolutionary history of the Gh45 cellulase, we used the identified orthologs for a gene tree reconstruction. If the ortholog search obtained more than one co-ortholog, we used the one that is most similar to the seed protein for the tree reconstruction. Sequences were aligned with Muscle v3.8.1551 (Edgar, 2004) and alignment columns comprising more than 50% gaps were removed with a custom perl script. The resulting multiple sequence alignment was used as input for a maximum likelihood tree reconstruction with IQ-TREE (Nguyen et al., 2015) v. 1.6.8. Branch support was assessed with 1000 bootstrap replicates using the ultrafast bootstrap approach. The SH-aLRT branch test was performed, and the optimal number of cores was automatically detected via IQ-TREE (- nt AUTO). The gene tree was visualized with iTOL (Letunic and Bork, 2021). Animal GH45-type cellulases are paraphyletic in this tree, and a topology test using the AU test (Shimodaira and Hasegawa, 1999) confirmed that a tree with monophyletic animal cellulases explained the data significantly worse (p-AU=9.2E-4). A second gene tree

containing all identified orthologs and co-orthologs was reconstructed with the same approach. iTOL was used to prune and visualize the gene tree (containing all oribatids and springtails) as well as to connect cellulase genes from the same species.

Phylogenies of springtails and oribatids

Phylogenies of springtail and oribatid organisms included in the MetaInvert project were computed separately with a supermatrix approach. BUSCO version 5.4.2 (Simão et al., 2015) with the precomputed BUSCO Arthropoda gene set (db10) was used to search for orthologs in all genome assemblies. Multiple sequence alignments were computed with MAFFT using local pairwise alignment and at maximum 1000 iterations (7.481)(Katoh and Standley, 2013), trimmed with clipkit (1.3.0) (Steenwyk et al., 2020) and concatenated with used FASconCAT-G (1.04) (Kück and Longo, 2014) into a supermatrix. Four phylogenetic trees per taxon group were reconstructed with IQ-TREE (Nguyen et al., 2015). Branch support was assessed with 1000 bootstrap replicates using the ultrafast bootstrap approach. The best-fit model was Q.insect+F+R9 for oribatids, and Q.insect+F+R10 for springtails, automatically chosen by ModelFinder according to BIC. The final consensus tree was computed with splitstree (4.19.0)(Huson and Bryant, 2006) by summarizing the four IQ-TREEs into a consensus tree. The consensus trees were outgroup-rooted using *Sarcoptes scabiei* (GCA_020844145.1) for oribatids and *Machilis hrabei* (GCA_003456935.1), *Drosophila albomicans* (GCA_009650485.1), and *Tyrophagus putrescentiae* from MetaInvert for springtails as outgroups.

Correspondence of GH45 tree and phylogenies

The BUSCO-based phylogenies of springtails and oribatids were outgroup-rooted and merged for the comparison with the Gh45 cellulase genetree. The tanglegram matches taxa by links and was computed with R with packages phytools v1.4-0 (Revell, 2012) and castor v1.7.6 (Louca and Doebeli, 2018).

3D structure comparison

The 3D structures of gh45 cellulase genes from *R.solani* and mustard beetle were retrieved from precomputed predictions from UniProt (accession numbers A0A0B7FQX1 and O97401). The structures were visualized and compared with VMD (Humphrey et al., 1996) and the extensions MultiSeq (Roberts et al., 2006) in combination with the alignment tool STAMP (Russell and Barton, 1992).

References

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol* **19**:455–477. doi:10.1089/cmb.2012.0021

Berg M, Stoffer M, Heuvel H. 2004. Feeding guilds in Collembola based on digestive enzymes. *Pedobiologia* **48**:589–601. doi:10.1016/j.pedobi.2004.07.006

Birikmen M, Bohnsack KE, Tran V, Somayaji S, Bohnsack MT, Ebersberger I. 2021. Tracing Eukaryotic Ribosome Biogenesis Factors Into the Archaeal Domain Sheds Light on the Evolution of Functional Complexity. *Front Microbiol* **12**:2598. doi:10.3389/fmicb.2021.739000

Bradford MA, Veen GF (Ciska), Bonis A, Bradford EM, Classen AT, Cornelissen JHC, Crowther TW, De Long JR, Freschet GT, Kardol P, Manrubia-Freixa M, Maynard DS, Newman GS, Logtestijn RSP, Viketoft M, Wardle DA, Wieder WR, Wood SA, van der Putten WH. 2017. A test of the hierarchical model of litter decomposition. *Nat Ecol Evol* **1**:1836–1845. doi:10.1038/s41559-017-0367-4

Briones MJI. 2018. The Serendipitous Value of Soil Fauna in Ecosystem Functioning: The Unexplained Explained. *Front Environ Sci* **6**.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**:59–60. doi:10.1038/nmeth.3176

Busch A, Danchin EGJ, Pauchet Y. 2019. Functional diversification of horizontally acquired glycoside hydrolase family 45 (GH45) proteins in Phytophaga beetles. *BMC Evol Biol* **19**:100. doi:10.1186/s12862-019-1429-9

Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. 2020. BlobToolKit – Interactive Quality Assessment of Genome Assemblies. *G3 Genes Genomes Genet* **10**:1361–1374. doi:10.1534/g3.119.400908

Chang WH, Lai AG. 2018. Mixed evolutionary origins of endogenous biomass-depolymerizing enzymes in animals. *BMC Genomics* **19**:483. doi:10.1186/s12864-018-4861-0

Chertov O, Komarov A, Shaw C, Bykhovets S, Frolov P, Shanin V, Grabarnik P, Priputina I, Zubkova E, Shashkov M. 2017. Romul_Hum—A model of soil organic matter formation coupling with soil biota activity. II. Parameterisation of the soil food web biota activity. *Ecol Model* **345**:125–139. doi:10.1016/j.ecolmodel.2016.10.024

Clérico EM, Maki JL, Gierasch LM. 2008. Use of synthetic signal sequences to explore the protein export machinery. *Pept Sci* **90**:307–319. doi:10.1002/bip.20856

Cragg SM, Beckham GT, Bruce NC, Bugg TD, Distel DL, Dupree P, Etxabe AG, Goodell BS, Jellison J, McGeehan JE, McQueen-Mason SJ, Schnorr K, Walton PH, Watts JE, Zimmer M. 2015. Lignocellulose degradation mechanisms across the Tree of Life. *Curr Opin Chem Biol*, Energy • Mechanistic biology **29**:108–119. doi:10.1016/j.cbpa.2015.10.018

Crowther TW, van den Hoogen J, Wan J, Mayes MA, Keiser AD, Mo L, Averill C, Maynard DS. 2019. The global soil community and its influence on biogeochemistry. *Science* **365**:eaav0550. doi:10.1126/science.aav0550

Davies GJ, Dodson GG, Hubbard RE, Tolley SP, Dauter Z, Wilson KS, Hjort C, Mikkelsen JM, Rasmussen G, Schülein M. 1993. Structure and function of endoglucanase V. *Nature* **365**:362–364. doi:10.1038/365362a0

de Vries FT, Liiri ME, Bjørnlund L, Bowker MA, Christensen S, Setälä HM, Bardgett RD. 2012. Land use alters the resistance and resilience of soil food webs to drought. *Nat Clim Change* **2**:276–280. doi:10.1038/nclimate1368

Deckmyn G, Flores O, Mayer M, Domene X, Schnepf A, Kuka K, Looy KV, Rasse DP, Briones MJI, Barot S, Berg M, Vanguelova E, Ostonen I, Vereecken H, Suz LM, Frey B, Frossard A, Tiunov A, Frouz J, Grebenc T, Öpik M, Javaux M, Uvarov A, Vindušková O, Krogh PH, Franklin O, Jiménez J, Yuste JC. 2020. KEYLINK: towards a more integrative soil representation for inclusion in ecosystem scale models. I. review and model concept. *PeerJ* **8**:e9750. doi:10.7717/peerj.9750

Dosch J, Bergmann H, Tran V, Ebersberger I. 2023. FAS: assessing the similarity between proteins using multi-layered feature architectures. *Bioinformatics* **39**:btad226. doi:10.1093/bioinformatics/btad226

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**:1792–1797. doi:10.1093/nar/gkh340

FAO, ITPS, GSBI, CBD, EC. 2020. State of knowledge of soil biodiversity - Status, challenges and potentialities, Report 2020. Rome, Italy: FAO. doi:10.4060/cb1928en

Filser J, Faber JH, Tiunov AV, Brussaard L, Frouz J, De Deyn G, Uvarov AV, Berg MP, Lavelle P, Loreau M, Wall DH, Querner P, Eijsackers H, Jiménez JJ. 2016. Soil fauna: key to new carbon models. *SOIL* **2**:565–582. doi:10.5194/soil-2-565-2016

Fischer BM, Meyer E, Maraun M. 2014. Positive correlation of trophic level and proportion of

462 sexual taxa of oribatid mites (Acari: Oribatida) in alpine soil systems. *Exp Appl Acarol*
463 **63**:465–479. doi:10.1007/s10493-014-9801-3
464 Fischer R, Ostafe R, Twyman RM. 2013. Cellulases from Insects In: Vilcinskas A, editor.
465 Yellow Biotechnology II: Insect Biotechnology in Plant Protection and Industry,
466 Advances in Biochemical Engineering/Biotechnology. Berlin, Heidelberg: Springer.
467 pp. 51–64. doi:10.1007/10_2013_206
468 Formenti G, Theissinger K, Fernandes C, Bista I, Bombarely A, Bleidorn C, Ciofi C, Crottini
469 A, Godoy JA, Höglund J, Malukiewicz J, Mouton A, Oomen RA, Paez S, Palsbøll PJ,
470 Pampoulie C, Ruiz-López María J., Svardal H, Theofanopoulou C, de Vries J,
471 Waldvogel A-M, Zhang Guojie, Mazzoni CJ, Jarvis ED, Bálint M, Formenti G,
472 Theissinger K, Fernandes C, Bista I, Bombarely A, Bleidorn C, Čiampor F, Ciofi C,
473 Crottini A, Godoy JA, Hoglund J, Malukiewicz J, Mouton A, Oomen RA, Paez S,
474 Palsbøll P, Pampoulie C, Ruiz-López María José, Svardal H, Theofanopoulou C, de
475 Vries J, Waldvogel A-M, Zhang Goujie, Mazzoni CJ, Jarvis E, Bálint M, Aghayan SA,
476 Alioto TS, Almudi I, Alvarez N, Alves PC, Amorim IR, Antunes A, Arribas P, Baldrian
477 P, Berg PR, Bertorelle G, Böhne A, Bonisoli-Alquati A, Boštjančić LL, Boussau B,
478 Breton CM, Buzan E, Campos PF, Carreras C, Castro LFi, Chueca LJ, Conti E,
479 Cook-Deegan R, Croll D, Cunha MV, Delsuc F, Dennis AB, Dimitrov D, Faria R,
480 Favre A, Fedrigo OD, Fernández R, Ficetola GF, Flot J-F, Gabaldón T, Galea Agius
481 DR, Gallo GR, Giani AM, Gilbert MTP, Grebenc T, Guschanski K, Guyot R, Hausdorf
482 B, Hawlitschek O, Heintzman PD, Heinze B, Hiller M, Husemann M, Iannucci A,
483 Irisarri I, Jakobsen KS, Jentoft S, Klinga P, Kloch A, Kratochwil CF, Kusche H,
484 Layton KKS, Leonard JA, Lerat E, Liti G, Manousaki T, Marques-Bonet T, Matos-
485 Maraví P, Matschiner M, Maumus F, Mc Cartney AM, Meiri S, Melo-Ferreira J,
486 Mengual X, Monaghan MT, Montagna M, Mysłajek RW, Neiber MT, Nicolas V, Novo
487 M, Ozretić P, Palero F, Pârvulescu L, Pascual M, Paulo OS, Pavlek M, Pegueroles
488 C, Pellissier L, Pesole G, Primmer CR, Riesgo A, Rüber L, Rubolini D, Salvi D,
489 Seehausen O, Seidel M, Secomandi S, Studer B, Theodoridis S, Thines M, Urban L,
490 Vasemägi A, Vella A, Vella N, Vernes SC, Vernesi C, Vieites DR, Waterhouse RM,
491 Wheat CW, Wörheide G, Wurm Y, Zammit G. 2022. The era of reference genomes in
492 conservation genomics. *Trends Ecol Evol* **37**:197–202.
493 doi:10.1016/j.tree.2021.11.008
494 Friedlingstein P, O'Sullivan M, Jones MW, Andrew RM, Gregor L, Hauck J, Le Quéré C,
495 Luijkx IT, Olsen A, Peters GP, Peters W, Pongratz J, Schwingshackl C, Sitch S,
496 Canadell JG, Ciais P, Jackson RB, Alin SR, Alkama R, Arneth A, Arora VK, Bates
497 NR, Becker M, Bellouin N, Bittig HC, Bopp L, Chevallier F, Chini LP, Cronin M, Evans
498 W, Falk S, Feely RA, Gasser T, Gehlen M, Gkritzalis T, Gloege L, Grassi G, Gruber
499 N, Gürses Ö, Harris I, Hefner M, Houghton RA, Hurtt GC, Iida Y, Ilyina T, Jain AK,
500 Jersild A, Kadono K, Kato E, Kennedy D, Klein Goldewijk K, Knauer J, Korsbakken
501 JI, Landschützer P, Lefèvre N, Lindsay K, Liu J, Liu Z, Marland G, Mayot N, McGrath
502 MJ, Metzl N, Monacci NM, Munro DR, Nakaoka S-I, Niwa Y, O'Brien K, Ono T,
503 Palmer PI, Pan N, Pierrot D, Pocock K, Poulter B, Resplandy L, Robertson E,
504 Rödenbeck C, Rodriguez C, Rosan TM, Schwinger J, Séférian R, Shutler JD,
505 Skjelvan I, Steinhoff T, Sun Q, Sutton AJ, Sweeney C, Takao S, Tanhua T, Tans PP,
506 Tian X, Tian H, Tilbrook B, Tsujino H, Tubiello F, van der Werf GR, Walker AP,
507 Wanninkhof R, Whitehead C, Willstrand Wranne A, Wright R, Yuan W, Yue C, Yue X,
508 Zaehle S, Zeng J, Zheng B. 2022. Global Carbon Budget 2022. *Earth Syst Sci Data*
509 **14**:4811–4900. doi:10.5194/essd-14-4811-2022
510 García-Palacios P, Maestre FT, Kattge J, Wall DH. 2013. Climate and litter quality differently
511 modulate the effects of soil fauna on litter decomposition across biomes. *Ecol Lett*
512 **16**:1045–1053. doi:10.1111/ele.12137
513 Greenslade P, Ireson JE. 1986. Collembola of the Southern Australian Culture Steppe and
514 Urban Environments: A Review of Their Pest Status and Key to Identification. *Aust J*
515 *Entomol* **25**:273–291. doi:10.1111/j.1440-6055.1986.tb01115.x
516 Griffiths HM, Ashton LA, Evans TA, Parr CL, Eggleton P. 2019. Termites can decompose

more than half of deadwood in tropical rainforest. *Curr Biol* **29**:R118–R119. doi:10.1016/j.cub.2019.01.012
Griffiths HM, Ashton LA, Parr CL, Eggleton P. 2021. The impact of invertebrate decomposers on plants and soil. *New Phytol* **231**:2142–2149. doi:10.1111/nph.17553
Han Z, Sieriebriennikov B, Susoy V, Lo W-S, Igreja C, Dong C, Berasategui A, Witte H, Sommer RJ. 2022. Horizontally Acquired Cellulases Assist the Expansion of Dietary Range in Pristionchus Nematodes. *Mol Biol Evol* **39**:msab370. doi:10.1093/molbev/msab370
Hong SM, Sung HS, Kang MH, Kim C-G, Lee Y-H, Kim D-J, Lee JM, Kusakabe T. 2014. Characterization of Cryptopygus antarcticus Endo-β-1,4-Glucanase from Bombyx mori Expression Systems. *Mol Biotechnol* **56**:878–889. doi:10.1007/s12033-014-9767-8
Humphrey W, Dalke A, Schulten K. 1996. VMD: visual molecular dynamics. *J Mol Graph* **14**:33–38, 27–28. doi:10.1016/0263-7855(96)00018-5
Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, Ruscheweyh H-J, Tappu R. 2016. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLOS Comput Biol* **12**:e1004957. doi:10.1371/journal.pcbi.1004957
Huson DH, Bryant D. 2006. Application of Phylogenetic Networks in Evolutionary Studies. *Mol Biol Evol* **23**:254–267. doi:10.1093/molbev/msj030
Jansson JK, Hofmockel KS. 2020. Soil microbiomes and climate change. *Nat Rev Microbiol* **18**:35–46. doi:10.1038/s41579-019-0265-7
Joly F-X, Coq S, Coulis M, David J-F, Hättenschwiler S, Mueller CW, Prater I, Subke J-A. 2020. Detritivore conversion of litter into faeces accelerates organic matter turnover. *Commun Biol* **3**:1–9. doi:10.1038/s42003-020-01392-4
Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* **30**:772–780. doi:10.1093/molbev/mst010
Kern M, McGeehan JE, Streeter SD, Martin RNA, Besser K, Elias L, Eborall W, Malyon GP, Payne CM, Himmel ME, Schnorr K, Beckham GT, Cragg SM, Bruce NC, McQueen-Mason SJ. 2013. Structural characterization of a unique marine animal family 7 cellobiohydrolase suggests a mechanism of cellulase salt tolerance. *Proc Natl Acad Sci* **110**:10189–10194. doi:10.1073/pnas.1301502110
King AJ, Cragg SM, Li Y, Dymond J, Guille MJ, Bowles DJ, Bruce NC, Graham IA, McQueen-Mason SJ. 2010. Molecular insight into lignocellulose digestion by a marine isopod in the absence of gut microbes. *Proc Natl Acad Sci* **107**:5345–5350. doi:10.1073/pnas.0914228107
Kirsch R, Gramzow L, Theißen G, Siegfried BD, ffrench-Constant RH, Heckel DG, Pauchet Y. 2014. Horizontal gene transfer and functional diversification of plant cell wall degrading polygalacturonases: Key events in the evolution of herbivory in beetles. *Insect Biochem Mol Biol* **52**:33–50. doi:10.1016/j.ibmb.2014.06.008
Kück P, Longo GC. 2014. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Front Zool* **11**:81. doi:10.1186/s12983-014-0081-x
Kumar S, Suleski M, Craig JM, Kasprowicz AE, Sanderford M, Li M, Stecher G, Hedges SB. 2022. TimeTree 5: An Expanded Resource for Species Divergence Times. *Mol Biol Evol* **39**:msac174. doi:10.1093/molbev/msac174
Lanfear R, Kokko H, Eyre-Walker A. 2014. Population size and the rate of evolution. *Trends Ecol Evol* **29**:33–41. doi:10.1016/j.tree.2013.09.009
Letunic I, Bork P. 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* **49**:W293–W296. doi:10.1093/nar/gkab301
Levy Karin E, Mirdita M, Söding J. 2020. MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome* **8**:48. doi:10.1186/s40168-020-00808-x

572 Lewin HA, Richards S, Lieberman Aiden E, Allende ML, Archibald JM, Bálint M, Barker KB,
573 Baumgartner B, Belov K, Bertorelle G, Blaxter ML, Cai J, Caperello ND, Carlson K,
574 Castilla-Rubio JC, Chaw S-M, Chen L, Childers AK, Coddington JA, Conde DA,
575 Corominas M, Crandall KA, Crawford AJ, DiPalma F, Durbin R, Ebenezer TE,
576 Edwards SV, Fedrigo O, Flicek P, Formenti G, Gibbs RA, Gilbert MTP, Goldstein
577 MM, Graves JM, Greely HT, Grigoriev IV, Hackett KJ, Hall N, Haussler D, Helgen
578 KM, Hogg CJ, Isobe S, Jakobsen KS, Janke A, Jarvis ED, Johnson WE, Jones SJM,
579 Karlsson EK, Kersey PJ, Kim J-H, Kress WJ, Kuraku S, Lawniczak MKN, Leebens-
580 Mack JH, Li X, Lindblad-Toh K, Liu X, Lopez JV, Marques-Bonet T, Mazard S, Mazet
581 JAK, Mazzoni CJ, Myers EW, O'Neill RJ, Paez S, Park H, Robinson GE, Roquet C,
582 Ryder OA, Sabir JSM, Shaffer HB, Shank TM, Sherkow JS, Soltis PS, Tang B,
583 Tedersoo L, Uliano-Silva M, Wang K, Wei X, Wetzer R, Wilson JL, Xu X, Yang H,
584 Yoder AD, Zhang G. 2022. The Earth BioGenome Project 2020: Starting the clock.
585 *Proc Natl Acad Sci U S A* **119**:e2115635118. doi:10.1073/pnas.2115635118
586 Louca S, Doebeli M. 2018. Efficient comparative phylogenetics on large trees. *Bioinformatics*
587 **34**:1053–1055. doi:10.1093/bioinformatics/btx701
588 Maraun M, Thomas T, Fast E, Treibert N, Caruso T, Schaefer I, Lu J-Z, Scheu S. 2023. New
589 perspectives on soil animal trophic ecology through the lens of C and N stable
590 isotope ratios of oribatid mites. *Soil Biol Biochem* **177**:108890.
591 doi:10.1016/j.soilbio.2022.108890
592 Melillo JM, Frey SD, DeAngelis KM, Werner WJ, Bernard MJ, Bowles FP, Pold G, Knorr MA,
593 Grandy AS. 2017. Long-term pattern and magnitude of soil carbon feedback to the
594 climate system in a warming world. *Science* **358**:101–105.
595 doi:10.1126/science.aan2874
596 Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective
597 Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol*
598 **32**:268–274. doi:10.1093/molbev/msu300
599 Nozaki M, Miura C, Tozawa Y, Miura T. 2009. The contribution of endogenous cellulase to
600 the cellulose digestion in the gut of earthworm (Pheretima hilgendorfi:
601 Megascolecidae). *Soil Biol Biochem* **41**:762–769. doi:10.1016/j.soilbio.2009.01.016
602 O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B,
603 Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V,
604 Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D,
605 Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P,
606 McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD,
607 Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE,
608 Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio
609 M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference sequence (RefSeq) database at
610 NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids*
611 *Res* **44**:D733-745. doi:10.1093/nar/gkv1189
612 Parey E, Louis A, Cabau C, Guiguen Y, Roest Crollius H, Berthelot C. 2020. Synteny-
613 Guided Resolution of Gene Trees Clarifies the Functional Impact of Whole-Genome
614 Duplications. *Mol Biol Evol* **37**:3324–3337. doi:10.1093/molbev/msaa149
615 Pauls SU, Nowak C, Bálint M, Pfenninger M. 2013. The impact of global climate change on
616 genetic diversity within populations and species. *Mol Ecol* **22**:925–946.
617 doi:https://doi.org/10.1111/mec.12152
618 Pausas JG, Bond WJ. 2020. On the Three Major Recycling Pathways in Terrestrial
619 Ecosystems. *Trends Ecol Evol* **35**:767–775. doi:10.1016/j.tree.2020.04.004
620 Phillips HRP, Guerra CA, Bartz MLC, Briones MJI, Brown G, Crowther TW, Ferlian O,
621 Gongalsky KB, Hoogen J van den, Krebs J, Orgiazzi A, Routh D, Schwarz B, Bach
622 EM, Bennett JM, Brose U, Decaëns T, König-Ries B, Loreau M, Mathieu J, Mulder C,
623 Putten WH van der, Ramirez KS, Rillig MC, Russell D, Rutgers M, Thakur MP, Vries
624 FT de, Wall DH, Wardle DA, Arai M, Ayuke FO, Baker GH, Beauséjour R, Bedano
625 JC, Birkhofer K, Blanchart E, Blossey B, Bolger T, Bradley RL, Callaham MA,
626 Capowiez Y, Caulfield ME, Choi A, Crotty FV, Crumsey JM, Dávalos A, Cosin DJD,

627 Dominguez A, Duhour AE, Eekeren N van, Emmerling C, Falco LB, Fernández R,
628 Fonte SJ, Fragoso C, Franco ALC, Fugère M, Fusilero AT, Gholami S, Gundale MJ,
629 López MG, Hackenberger DK, Hernández LM, Hishi T, Holdsworth AR, Holmstrup M,
630 Hopfensperger KN, Lwanga EH, Huhta V, Hurisso TT, Iannone BV, Iordache M,
631 Joschko M, Kaneko N, Kanianska R, Keith AM, Kelly CA, Kernecker ML, Klaminder
632 J, Koné AW, Kooch Y, Kukkonen ST, Lalthanzara H, Lammel DR, Lebedev IM, Li Y,
633 Lidon JBJ, Lincoln NK, Loss SR, Marichal R, Matula R, Moos JH, Moreno G, Morón-
634 Ríos A, Muys B, Neirynck J, Norgrove L, Novo M, Nuutinen V, Nuzzo V, P95 MR,
635 Pansu J, Paudel S, Pérès G, Pérez-Camacho L, Piñeiro R, Ponge J-F, Rashid MI,
636 Rebollo S, Rodeiro-Iglesias J, Rodríguez MÁ, Roth AM, Rousseau GX, Rozen A,
637 Sayad E, Schaik L van, Scharenbroch BC, Schirrmann M, Schmidt O, Schröder B,
638 Seeber J, Shashkov MP, Singh J, Smith SM, Steinwandter M, Talavera JA, Trigo D,
639 Tsukamoto J, Valença AW de, Vanek SJ, Virto I, Wackett AA, Warren MW, Wehr
640 NH, Whalen JK, Wironen MB, Wolters V, Zenkova IV, Zhang W, Cameron EK,
641 Eisenhauer N. 2019. Global distribution of earthworm diversity. *Science* **366**:480–
642 485. doi:10.1126/science.aax4851
643 Pollierer MM, Scheu S. 2021. Stable isotopes of amino acids indicate that soil decomposer
644 microarthropods predominantly feed on saprotrophic fungi. *Ecosphere* **12**:e03425.
645 doi:10.1002/ecs2.3425
646 Post WM, Peng T-H, Emanuel WR, King AW, Dale VH, DeAngelis DL, others. 1990. The
647 global carbon cycle. *Am Sci* **78**:310–326.
648 Potapov A, Bellini BC, Chown SL, Deharveng L, Janssens F, Kováč L, Kuznetsova N, Ponge
649 J-F, Potapov M, Querner P. 2020. Towards a global synthesis of Collembola
650 knowledge: challenges and potential solutions. *Soil Org* **92**:161–188.
651 Potapov AM, Beaulieu F, Birkhofer K, Bluhm SL, Degtyarev MI, Devetter M, Goncharov AA,
652 Gongalsky KB, Klarner B, Korobushkin DI, Liebke DF, Maraun M, Mc Donnell RJ,
653 Pollierer MM, Schaefer I, Shrubovych J, Semenyuk II, Sendra A, Tuma J, Tůmová M,
654 Vassilieva AB, Chen T-W, Geisen S, Schmidt O, Tiunov AV, Scheu S. 2022. Feeding
655 habits and multifunctional classification of soil-associated consumers from protists to
656 vertebrates. *Biol Rev* **97**:1057–1117. doi:10.1111/brv.12832
657 Potapov AM, Guerra CA, van den Hoogen J, Babenko A, Bellini BC, Berg MP, Chown SL,
658 Deharveng L, Kováč Ľ, Kuznetsova NA, Ponge J-F, Potapov MB, Russell DJ,
659 Alexandre D, Alatalo JM, Arbea JI, Bandyopadhyaya I, Bernava V, Bokhorst S,
660 Bolger T, Castaño-Meneses G, Chauvat M, Chen T-W, Chomel M, Classen AT,
661 Cortet J, Čuchta P, Manuela de la Pedrosa A, Ferreira SSD, Fiera C, Filser J,
662 Franken O, Fujii S, Koudji EG, Gao M, Gendreau-Berthiaume B, Gomez-Pamies DF,
663 Greve M, Tanya Handa I, Heiniger C, Holmstrup M, Homet P, Ivask M, Janion-
664 Scheepers C, Jochum M, Joimel S, Claudia S. Jorge B, Jucevica E, Ferlian O, Iuñes
665 de Oliveira Filho LC, Klauberg-Filho O, Baretta D, Krab EJ, Kuu A, de Lima ECA, Lin
666 D, Lindo Z, Liu A, Lu J-Z, Luciañez MJ, Marx MT, McCary MA, Minor MA, Nakamori
667 T, Negri I, Ochoa-Hueso R, Palacios-Vargas JG, Pollierer MM, Querner P,
668 Raschmanová N, Rashid MI, Raymond-Léonard LJ, Rousseau L, Saifutdinov RA,
669 Salmon S, Sayer EJ, Scheunemann N, Scholz C, Seeber J, Shveenkova YB,
670 Stebaeva SK, Sterzynska M, Sun X, Susanti WI, Taskaeva AA, Thakur MP, Tsiafouli
671 MA, Turnbull MS, Twala MN, Uvarov AV, Venier LA, Widenfalk LA, Winck BR,
672 Winkler D, Wu D, Xie Z, Yin R, Zeppelini D, Crowther TW, Eisenhauer N, Scheu S.
673 2023. Globally invariant metabolism but density-diversity mismatch in springtails. *Nat*
674 *Commun* **14**:674. doi:10.1038/s41467-023-36216-6
675 Pryszcz LP, Gabaldón T. 2016. Redundans: an assembly pipeline for highly heterozygous
676 genomes. *Nucleic Acids Res* **44**:e113. doi:10.1093/nar/gkw294
677 Revell LJ. 2012. phytools: An R package for phylogenetic comparative biology (and other
678 things). *Methods Ecol Evol* **3**:217–223. doi:10.1111/j.2041-210X.2011.00169.x
679 Roberts E, Eargle J, Wright D, Luthey-Schulten Z. 2006. MultiSeq: unifying sequence and
680 structure data for evolutionary analysis. *BMC Bioinformatics* **7**:382.
681 doi:10.1186/1471-2105-7-382

682 Rosenberg Y, Bar-On YM, Fromm A, Ostikar M, Shoshany A, Giz O, Milo R. 2023. The
683 global biomass and number of terrestrial arthropods. *Sci Adv* **9**:eabq4049.
684 doi:10.1126/sciadv.abq4049
685 Russell RB, Barton GJ. 1992. Multiple protein sequence alignment from tertiary structure
686 comparison: assignment of global and residue confidence levels. *Proteins* **14**:309–
687 323. doi:10.1002/prot.340140216
688 Savory F, Leonard G, Richards TA. 2015. The Role of Horizontal Gene Transfer in the
689 Evolution of the Oomycetes. *PLOS Pathog* **11**:e1004805.
690 doi:10.1371/journal.ppat.1004805
691 Schaefer I, Caruso T. 2019. Oribatid mites show that soil food web complexity and close
692 aboveground-belowground linkages emerged in the early Paleozoic. *Commun Biol*
693 **2**:1–8. doi:10.1038/s42003-019-0628-7
694 Shear WA, Bonamo PM, Grierson JD, Rolfe WDI, Smith EL, Norton RA. 1984. Early Land
695 Animals in North America: Evidence from Devonian Age Arthropods from Gilboa,
696 New York. *Science* **224**:492–494. doi:10.1126/science.224.4648.492
697 Shelomi M, Heckel DG, Pauchet Y. 2016. Ancestral gene duplication enabled the evolution
698 of multifunctional cellulases in stick insects (Phasmatodea). *Insect Biochem Mol Biol*
699 **71**:1–11. doi:10.1016/j.ibmb.2016.02.003
700 Shelomi M, Watanabe H, Arakawa G. 2014. Endogenous cellulase enzymes in the stick
701 insect (Phasmatodea) gut. *J Insect Physiol* **60**:25–30.
702 doi:10.1016/j.jinsphys.2013.10.007
703 Shimodaira H, Hasegawa M. 1999. Multiple Comparisons of Log-Likelihoods with
704 Applications to Phylogenetic Inference. *Mol Biol Evol* **16**:1114.
705 doi:10.1093/oxfordjournals.molbev.a026201
706 Shin NR, Doucet D, Pauchet Y. 2022. Duplication of Horizontally Acquired GH5_2 Enzymes
707 Played a Central Role in the Evolution of Longhorned Beetles. *Mol Biol Evol*
708 **39**:msac128. doi:10.1093/molbev/msac128
709 Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO:
710 assessing genome assembly and annotation completeness with single-copy
711 orthologs. *Bioinformatics* **31**:3210–3212. doi:10.1093/bioinformatics/btv351
712 Song JM, Hong SK, An YJ, Kang MH, Hong KH, Lee Y-H, Cha S-S. 2017. Genetic and
713 Structural Characterization of a Thermo-Tolerant, Cold-Active, and Acidic Endo-β-
714 1,4-glucanase from Antarctic Springtail, Cryptopygus antarcticus. *J Agric Food Chem*
715 **65**:1630–1640. doi:10.1021/acs.jafc.6b05037
716 Steenwyk JL, Iii TJB, Li Y, Shen X-X, Rokas A. 2020. ClipKIT: A multiple sequence
717 alignment trimming software for accurate phylogenomic inference. *PLOS Biol*
718 **18**:e3001007. doi:10.1371/journal.pbio.3001007
719 Steinegger M, Salzberg SL. 2020. Terminating contamination: large-scale search identifies
720 more than 2,000,000 contaminated entries in GenBank. *Genome Biol* **21**:115.
721 doi:10.1186/s13059-020-02023-1
722 Sünnemann M, Siebert J, Reitz T, Schädler M, Yin R, Eisenhauer N. 2021. Combined
723 effects of land-use type and climate change on soil microbial activity and invertebrate
724 decomposer activity. *Agric Ecosyst Environ* **318**:107490.
725 doi:10.1016/j.agee.2021.107490
726 Tran N-V, Greshake Tzovaras B, Ebersberger I. 2018. PhyloProfile: dynamic visualization
727 and exploration of multi-layered phylogenetic profiles. *Bioinforma Oxf Engl* **34**:3041–
728 3043. doi:10.1093/bioinformatics/bty225
729 Ullah MR, Carrillo Y, Dijkstra FA. 2023. Relative contributions of fungi and bacteria to litter
730 decomposition under low and high soil moisture in an Australian grassland. *Appl Soil*
731 *Ecol* **182**:104737. doi:10.1016/j.apsoil.2022.104737
732 Undheim EAB, Jenner RA. 2021. Phylogenetic analyses suggest centipede venom arsenals
733 were repeatedly stocked by horizontal gene transfer. *Nat Commun* **12**:818.
734 doi:10.1038/s41467-021-21093-8
735 van den Hoogen J, Geisen S, Routh D, Ferris H, Traunspurger W, Wardle DA, de Goede
736 RGM, Adams BJ, Ahmad W, Andriuzzi WS, Bardgett RD, Bonkowski M, Campos-

Herrera R, Cares JE, Caruso T, de Brito Caixeta L, Chen X, Costa SR, Creamer R, Mauro da Cunha Castro J, Dam M, Djigal D, Escuer M, Griffiths BS, Gutiérrez C, Hohberg K, Kalinkina D, Kardol P, Kergunteuil A, Korthals G, Krashevska V, Kudrin AA, Li Q, Liang W, Magilton M, Marais M, Martín JAR, Matveeva E, Mayad EH, Mulder C, Mullin P, Neilson R, Nguyen TAD, Nielsen UN, Okada H, Rius JEP, Pan K, Peneva V, Pellissier L, Carlos Pereira da Silva J, Pitteloud C, Powers TO, Powers K, Quist CW, Rasmann S, Moreno SS, Scheu S, Setälä H, Sushchuk A, Tiunov AV, Trap J, van der Putten W, Vestergård M, Villenave C, Waeyenberge L, Wall DH, Wilschut R, Wright DG, Yang J, Crowther TW. 2019. Soil nematode abundance and functional group composition at a global scale. *Nature* **572**:194–198. doi:10.1038/s41586-019-1418-6

Watanabe H, Noda H, Tokuda G, Lo N. 1998. A cellulase gene of termite origin. *Nature* **394**:330–331. doi:10.1038/28527

Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biol* **20**:257. doi:10.1186/s13059-019-1891-0

Wybouw N, Pauchet Y, Heckel DG, Van Leeuwen T. 2016. Horizontal Gene Transfer Contributes to the Evolution of Arthropod Herbivory. *Genome Biol Evol* **8**:1785–1801. doi:10.1093/gbe/evw119

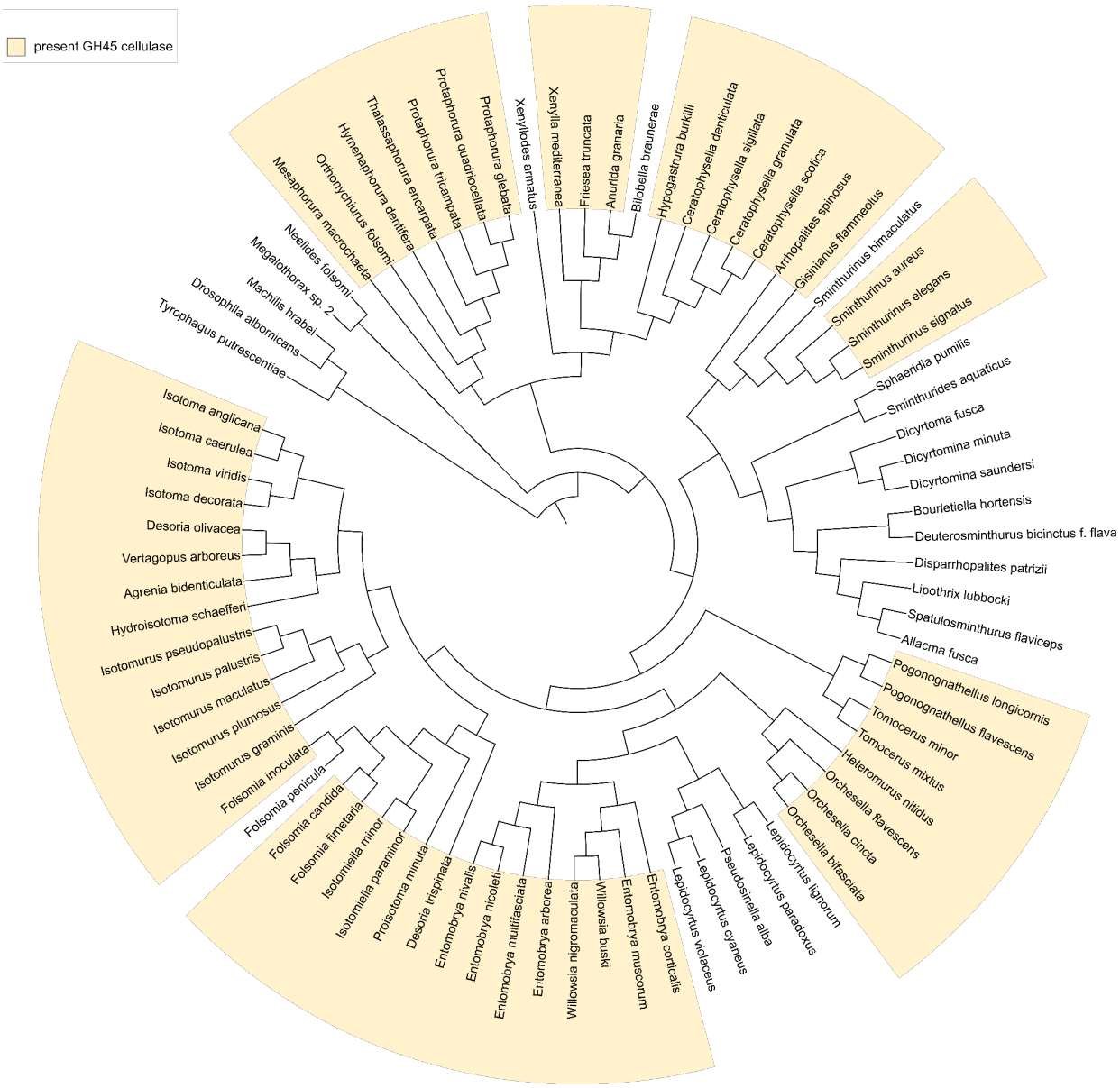
Zhou Z, Wang C, Luo Y. 2018. Effects of forest degradation on microbial communities and soil carbon cycling: A global meta-analysis. *Glob Ecol Biogeogr* **27**:110–124. doi:10.1111/geb.12663

Acknowledgements

The research was funded by the Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz (LOEWE) Program of the Hessian Ministry of Higher Education, Research, Science and the Arts through the LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG), and by the BA 4843/4-1 grant of the German Research Foundation (DFG). MB, IE, HM, SS, IS, AP, MP, YP, GC conceptualized the study. IE, HM, FA, MB, CS, JR, RL, KH, GC contributed methodology and materials. HM, FA, IE performed the analyses and visualized results. IE supervised the analyses. MB and IE wrote the original draft of the manuscript. HM, GC, TH, KH, RL, YP, MP, AP, JR, IS, SS, CS, IE and MB reviewed and finalized the manuscript. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Genome data will be available via NCBI BioProject PRJNA758215. The authors declare that they have no competing interests.

Supplementary Materials

Supplementary Figures



774

775 **Fig. S1.** Maximum likelihood phylogeny of springtails. Species with GH45-cellulase are
776 highlighted in yellow.

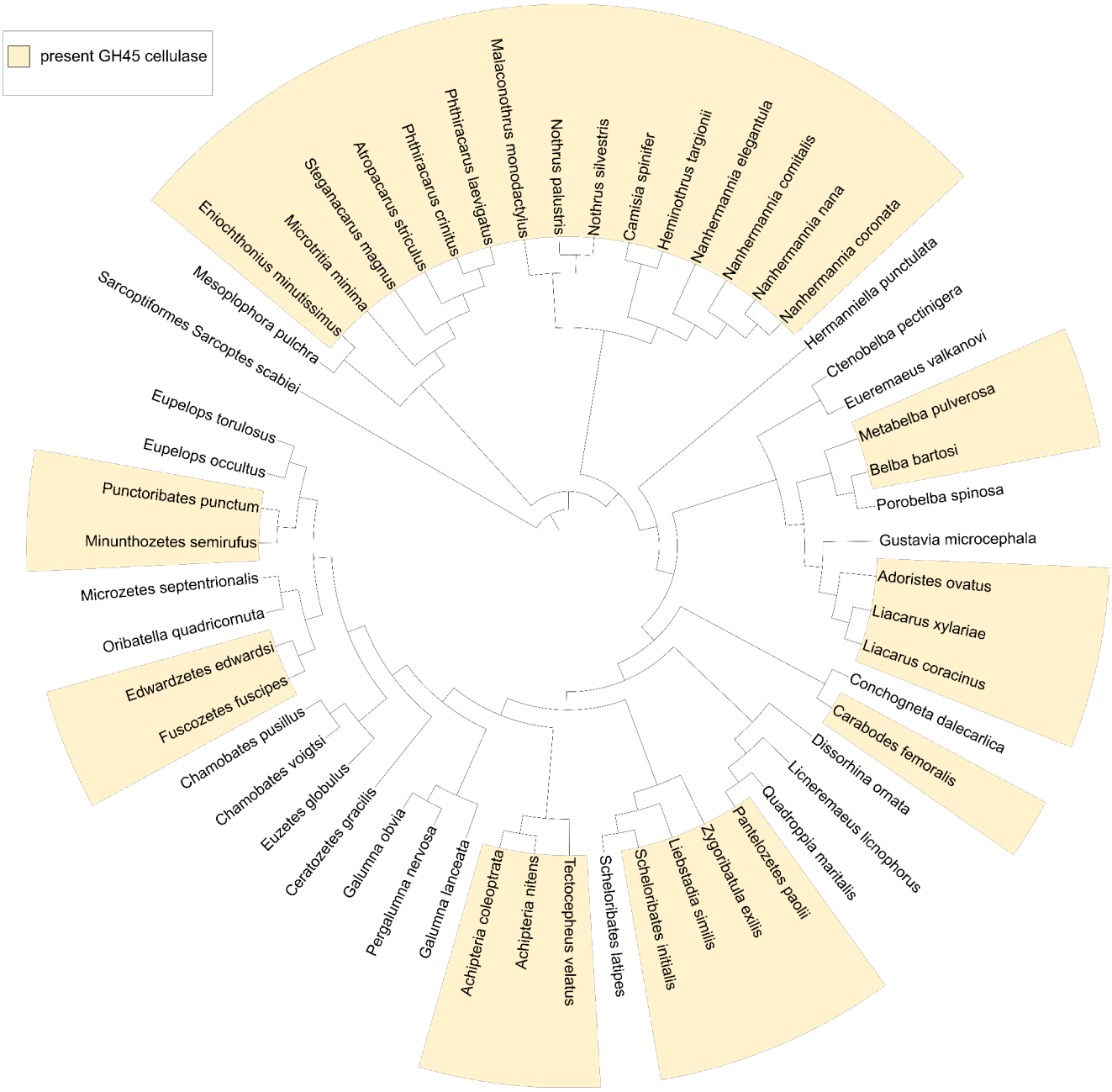


Fig. S2. Maximum likelihood phylogeny of oribatid mites. Species with GH45-cellulase are highlighted in yellow.

Fig. S3. Maximum likelihood phylogeny of all GH45-cellulase genes identified in soil invertebrates. Lines connect co-orthologs found in the same species. (uploaded separately)

Supplementary Tables

Table S1. Screened RefSeq genomes. Bacterial, archaeal, fungal, invertebrate and vertebrate genomes screened for GH45 ortholog presence. (large table, uploaded separately)

Table S2. Screened MetaInvert genomes. Soil invertebrate genomes of soil invertebrates screened with fDOG-Assembly for GH45 ortholog presence.

790

libid	countgroup	Scientific name	taxid	busco completness	Gh45 cellulase
a_17	Chilopoda	*Cryptops parisi*	173049	0,77	Absent
a_84	Chilopoda	*Geophilus carpophagus*	173285	0,82	Absent
P3_3	Chilopoda	*Geophilus flavus*	856749	0,72	Absent
a96	Chilopoda	*Geophilus truncorum*	173284	0,84	Absent
a_4	Chilopoda	*Haplophilus subterraneus*	173289	0,65	Absent
a_5	Chilopoda	*Henia vesuviana*	126936	0,8	Absent
a_21	Chilopoda	*Pachymerium ferrugineum*	115410	0,84	Absent
a_83	Chilopoda	*Stenotaenia linearis*	1569481	0,76	Absent
a42	Chilopoda	*Strigamia acuminata*	1255758	0,76	Absent
a39	Chilopoda	*Strigamia crassipes*	1428135	0,79	Absent
a_23	Chilopoda	*Strigamia transsilvanica*	1579475	0,7	Absent
a_66	Collembola	*Agrenia bidenticulata*	1933610	0,82	Present
MI_479	Collembola	*Allacma fusca*	39272	0,84	Absent
a_35	Collembola	*Anurida granaria*	187597	0,79	Present
a107	Collembola	*Arrhopalites spinosus*	187608	0,6	Present
a_42	Collembola	*Bilobella braunerae*	106916	0,57	Absent
a102	Collembola	*Bourletiella hortensis*	574228	0,76	Absent
P1_24	Collembola	*Ceratophysella denticulata*	928250	0,75	Present
a_34	Collembola	*Ceratophysella granulata*	1218962	0,78	Present
a52	Collembola	*Ceratophysella scotica*	187617	0,62	Present
a49	Collembola	*Ceratophysella sigillata*	1218965	0,66	Present
a72	Collembola	*Desoria olivacea*	370026	0,55	Present
MI_424	Collembola	*Desoria trispinata*	1184801	0,61	Present
a3_16	Collembola	*Deuterosminthurus bicinctus f. flava*	2041938	0,67	Absent
a_31	Collembola	*Dicyrtoma fusca*	1385863	0,8	Absent
a_32	Collembola	*Dicyrtomina minuta*	1387116	0,77	Absent
a_51	Collembola	*Dicyrtomina saundersi*	438492	0,75	Absent
P4_21	Collembola	*Disparrhopalites patrizii*	999999006	0,83	Absent
a_64	Collembola	*Entomobrya cf arborea*	30001	0,77	Present
a64	Collembola	*Entomobrya corticalis*	1503966	0,77	Present
a_33	Collembola	*Entomobrya muscorum*	2041940	0,53	Present
a_48	Collembola	*Entomobrya nicoleti*	2041941	0,7	Present
a71	Collembola	*Entomobrya nivalis*	1387109	0,66	Present
a108	Collembola	*Entomobrya Typ multifasciata*	247613	0,85	Present
a103	Collembola	*Folsomia candida*	158441	0,91	Present
a3_12	Collembola	*Folsomia fimetaria*	1387114	0,72	Present
a_38	Collembola	*Folsomia inoculata*	2041942	0,77	Present
MI_48	Collembola	*Folsomia penicula*	266765	0,59	Absent
a60	Collembola	*Friesea truncata*	187628	0,7	Present

a_40	Collembola	*Gisinianus flammeolus*	2449080	0,85	Present
P2_15	Collembola	*Heteromurus nitidus*	254095	0,81	Present
a98	Collembola	*Hydroisotoma schaefferi*	301519	0,85	Present
a58	Collembola	*Hymenaphorura dentifera*	999999008	0,72	Present
a_70	Collembola	*Hypogastrura burkilli*	1725397	0,71	Present
a105	Collembola	*Isotoma anglicana*	247611	0,66	Present
a_26	Collembola	*Isotoma caerulea*	308473	0,72	Present
a_71	Collembola	*Isotoma decorata*	57735	0,57	Present
P3_1	Collembola	*Isotoma viridis*	187635	0,75	Present
P1_25	Collembola	*Isotomiella minor*	370032	0,83	Present
a57	Collembola	*Isotomiella paraminor*	370031	0,58	Present
a_36	Collembola	*Isotomurus graminis*	1184803	0,72	Present
a_69	Collembola	*Isotomurus maculatus*	36143	0,66	Present
a97	Collembola	*Isotomurus palustris*	36144	0,81	Present
P1_26	Collembola	*Isotomurus plumosus*	1410395	0,53	Present
a_25	Collembola	*Isotomurus pseudopalustris*	36142	0,79	Present
a_53	Collembola	*Lepidocyrtus cyaneus*	247612	0,59	Absent
MI_426	Collembola	*Lepidocyrtus lignorum*	707889	0,59	Absent
a_27	Collembola	*Lepidocyrtus paradoxus*	49179	0,67	Absent
P2_2	Collembola	*Lepidocyrtus violaceus*	707891	0,81	Absent
a_46	Collembola	*Lipothrix lubbocki*	1387126	0,82	Absent
a4_24	Collembola	*Megalothorax sp. 2*	2340290	0,83	Absent
a68	Collembola	*Mesaphorura macrochaeta*	2651973	0,68	Present
MI_445	Collembola	*Neelides folsomi*	332381	0,75	Absent
a_29	Collembola	*Orchesella bifasciata*	576794	0,71	Present
a53	Collembola	*Orchesella cincta*	48709	0,53	Present
a_28	Collembola	*Orchesella flavescens*	48711	0,76	Present
a106	Collembola	*Orthonychiurus folsomi*	2581074	0,79	Present
a_49	Collembola	*Pogonognathellus flavescens*	511703	0,78	Present
a54	Collembola	*Pogonognathellus longicornis*	707266	0,84	Present
P2_17	Collembola	*Proisotoma minuta*	301521	0,84	Present
a61	Collembola	*Protaphorura glebata*	187683	0,79	Present
a_61	Collembola	*Protaphorura quadriocellata*	187683	0,72	Present
P2_19	Collembola	*Protaphorura tricampata*	187683	0,74	Present
a_41	Collembola	*Pseudosinella alba*	1302326	0,57	Absent
a104	Collembola	*Sminthurides aquaticus*	281415	0,8	Absent
a_39	Collembola	*Sminthurinus aureus*	1496267	0,84	Present
P4_22	Collembola	*Sminthurinus bimaculatus*	187699	0,81	Absent
P4_20	Collembola	*Sminthurinus elegans*	1190784	0,66	Present
a100	Collembola	*Sminthurinus signatus*	2584529	0,86	Present
a101	Collembola	*Spatulosminthurus flaviceps*	999999007	0,81	Absent
a3_8	Collembola	*Sphaeridia pumilis*	212016	0,87	Absent

P2_22	Collembola	*Thalassaphorura encarpata*	2583954	0,81	Present
P2_23	Collembola	*Tomocerus minor*	187706	0,58	Present
a55	Collembola	*Tomocerus mixtus*	58788	0,84	Present
a_43	Collembola	*Vertagopus arboreus*	2041954	0,62	Present
a_54	Collembola	*Willowsia buski*	1458441	0,77	Present
a_55	Collembola	*Willowsia nigromaculata*	1302335	0,8	Present
a99	Collembola	*Xenylla mediterranea*	2567731	0,78	Present
a62	Collembola	*Xenyllodes armatus*	187716	0,71	Absent
a94	Diplopoda	*Chordeuma sylvestre*	1569510	0,7	Absent
a_14	Diplopoda	*Cylindroiulus punctatus*	61981	0,68	Absent
a_96	Diplopoda	*Glomeris hexasticha*	1392624	0,66	Absent
a_10	Diplopoda	*Glomeris marginata*	62006	0,67	Absent
a_18	Diplopoda	*Julus scandinavius*	1008810	0,68	Absent
a_93	Diplopoda	*Julus scanicus*	541046	0,72	Absent
a_86	Diplopoda	*Kryphioiulus occultus*	1008825	0,72	Absent
a91	Diplopoda	*Megaphyllum sjaelandicum*	52423	0,7	Absent
a_95	Diplopoda	*Melogona broelemanni*	1147011	0,54	Absent
a95	Diplopoda	*Mycogona germanica*	999999013	0,67	Absent
a_13	Diplopoda	*Ommatoiulus sabulosus*	1008866	0,69	Absent
a_92	Diplopoda	*Ophyiulus pilosus*	118470	0,61	Absent
a_2	Diplopoda	*Polydesmus angustus*	1068628	0,72	Absent
a_15	Diplopoda	*Polydesmus complanatus*	510027	0,84	Absent
a_20	Diplopoda	*Proteroiulus fuscus*	88024	0,7	Absent
a_91	Diplopoda	*Rossiulus vilnensis*	999999014	0,72	Absent
a_12	Diplopoda	*Xestoiulus laeticollis*	1522044	0,74	Absent
MI_473	Enchytraeidae	*Cognettia cognettii*	1502715	0,58	Absent
a4_3	Enchytraeidae	*Enchytraeus crypticus*	913645	0,53	Absent
MI_474	Enchytraeidae	*Oconnorella tubifera*	913705	0,6	Absent
a35	Gamasina	*Phytoseiulus persimilis*	44414	0,8	Absent
a36	Gamasina	*Stratiolaelaps miles*	406085	0,83	Absent
a4_15	Nematoda	*Acrobeloides thornei*	96599	0,6	Absent
P2_7	Nematoda	*Aphelenchus avenae*	70226	0,59	Absent
MI_396	Nematoda	*Discolaimus major*	211252	0,61	Absent
a85	Nematoda	*Mesodorylaimus bastiani*	344383	0,55	Present
MI_395	Nematoda	*Panagrellus redivivus*	6233	0,55	Absent
a3_9	Nematoda	*Panagrolaimus detritophagus*	310956	0,55	Absent
P3_2	Nematoda	*Phasmarhabditis papillosa*	6243	0,64	Absent
a75	Nematoda	*Prionchulus punctatus*	293874	0,58	Present
D23	Nematoda	*Prismatolaimus dolichurus*	288633	0,53	Present
P2_4	Oribatida	*Achipteria coleoptrata*	229769	0,78	Present
P1_2	Oribatida	*Achipteria nitens*	229768	0,8	Present
a1	Oribatida	*Adoristes ovatus*	708363	0,72	Present

a2	Oribatida	*Atropacarus striculus*	229743	0,69	Present
a3	Oribatida	*Belba bartosi*	2241992	0,8	Present
a4	Oribatida	*Camisia spinifer*	198258	0,8	Present
MI_457	Oribatida	*Carabodes femoralis*	229793	0,74	Present
P1_32	Oribatida	*Ceratozetes gracilis*	1686620	0,77	Absent
a6	Oribatida	*Chamobates pusillus*	503572	0,75	Absent
a3_18	Oribatida	*Chamobates voigtsi*	198262	0,66	Absent
P4_1	Oribatida	*Conchogneta dalecarlica*	999999009	0,81	Absent
a7	Oribatida	*Ctenobelba pectinigera*	1401282	0,65	Absent
P1_12	Oribatida	*Dissorhina ornata*	2202870	0,8	Absent
a10	Oribatida	*Edwardzetes edwardsi*	2202872	0,78	Present
D2	Oribatida	*Eniochthonius minutissimus*	229763	0,63	Present
a11	Oribatida	*Eueremaeus valkanovi*	1401269	0,74	Absent
P1_14	Oribatida	*Eupelops occultus*	2234141	0,77	Absent
a13	Oribatida	*Eupelops torulosus*	198282	0,77	Absent
P4_3	Oribatida	*Euzetes globulus*	334610	0,79	Absent
P1_4	Oribatida	*Fuscozetes fuscipes*	1686651	0,77	Present
P1_7	Oribatida	*Galumna lanceata*	229834	0,81	Absent
D3	Oribatida	*Galumna obvia*	885392	0,79	Absent
P1_8	Oribatida	*Gustavia microcephala*	1685391	0,65	Absent
MI_402	Oribatida	*Heminothrus targionii*	2664691	0,77	Present
a15	Oribatida	*Hermanniella punctulata var. septentrionalis*	885393	0,77	Absent
a16	Oribatida	*Liacarus coracinus*	198285	0,7	Present
a17	Oribatida	*Liacarus xylariae*	198284	0,81	Present
MI_463	Oribatida	*Licneremaeus licnophorus*	999999011	0,77	Absent
P1_16	Oribatida	*Liebstadia similis*	1250587	0,8	Present
P4_5	Oribatida	*Malaconothrus monodactylus*	1797415	0,79	Present
MI_464	Oribatida	*Mesoplophora pulchra*	334620	0,79	Absent
MI_465	Oribatida	*Metabelba pulverosa*	229776	0,73	Present
a18	Oribatida	*Microtritia minima*	229747	0,75	Present
a19	Oribatida	*Microzetes septentrionalis*	999999012	0,74	Absent
P1_10	Oribatida	*Minunthozetes semirufus*	1979919	0,77	Present
P2_8	Oribatida	*Nanhermannia comitalis*	1979898	0,7	Present
D4	Oribatida	*Nanhermannia coronata cf.*	198290	0,75	Present
a20	Oribatida	*Nanhermannia elegantula*	66595	0,81	Present
MI_408	Oribatida	*Nanhermannia nana*	198291	0,78	Present
P1_11	Oribatida	*Nothrus palustris*	198293	0,74	Present
MI_467	Oribatida	*Nothrus silvestris*	66602	0,78	Present
P1_13	Oribatida	*Oribatella quadricornuta*	198298	0,75	Absent
P1_18	Oribatida	*Pantelozetes paolii*	1979943	0,74	Present
MI_412	Oribatida	*Pergalumna nervosa*	708370	0,78	Absent

a88	Oribatida	*Phthiracarus crinitus*	229740	0,82	Present
MI_492	Oribatida	*Phthiracarus laevigatus*	229740	0,81	Present
a22	Oribatida	*Porobelba spinosa*	2886740	0,79	Absent
a4_7	Oribatida	*Punctoribates punctum*	1720615	0,78	Present
P4_9	Oribatida	*Quadroppia maritalis*	1250640	0,7	Absent
a24	Oribatida	*Scheloribates initialis*	1979935	0,73	Present
D8	Oribatida	*Scheloribates latipes*	1979937	0,72	Absent
D7	Oribatida	*Steganacarus magnus*	52000	0,79	Present
a4_11	Oribatida	*Tectocepheus velatus*	229869	0,69	Present
P1_144	Oribatida	*Zygoribatula exilis*	1251916	0,76	Present
a4_20	Tardigrada	*Isohypsibius dastychi*	947160	0,64	Present
a3_3	Tardigrada	*Paramacrobiotus richtersi*	697321	0,57	Present

Table S3. Metazoan genomes with GH45. Taxonomic assignments of animal sequences identified by fDOG as orthologs of cellulases with a GH45-type Pfam domain.

Scientific name	NCBI ID	Accession number	Protein ID	Gene ID	Taxonomic assignment
Bradysia coprophila	38358	GCF_014529535.1	XP_037026636.1	119067642	*Bradysia odoriphaga*
			XP_037050424.1	119084512	Protostomia
			XP_037027558.1	119068175	*Bradysia odoriphaga*
Leptinotarsa decemlineata	7539	GCF_000500325.1	XP_023016322.1	111505702	Chrysomelinae
			XP_023029513.1	111517551	*Gonioctena quinquepunctata*
			XP_023016323.1	111505703	Chrysomelinae
			XP_023029514.1	111517551	*Gonioctena quinquepunctata*
			XP_023022929.1	111511149	Chrysomelini
			XP_023016326.1	111505705	Chrysomelini
Diabrotica virgifera virgifera	50390	GCF_003013835.1	XP_028147313.1	114340743	Chrysomelidae
			XP_028139473.1	114333726	Chrysomelidae
			XP_028143849.1	114337572	Chrysomelidae
			XP_028147314.1	114340743	Chrysomelidae
Anoplophora glabripennis	217634	GCF_000390285.2	XP_018561275.1	108903540	*Anoplophora chinensis*

			XP_018561265.1	108903530	Lamiinae
Dendroctonus ponderosae	77166	GCF_000355655.1	XP_019754618.1	109533680	Dryophthorinae
			XP_019754620.1	109533682	Dryophthorinae
			XP_019771468.1	109545306	Cucujiformia
			XP_019754619.1	109533681	Dryophthorinae
			XP_019766961.1	109542255	Chrysomelinae
Sitophilus oryzae	7048	GCF_002938485.1	XP_030751361.1	115878892	Dryophthorinae
			XP_030747083.1	115875708	*Rhynchophorus ferrugineus*
Thrips palmi	161013	GCF_012932325.1	XP_034236588.1	117642458	*Frankliniella occidentalis*
Frankliniella occidentalis	133901	GCF_000697945.2	XP_026287984.1	113213214	*Thrips palmi*
			XP_026287985.1	113213214	*Thrips palmi*
			XP_026289264.1	113214189	*Thrips palmi*
Folsomia candida	158441	GCF_002217175.1	XP_021945337.1	110843646	Entomobryomorpha
			XP_021948187.1	110845935	Protostomia

Table S4. Fungal GH45 inputs. Core group of fungal species containing GH45 orthologs, computed by fDOG. This ortholog group was used as input for the final ortholog searches with fDOG and fDOG-Assembly by querying RefSeq genome assemblies for GH45 presence.

Scientific name	NCBI ID	Accession number RefSeq gene set	Gene IDs
Rhizoctonia solani	456999	GCF_016906535.1	XP_043186467.1
Pleurotus ostreatus	5322	GCF_014466165.1	XP_036634094.1
Marasmius oreades	181124	GCF_018924745.1	XP_043007043.1
Pseudozyma flocculosa PF-1	1277687	GCF_000417875.1	XP_007880076.1
Kalmanozyma brasiliensis GHG001	1365824	GCF_000497045.1	XP_016292070.1
Ustilago maydis 521	237631	GCF_000328475.2	XP_011388317.1

Table S5. **Excluded orthologs**. RefSeq candidate orthologs excluded from the final orthology inference results due to missing the Pfam domain characteristic of GH45 cellulases.

Scientific name	NCBI ID	Class	RefSeq accession number	Gene ID from inferred ortholog
Osmia lignaria	473952	Insecta	GCF_012274295.1	XP_034173034.1
Bombus terrestris	30195	Insecta	GCF_000214255.1	XP_020720566.1 XP_012168823.1 XP_012168816.1 XP_012168808.1
Colletes gigas	935657	Insecta	GCF_013123115.1	XP_043266291.1 XP_043266289.1
Dufourea novaeangliae	178035	Insecta	GCF_001272555.1	XP_015435417.1
Rhagoletis zephyri	28612	Insecta	GCF_001687245.1	XP_017480646.1
Streptomyces lacrimifluminis	1500077	Actinomycetes	GCF_014646095.1	WP_189152206.1
Actinoplanes globisporus DSM 43857	1120949	Actinomycetes	GCF_000379645.1	WP_169516340.1
Brevibacterium jeotgali	1262550	Actinomycetes	GCF_007828155.1	WP_101587258.1
Aneurinibacillus danicus	267746	Bacilli	GCF_007991215.1	WP_146809708.1
Paenibacillus thalictri	2527873	Bacilli	GCF_004307995.1	WP_131011676.1
Chondromyces apiculatus DSM 436	1192034	Deltaproteobacteria	GCF_000601485.1	WP_197041519.1